frontiers | Research Topics

Artificial intelligence and bioinformatics applications for omics and multi-omics studies

Edited by
Angelo Facchiano, Margherita Mutarelli and Dominik Heider

Published in
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-8325-4445-7
DOI 10.3389/978-2-8325-4445-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Artificial intelligence and bioinformatics applications for omics and multi-omics studies

Topic editors

Angelo Facchiano — Institute of Food Sciences, National Research Council (CNR), Italy

Margherita Mutarelli — Institute of Applied Sciences and Intelligent Systems, Department of Physical Sciences and Technologies of Matter, National Research Council (CNR), Italy

Dominik Heider — University of Marburg, Germany

Citation

Facchiano, A., Mutarelli, M., Heider, D., eds. (2024). *Artificial intelligence and bioinformatics applications for omics and multi-omics studies*.

Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4445-7

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 30 January 2024
DOI 10.3389/fgene.2024.1371473

OPEN ACCESS

EDITED AND REVIEWED BY
Quan Zou,
University of Electronic Science and Technology of China, China

*CORRESPONDENCE
Angelo Facchiano,
angelo.facchiano@isa.cnr.it

RECEIVED 16 January 2024
ACCEPTED 18 January 2024
PUBLISHED 30 January 2024

CITATION
Facchiano A, Heider D and Mutarelli M (2024), Editorial: Artificial intelligence and bioinformatics applications for omics and multi-omics studies.
Front. Genet. 15:1371473.
doi: 10.3389/fgene.2024.1371473

Editorial: Artificial intelligence and bioinformatics applications for omics and multi-omics studies

Angelo Facchiano[1]*, Dominik Heider[2] and Margherita Mutarelli[3]

[1]Institute of Food Sciences, National Research Council (CNR), Avellino, Italy, [2]Department of Mathematics & Computer Science, University of Marburg, Marburg, Germany, [3]Institute of Applied Sciences and Intelligent Systems, National Research Council (CNR), Pozzuoli, Italy

KEYWORDS

artificial intelligence, multi-omics, systems biology, biomedical data science, machine learning

Editorial on the Research Topic
Artificial intelligence and bioinformatics applications for omics and multi-omics studies

Introduction

The omics sciences have revolutionized research in areas such as biology, biotechnology, medicine, and agri-food sciences. The production of large-scale datasets has led to strong demand for appropriate computational tools for their management, analysis, and interpretation. In the era of Big Data, this need has surged. The use of artificial intelligence is now widespread in the biomedical field. Of enormous impact are recent developments in the field of protein three-dimensional structure prediction, once considered achievable only with experimental techniques, however showing many limitations. The application prospects in all domains of molecular biology, genomics, and omics sciences are now a tangible reality.

Following up on a Research Topic already carried out in past years (Chicco et al., 2020), we introduce a new Research Topic of articles to present Artificial Intelligence and new bioinformatics applications and computational approaches for analyzing omics data, or the application of existing tools, toward a more complete interpretation of biological phenomena, with applications in personalized medicine and biotechnology.

The Research Topic includes 12 articles, of which 9 are classified as Original Research, 2 as Brief Research Report, and 1 as Methods.

Original research articles

Sun et al. introduced a new machine learning model, which integrates multi-omics data for accurate cancer subtype recognition. This model combines an adversarial generation network and the self-attention mechanism. By learning from multi-omics data, the new model efficiently identifies cancer subtypes, outperforming traditional methods. The study

demonstrates its effectiveness across various cancer datasets, highlighting its potential for improving cancer diagnosis and treatment strategies.

Ochoa and Hernandez-Lemus explored the functional impact of multi-omic interactions in breast cancer subtypes. They propose a comprehensive analysis framework to semi-automatically generate network models of regulatory constraints influencing biological functions. By analyzing multi-omics data, they identified significant functions enriched in various breast cancer molecular subtypes, highlighted new regulatory features, and demonstrated the capability of multi-omic regulatory networks to provide reliable models for understanding the connections between omics, thereby aiding in systematic generation of mechanistic hypotheses in cancer biology.

Liu et al. introduced a novel model for classifying gastric cancer subtypes. It utilizes a residual graph convolutional network, combining multi-omics data and patient similarity networks. The study demonstrates that the new approach significantly outperforms traditional methods in predictive performance. This approach offers potential advancements in understanding gastric cancer subtypes and could assist in developing more targeted treatments.

Luo et al. presented SupCAM, a method that improves the identification of chromosome clusters for karyotyping. The new approach involves pre-training the backbone network with supervised contrastive learning on ChrCluster, incorporating variable image composition by category, and introducing self-marginal loss. Fine-tuning the network results in a final model, with SupCAM achieving a 94.99% accuracy on the ChrCluster dataset, surpassing previous methods.

Zheng et al. proposed a machine learning method for predicting if pairs of enhancers and promoters physically interact. They built a model called HARD from the names of the four (epi)genomic signals included: the histone modification H3K27ac and ATAC-seq to represent chromatin accessibility, RAD21 subunit of cohesin that is important in loop formation and the Distance between the promoter and the enhancer and classify them using a random forest algorithm. The method was tested on enhancer-promoter interaction benchmarks from the BENGI database (Moore et al., 2020) and compared with two existing methods outperforming them in the majority of measures, thus proving to be a useful new approach to this important although complex task.

Ettetuani et al. presented article focused on gene expression analysis, using *p*-values to identify significant genes, gene ontology terms and similarity scores to understand biological pathways, regulation, and gene networks, and machine learning for gene prioritization. The study proposes using deep neural network algorithms for gene clustering based on regulatory pathways. The work validates findings through the detection of genetic interactions. Specific tissues with normalized gene expression and occurrence frequencies are considered, particularly in the context of glomerular diseases. The results highlight the relevance of genes like EGR1, IL33, BMP2, and SLAMF8 in glomerular diseases.

Wang et al. proposed a novel framework for predicting Alzheimer's disease risk genes by considering spatial and temporal features of gene expression data. Utilizing gene expression data from various tissues and age groups, a support vector machine model is developed. The work identified 19 crucial features from an initial set of 64, and 15 potential risk genes with a probability exceeding 90%, offering a promising approach for understanding Alzheimer's disease genetic etiology.

Jia et al. presented a deep learning tool to predict N4-acetylcytidine (ac4C), a post-transcriptional RNA modification highly conserved with a relevant function in transcription regulation and protein translation and associated with several human diseases. The authors tested different encoding approaches and classification models and found that a simple one-hot encoding and a downsampled ensemble deep learning network consisting of a modified DenseNet and Squeeze-and-Excitation Networks with a convolutional residual structure in parallel with the dense block gave the best performance results. The model outperformed two existing methods in a fair comparison, proving it is a promising new resource in predicting this important nucleoside modification.

The manuscript by Rahmani et al. described the development of an AI-based R package called MBMethPred to classify childhood medulloblastoma (MB) subtypes from DNA methylation and gene expression data. The two data types were combined using a similarity network fusion approach and feature selection was performed with random forests. The authors then applied six different machine-learning algorithms for subtype predictions, all scoring very good with a variety of performance measures and the selected biomarkers were challenged for biological and clinical relevance using survival and network analysis. The study represents a useful advancement towards the goal of accurate classification of molecular subgroups in MB patients that are vital to choose the best therapeutical plans for them.

Methods articles

Klinkhammer et al. presented an article describing the development of a boosting algorithm, called snpboost, for creating polygenic risk scores (PRS) directly from genetic data, with the aim of improving predictive accuracy in clinical risk stratification. The approach efficiently addresses the high-dimensional nature of genotype data and outperforms other methods in terms of predictive performance and computational time.

Klau et al. focused on improving disease risk prediction using polygenic risk scores (PRS). They investigated whether incorporating multiple PRS from different diseases and applying machine learning models can enhance predictive accuracy compared to traditional single-PRS models using regression. Their results show that multi-PRS models, especially when combined with deep learning techniques, significantly outperform single-PRS models in predicting risks for diseases like cancer, diabetes, and cardiovascular diseases. This advancement could lead to more effective disease prediction and personalized medicine approaches.

Brief research report article

Waechter et al. investigated the effectiveness of two different 16S rRNA primer sets for sequencing human fecal microbiomes using the Nanopore platform. They compared the conventional 27F primer included in the 16S Barcoding Kit by Oxford Nanopore Technologies and a more degenerate 27F primer. The study reveals

significant differences in the detection of taxonomic diversity and relative abundance of various taxa between these primer sets. The more degenerate primer set appears to provide a more accurate and diverse representation of the fecal microbiome compared to the conventional primer set.

Author contributions

AF: Writing–original draft, Writing–review and editing. DH: Writing–original draft, Writing–review and editing. MM: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Chicco, D., Heider, D., and Facchiano, A. (2020). Editorial: Artificial intelligence bioinformatics: development and application of tools for omics and inter-omics studies. *Front. Genet.* 11, 309. doi:10.3389/fgene.2020.00309

Moore, J. E., Pratt, H. E., Purcaro, M. J., and Weng, Z. (2020). A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol.* 21, 17–16. doi:10.1186/s13059-019-1924-8

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/fgene.2022.1032768

OPEN ACCESS

EDITED BY
Dominik Heider,
University of Marburg, Germany

REVIEWED BY
Markus List,
Technical University of Munich, Germany
Olga Zolotareva,
Technical University of Munich, Germany

*CORRESPONDENCE
Ping Gong,
gongping@xzhmu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 31 August 2022
ACCEPTED 15 December 2022
PUBLISHED 04 January 2023

CITATION
Sun Q, Cheng L, Meng A, Ge S, Chen J, Zhang L and Gong P (2023), SADLN: Self-attention based deep learning network of integrating multi-omics data for cancer subtype recognition. *Front. Genet.* 13:1032768. doi: 10.3389/fgene.2022.1032768

SADLN: Self-attention based deep learning network of integrating multi-omics data for cancer subtype recognition

Qiuwen Sun[1†], Lei Cheng[1†], Ao Meng[1], Shuguang Ge[2], Jie Chen[3], Longzhen Zhang[3] and Ping Gong[1*]

[1]School of Medical Imaging, Xuzhou Medical University, Xuzhou, China, [2]School of Information and Control Engineering, University of Mining and Technology, Xuzhou, China, [3]Department of Radiation Oncology, Affiliated Hospital of Xuzhou Medical University, Xuzhou, China

Integrating multi-omics data for cancer subtype recognition is an important task in bioinformatics. Recently, deep learning has been applied to recognize the subtype of cancers. However, existing studies almost integrate the multi-omics data simply by concatenation as the single data and then learn a latent low-dimensional representation through a deep learning model, which did not consider the distribution differently of omics data. Moreover, these methods ignore the relationship of samples. To tackle these problems, we proposed SADLN: A self-attention based deep learning network of integrating multi-omics data for cancer subtype recognition. SADLN combined encoder, self-attention, decoder, and discriminator into a unified framework, which can not only integrate multi-omics data but also adaptively model the sample's relationship for learning an accurately latent low-dimensional representation. With the integrated representation learned from the network, SADLN used Gaussian Mixture Model to identify cancer subtypes. Experiments on ten cancer datasets of TCGA demonstrated the advantages of SADLN compared to ten methods. The Self-Attention Based Deep Learning Network (SADLN) is an effective method of integrating multi-omics data for cancer subtype recognition.

KEYWORDS

self-attention, deep learning, multi-omics data, Gaussian mixture model, cancer subtype recognition

1 Introduction

Cancer is one of the most common and fatal diseases with high heterogeneity, that is same cancer will produce subtypes with different phenotypes, which will affect the clinical treatment and prognosis (Bray et al., 2018; Siegel et al., 2020). Therefore, the recognition of the cancer subtype is of great significance for the choice of treatment and prognosis of cancer patients (Hong Zhao et al., 2014). With the developments of high-throughput sequencing technology, there yield large amounts of multi-omics data, such as miRNA expression data, mRNA expression data, DNA methylation data, and copy number

variation etc. (Song et al., 2020). These multi-omics data can be obtained by some publicly available projects. For example, The Cancer Genome Atlas (TCGA) (Sayáns et al., 2019) stories more than 30 cancers over 11,000 patients' data and provides valuable opportunities for cancer subtype recognition. Existing studies have demonstrated that incorporating multi-omics data can obtain better performances and improve the understanding of cancer progression compared to using single-omic data (Hawkins et al., 2010; Kristensen et al., 2014; Hasin et al., 2017). Therefore, there is a strong need for integrated analysis of multi-omics data in cancer subtype recognition (Simidjievski et al., 2019; Xu et al., 2019; Picard et al., 2021).

The clustering algorithm is often used to recognize cancer subtypes. Researchers have proposed many clustering methods for multi-omics data integration. These methods can be divided into three categories: early integration, late integration, and intermediate integration (Rappoport and Shamir, 2018).

Early integration methods simply concatenate different omics' feature matrices to a single matrix and use the single omics clustering algorithm to subtype the matrix (Rappoport and Shamir, 2018). For example, K-means, LRAcluster, and Spectral clustering all belong to this category. Early integration methods do not consider the differences in the distribution and information contribution of each omics data, they increase the dimension of input data and exacerbate the dimension problem. In late integration, each omic data is clustered separately and the clustering solutions are integrated to obtain a single clustering solution. For example, COCA (Le et al., 2016) and PINS (Nguyen et al., 2017) belong to this category. Late integration methods ensure robustness against noise and bias, but the performance may be greatly affected when each omics data have different degrees of information contribution.

On the other hand, intermediate integration attempts to build a model that integrates all omics, including the method of integrating sample similarity, the method of using joint size reduction, and the method of using data statistical modeling. Similarity-based ensemble methods construct and fuse the sample similarity at each omics level to obtain consistent sample-sample relationships, and then perform cluster analysis. Typical methods include SNF (Wang et al., 2014) and NEMO (Rappoport and Shamir, 2019). These methods are very sensitive to data noise or network parameters due to the instability of the kernel distance function. An ensemble method based on dimensionality reduction is used to project each omics data into a common low-dimensional space, typical methods are CCA and MCCA (Witten and Tibshirani, 2009). However, these methods are susceptible to data noise and feature heterogeneity. Statistics-based ensemble methods build a statistical model to tackle ensemble challenges, including cluster (Shen et al., 2009), iClusterPlus and iClusterBayes.

As machine learning development, deep learning has been widely used in healthcare, such as imaging-based computer-aided diagnosis (Yu et al., 2021), digital pathology (Parodi et al., 2015), drug design (Peng et al., 2020), prediction of hospital admission (Zhang et al., 2022), classification of cancer (Zeng et al., 2021), and so on. With the advancement of the high learning capability and flexibility of deep neural networks, more and more deep learning based multi-omics integration methods have been proposed for cancer subtype recognition (Poirion et al., 2018; Guo et al., 2019). Most of them adopted autoencoder (AE) architecture, such as multi-omics autoencoder integration (MAUI) (Song et al., 2021), stacked sparse autoencoder (SSAE) (Xu et al., 2016), denoising autoencoder for accurate cancer prognosis prediction (DCAP) (Chai et al., 2021), which can efficiently leverage multi-omics datasets to learn latent factors of observed data in lower dimensions. However, these methods are almost based on early integration and ignore the distributions of different omics which would underestimate heterogeneous omics data (Wang et al., 2020). To solve these problems, some researchers have proposed deep learning based middle integration methods (Sharifi-Noghabi et al., 2019; Adossa et al., 2021; Picard et al., 2021). These methods separately learned each omics data through some subnetwork, and then integrated the output of every sub-network into a unified representation. For example, Tong et al. (2020) proposed ConcatAE, a method of concatenating features learned from each omics using an autoencoder. Yang et al. (2021a) proposed Subtype-GAN, an approach that used multi-input multi-output neural networks separately to model multi-omics data. Although these methods have demonstrated good performance in cancer subtype recognition, they ignore the relationship between samples when learning valuable feature representation. Different omics data types could provide unique characteristics to the patients' space. Therefore, it is crucial to utilize the relationship of patients to further boost learning performance.

More recently, attention mechanism has become a new technology in the field of deep learning. The dominant thought is to measure the similarity between the Key and the Query (Mercer and Neufeld, 2021). Attention mechanism has been applied in speech NLP, image and other fields (Luo et al., 2018; Yuan et al., 2018; Li et al., 2020a; Liu et al., 2020), since it can select the most informative features of an input, adaptively consider the importance of a single feature and allow the model to make a more accurate judgment. As a special, self-attention (Shaw et al., 2018; Hou et al., 2019), which calculates the response at a position in the sequence by attending to all positions within the same sequence has achieved notable success in modeling complicated relations (Gao et al., 2019). For instance, it displays the superiority in machine translation (Zhang et al., 2020), sentence embedding (Li et al., 2020b) of modeling arbitrary word dependency and has been successfully applied to capture node similarities in graph embedding (Mustafa Abualsaud, 2019). Research shows that the attention-based encoder is more fit for learning high-level features (Chen et al., 2021).

To this end, we proposed SADLN: a self-attention based deep learning network integrating multi-omics data for cancer subtype

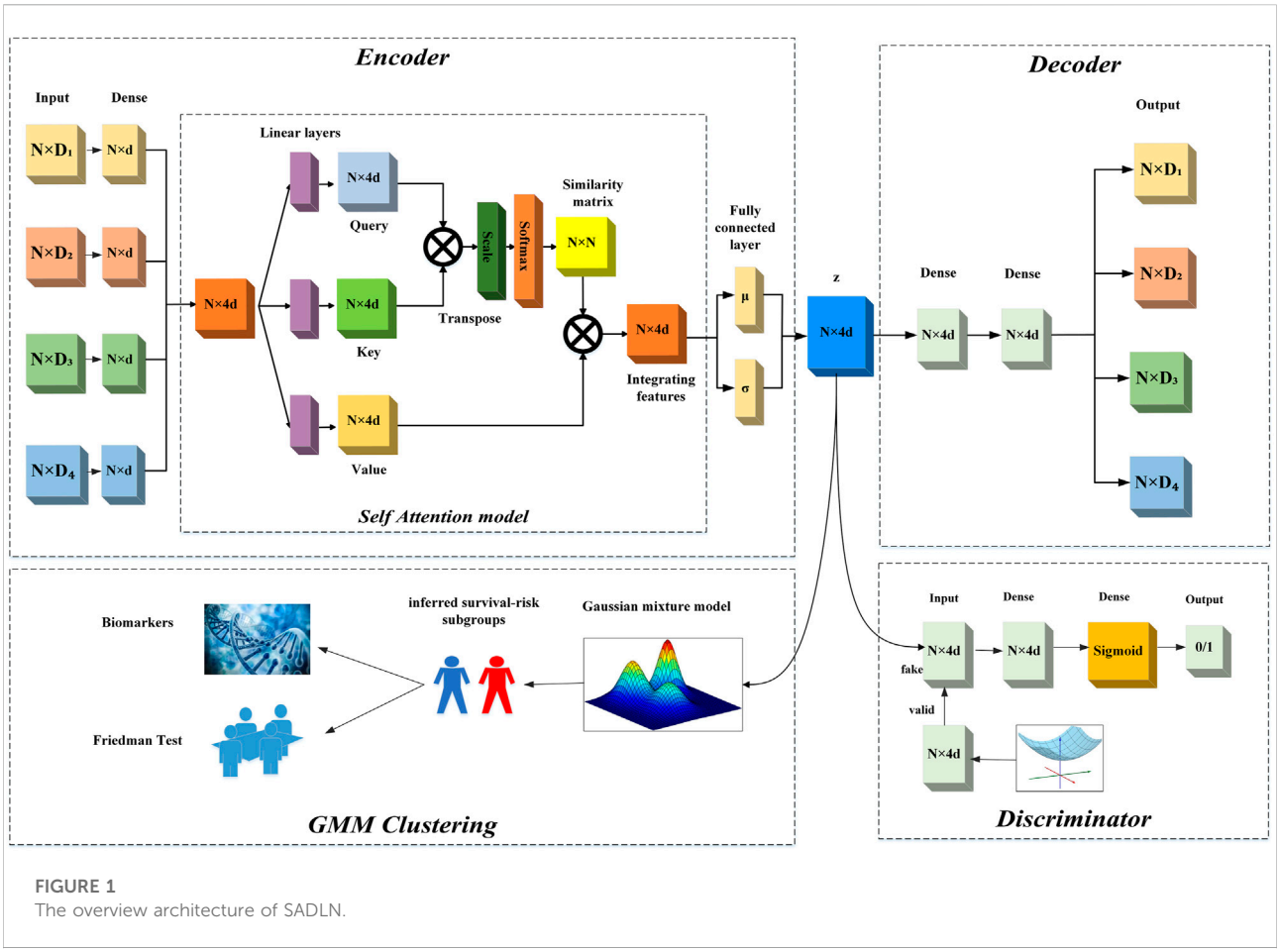


FIGURE 1
The overview architecture of SADLN.

recognition. SADLN is a middle integration method by consolidates the adversarial generation network and the self-attention mechanism to describe the different distributions of multi-omics data and fusion samples' relationship. It used an independent sub network to learn omics-specific features and concatenated omics-specific features to an integration representation. Then used a self-attention to learn the relationship of samples on the integration representation and obtained a feature representation that fused the sample relationship. Finally, it used the Gaussian Mixture Model (GMM) to obtain the subtyping label of each sample.

The main contribution is summarized as follows:

1) We proposed a novel deep learning method, SADLN, which combines encoder, self-attention, decoder, and discriminator into a unified framework. It can simultaneously integrate multi-omics representation and sample relations.
2) We firstly introduced the self-attention into the deep learning based method for the cancer subtyping recognition task which allows the model to automatically learn the similarity of samples for better representation.
3) We conducted experiments on ten cancer datasets of TCGA, and SADLN achieved outstanding performance compared with ten integration methods. It provided the theoretical basis and a new method for clinical diagnosis and precise treatment of cancer, which has great theoretical significance and clinical application value.

2 Methodology

Our proposed method consists of two steps. Firstly, we used the SADLN model to learn an integrated feature representation from multi-omics data. Secondly, with the learned feature representation, we used the GMM to identify sample's subtypes. In the SADLN model, the input is the sample's multi-omics data and the output is the sample's integrated low-dimensional feature representation. The model consists of three main blocks: self-attention based encoder, decoder and discriminator. Figure 1 gives the overview architecture of our proposed method. In the following, we describe each block in more detail.

2.1 Self-attention based encoder

To be able to generate higher quality data distribution, we design a self-attention based encoder in our SADLN model as shown in Figure 1. The self attention based encoder transforms the multi-omics data into a low-dimensional latent space representation z with distribution $\mathbf{N}(\mu, \sigma)$ using multiple independent network layers, a fully connected layer and the self attention layer. We used four sub-independent dense network to extract features from each original omics data. For each sub-independent layer, let $\mathbf{x}^m = \{x_1^m, \ldots, x_N^m\} \in R^{N\times D}$ denotes the input of the network for the m-th omics data, $\mathbf{y}^m = \{y_1^m, \ldots, y_N^m\} \in R^{N\times d}$ denotes the output of the m-th omics through the sub-independent layer, where N is the number of data samples, D and d are the feature dimension of the input data and the output data respectively. $\mathbf{y}^m$ can be express as:

$$\mathbf{y}^m = \mathbf{w}_m \mathbf{x}^m + \mathbf{b}^m \tag{1}$$

where $\mathbf{w}_m$ is the weight matrix, $\mathbf{b}_m$ is the bias.

To fusion features from different omics data, we concatenate four features matrices into a feature representation matrix. The integrating feature matrix $\mathbf{Y}$ can be expressed as:

$$\mathbf{Y} = Concat\left(\mathbf{y}^1, \ldots, \mathbf{y}^4\right) \tag{2}$$

For example, if the outputs of the sub-networks is a $N \times d$ feature matrix, after concatenation, the output will be one $N \times 4d$ feature representation matrix. To prevent the model overfitting, we appended batch normalization layers and used the Gaussian Error Linear Unit (GELU) function as the non-linear activation function. That is:

$$\mathbf{Y}' = GELU\left(\mathbf{Y}\right) \tag{3}$$

Although the concatenation operation can integrate multi-omics data, the relationship between samples is not considered. In this study, we introduced self-attention mechanism to construct the relationship between samples. Self-attention is typically used to model the relationship of words in a sentence, we treat each sample's features vector as a word and learn the samples' weight matrix through the sample's feature vectors.

Let $d_k = 4d$, $\mathbf{K} = [k_1, k_2, \ldots, k_N] \in R^{N\times d_k}$ is a set of keys, $\mathbf{Q} = [q_1, q_2, \ldots, q_N] \in R^{N\times d_k}$ is a set of queries, $\mathbf{V} = [v_1, v_2, \ldots, v_N] \in R^{N\times d_k}$ is a set of values, $\mathbf{K} = \mathbf{Q} = \mathbf{V} = \mathbf{Y}'$, $\mathbf{K} = \mathbf{Y}'\mathbf{W}^K$, $\mathbf{Q} = \mathbf{Y}'\mathbf{W}^Q$, $\mathbf{V} = \mathbf{Y}'\mathbf{W}^V$. $\mathbf{W}^K$, $\mathbf{W}^Q$, $\mathbf{W}^V$ are the parameters of linear projection layers. $Z = \{z_1, z_2, \ldots, z_N\} \in R^{N\times d_k}$ denotes the finally integrating representation, the jth feature vector z_j is computed as the following steps (Yang et al., 2021b). Firstly, we use the dot-product between q_i and k_j to compute the similarity of the sample i and j. To ensure the result does not get excessively large, we scale it by $\sqrt{d_k}$. That is:

$$r_{i,j} = \frac{q_i \times k_j^T}{\sqrt{d_k}} \tag{4}$$

Secondly, softmax function was used to obtain the similarity weight. That is:

$$\omega_i = softmax\left\{\frac{q_i \times k_1^T}{\sqrt{d_k}}, \frac{q_i \times k_2^T}{\sqrt{d_k}}, \ldots, \frac{q_i \times k_N^T}{\sqrt{d_k}}\right\} \tag{5}$$

Thirdly, the integrated feature vector z_i of sample i can be obtained by a weighted sum of the values. That is:

$$z_i = Attention\left(q_i, \mathbf{K}, \mathbf{V}\right) = \sum_{j=1}^{N} \omega_i v_j \tag{6}$$

Finally, the integrated feature representation can be express as:

$$\mathbf{Z} = Attention\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = [z_1, z_2, \ldots z_N] \in R^{N\times d_k} \tag{7}$$

To keep the data distribution unchanged, we added batch normalization layers after the self-attention model.

Suppose $\mathbf{Z}$ obeys Gaussian distribution $\mathbf{Z} \sim N(\mu, \sigma^2)$, where μ is the mean and σ^2 is the variance. In this paper, we obtained μ and σ^2 through two fully-connected layers.

2.2 Decoder

Decoder, in our SADLN model attempts to reconstruct the original multi-omics data from the integrating representation $\mathbf{Z}$. As shown in the upper right halves of Figure 1, it contains fully connected layers and an output layer. Let $\mathbf{X}_I = \{\mathbf{x}_I^1, \mathbf{x}_I^2, \mathbf{x}_I^3, \mathbf{x}_I^4\}$ denotes the input of encoder, $\mathbf{X}_O = \{\mathbf{x}_O^1, \mathbf{x}_O^2, \mathbf{x}_O^3, \mathbf{x}_O^4\}$ denotes the output of decoder. To minimize the error between the input $\mathbf{X}_I$ and the output $\mathbf{X}_O$ (Badrinarayanan et al., 2017), the square Euclidean distance was applied to calculate the loss $L_{Decoder}$, it can be expressed as:

$$L_{Decoder} = \|\mathbf{X}_I - \mathbf{X}_O\|_2^2 = \frac{1}{4}\sum_{k=1}^{4} \|\mathbf{x}_I^k - \mathbf{x}_O^k\|_2^2 \tag{8}$$

2.3 Discriminator

To force the distribution of the integrated feature representation matches the prior Gaussian distribution, we added a discriminator D to the model, which is a part of the GAN network. A typical GAN network is composed of a generator G and a discriminator D. In this work, we regard the self-attention base encoder part as the generator G, the input of the discriminator D is the output of the encoder part, and the randomly sampled data with Gaussian distribution. Let $G(z)$ denote the function of the generator, and $P(z)$ denote the prior Gaussian distribution. The discriminator D is used to distinguish the samples from $P(z)$ or the $G(z)$ (Yang et al., 2021a). Through adversarial learning, $G(z)$ is as close to $P(z)$ as possible.

The objective function optimization of discriminator D adopts the method of maximization and minimization. It can be expressed as:

$$\min_G \max_D E_{z'\sim P(z)}\left[logD\left(z'\right)\right] + E_{z\sim G(z)}\left[log\left(1-D\left(z\right)\right)\right] \quad (9)$$

where E represents the expected value of the distribution function. We use the binary_crossentropy function to train the discriminator learning process. The loss of the discriminator is:

$$L_{Discr} = -E_{z'\sim P(z)}\left[log\, D\left(z'\right)\right] - E_{z\sim G(z)}\left[log\left(1-D\left(z\right)\right)\right] - E_{z\sim G(z)}\left[log\left(D\left(z\right)\right)\right] \quad (10)$$

Our model parameters of the whole network are jointly trained by minimizing the following total loss:

$$L = \lambda_1 L_{Decoder} + \lambda_2 L_{Discr} \quad (11)$$

where $L_{Decoder}$ and L_{Discr} are defined in Eq.8 and Eq. 11, respectively. λ_1 and $\lambda_2 \in [0, 1]$ are trade-off parameters.

2.4 The GMM clustering of SADLN

For the generated feature representation $Z = \{z_n\}_{n=1}^N$, we use GMM to identify sample's subtypes. GMM is a probabilistic clustering method, which also belongs to the generative model. It assumes that all the data points are generated from a mixture of a finite number of Gaussian distributions (Gu et al., 2020). GMM model has excellent clustering performance. In this paper, we use GMM as the clustering module. Let K denotes the number of clusters, $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ represent the weight of each cluster, $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ is the mean vector, $\sum = (\sum_1, \sum_2, \ldots, \sum_k)$ is the covariance vector, $Z = \{z_n\}_{n=1}^N$ is the final integrated feature representation, $p(z_n)$ is the probability distribution function as a mixture of K Gaussian distributions. That is:

$$p\left(z_n\right) = \sum_{k=1}^{K} \pi_k p_k\left(z_n\right) = \sum_{k=1}^{K} \pi_k N\left(z_n|\mu_k, \sum_k\right) \quad (12)$$

GMM used the EM algorithm to update the parameters π, μ and $\sum$. According to the maximum probability density of the sample in different clusters, the most suitable subtype labels are obtained.

3 Experiments and analysis

3.1 Network structure and hyperparameter setting

The SADLN model has 19 layers, including 10 layers of the encoder, five layers of the decoder, and four layers of the discriminator. The specific network structure of SADLN is shown in Table 1. The model is built based on python 3.6.12, Keras 2.2.4, and TensorFlow 1.14.0 (the CPU version). The operating system is Windows 10. In terms of hardware, the CPU is Intel(R) Core (TM) i7-105 10U.

Optimizing hyperparameters are the key to training neural network models. Choosing appropriate hyperparameters can significantly improve the performance of the model. In this paper, the hyperparameters of the SADLN model mainly include the feature dimension of the independent sub network (d), the initial epoch, batch size, random seed, optimizer, activation function, learning rate and loss. Table 2 shows the hyperparameter settings of the SADLN model.

3.2 Datasets and evaluation metrics

To evaluate the performance of our proposed method SADLN, we used ten TCGA cancers datasets provided by (Yang et al., 2021a) from https://github.com/haiyang1986/Subtype-GAN. The datasets include BRCA, LUAD, BLCA, PAAD, KIRC, STAD, UVM, GBM, SKCM, and UCEC. These ten datasets contain sufficient samples and have reasonable numbers of subtypes. There are four types of omics data for each cancer: copy number, DNA methylation, mRNA and miRNA. The datasets have been preprocessed and feature selection was performed. The preprocessing steps of four types data are as follows (Hoadley et al., 2018). The DNA methylation data were combined from two generations of Infinium arrays, HumanMethylation27 (HM27) and HumanMethylation450 (HM450). Firstly, the HM27 data against the HM450 data was normalized of 0–1 for β-values using a probe-by-probe proportional rescaling method. Then, 3,139 CpG sites were selected that were methylated at a β-value of ≥ 0.3. For mRNA and miRNA data, firstly, the log transformation was performed separately, then poorly expressed genes were excluded based on median-normalized counts, and finally variance filtering was used to reduced features. Pre-processing led to 3,217 mRNA and 382 miRNA features. For copy number data, firstly, genomic regions along a chromosome defined by consecutive positions with a maximum Euclidean distance (based on copy number log-ratio segmented values) between any adjacent two probes smaller than 0.01 were formed; this resulted in a total of 3,105 copy number regions. Then each region was represented by its medoid signature, led to 3,105 copy number features. Finally, 3,105 copy number features, 3,217 mRNA features, 383 miRNA features and 3139 DNA methylation features were extracted from the original data source.

We used two evaluation metrics to evaluate the effect of cancer subtype recognition: survival analysis and clinical enrichment analysis. Survival analysis was obtained by the Cox log-rank test (Rainer and Muche and hosmer, 2001) to measure differential survival between subtypes. Smaller p-value indicates significant differences in survival profiles of different

TABLE 1 The network structure of SADLN.

Architectures	SADLN	
Self-attention based encoder	3,105 + 3,217 + 383+3,139 (Input)	
	25 + 25+25 + 25 (concatenate)	
	100 (Batch normalization)	
	100 (Activation)	
	100 (Attention)	
	100 (Batch normalization)	
	100 (Fully-connected)	
	100 (Fully-connected, Mean)	100 (Fully-connected, VAR)
	100 (Output)	
Decoder	100 (Input)	
	100 (Fully-connected)	
	100 (Batch normalization)	
	100 (Activation)	
	3,105 + 3,217 + 383 + 3,139 (Output)	
Discriminator	100 (Input)	
	1 (Fully-connected)	
	1 (Sigmoid)	
	1 (Output)	

TABLE 2 Hyperparameter settings of SADLN model.

Hyperparameter	Setting
d	25
Epoch	600
Batch size	64
Random seed	2
Optimizer	Adam optimizer
Activation function	Gaussian error linear unit
Learning rate (lr)	1e-4, 2e-4, 3e-4, 4e-4, 5e-4, 1e-5, 2e-5, 3e-5, 4e-5, 5e-5
Loss λ_1	1
Loss λ_2	0.0001

subtypes. In the clinical enrichment analysis, the differences in clinical indicators between subtypes were measured by the *p*-value obtained by Kruskal-Wallis test and Chi-square test for numerical and discrete clinical labels of cancer, respectively. Smaller *p*-value indicates significant differences between subtypes on this clinical label. Six clinical labels (Rappoport and Shamir, 2018) including age at diagnosis, gender, pathologic T, pathologic N, pathologic M, and pathologic stage were used for testing. The four latter parameters are discrete pathological parameters, measuring the size and extend of the primary tumor (T), the number of nearby lymph nodes that have cancer (N), whether the cancer has

TABLE 3 The $-\log10p$ values of ablation studies in ten cancer datasets on TCGA (bold indicates that this method performs best on the corresponding cancer dataset).

Cancer	SADLN	SADLN (NO SA)	SADLN ($\lambda_2 = 0$)	SADLN ($\lambda_1 = 0$)
BLCA	2.4	**2.5**	1.7	0.3
BRCA	**2.6**	2.3	0.4	0.4
GBM	**1.8**	1.7	0.4	0.2
KIRC	**6.6**	5.7	5.4	1.9
LUAD	**3.1**	2.4	0.3	0.04
PAAD	**3.2**	1.7	2.2	0.3
SKCM	3.0	**4.5**	0.9	0.8
STAD	1.4	1.3	**2.0**	0.3
UCEC	4.0	**5.4**	4.7	1.8
UVM	**4.5**	4.2	2.2	0.2

metastasized (M)and the total progression (pathologic stage). Cancer's clinical parameters were not all available, such as GBM and UCEC only have two clinical parameters.

To avoid the influence of small cluster size on the accuracy of evaluation metrics, the permutation test (Rappoport and Shamir, 2018) was applied to calculate the *p*-value of Cox log-rank test in survival analysis and Chi-square test in clinical enrichment analysis. Permutation test obtains an empirical *p*-value using the test statistic by permuting the cluster labels between samples. To perform permutation tests, we randomly permuted the clustering assignments of the different samples. For the log-rank test, the number of permutations we performed for each clustering solution was first $\min((\max\frac{10}{\text{original } p-value}, 1e4), 1e6)$ and then another $1e5$ permutations until the stopping condition was met. The stopping condition was having both the lower and upper ends of the 95% confidence interval for the *p*-value to be within 10% of its estimate, and such that the interval did not cross .05. For the clinical enrichment test, we continued on performing $1e3$ permutations until the 95% confidence interval did not cross 0.05, up to a maximum of $1e5$ iterations. This maximum number of iterations was only needed in case the *p*-value was extremely close to 0.05.

3.3 Ablation studies

To evaluate the contributions of key component of our model, we perform ablation studies in this section. There are three key modules in SADLN, self attention, decoder and discriminator. We separately removed these modules from SADLN, Table 3 gives the results of ablation studies in ten cancer datasets on TCGA.

From Table 3, we can see that, compared with the model without the attention module, namely SADLN (NO SA), SADLN achieved better values on seven cancer datasets (BRCA, KIRC, GBM, LUAD, PAAD, STAD, and UVM). Compared with removing the discriminator module ($\lambda_2 = 0$), SADLN obtained better value on eight cancer datasets (BRCA, GBM, KIRC, LUAD, PAAD, SKCM, UCEC, and UVM). The $-\log10p$ values of removing the decoder module ($\lambda_1 = 0$) are lower than SADLN. These results indicates that three modules play an important role in addressing the issue of feature generation.

3.4 Comparison with other state-of-the-art algorithms

To verify the performance of SADLN, we compared it with ten state-of-the-art methods. Three deep learning based methods include AE, VAE and Subtype-GAN and seven non-deep learning based methods include K-means, LRAcluster, iCluster, Spectral, NEMO (Rappoport and Shamir, 2019), MCCA (Witten and Tibshirani, 2009) and SNF (Wang et al., 2014). These ten methods can represent different types of approaches for integrating multi-omics data. AE and VAE belong to early integration methods, both input and output are integrated multi-omics data. Subtype-GAN belong to middle integration method, the input and output are multi-omics features. For ten comparison algorithms, (Yang et al., 2021a) detailed the network structure, parameter selection and execution details its Supplementary Materials Note 1 and Note 2. In this study, we rigorously implement these algorithms following the guidelines of (Yang et al., 2021a).

To reduce the influence of different clustering numbers on the results of subtyping, following the work (Yang et al., 2021a), we set the cluster number of BRCA, LUAD, BLCA, PAAD, KIRC, STAD, UVM, GBM, SKCM and UCEC were 5, 3, 5, 2, 4, 3, 4, 4, 4, 4, respectively. These cluster numbers of different cancers have

TABLE 4 The cluster number and subtypes of ten cancers.

Cancer	Cluster number	Subtypes
BRCA	5	LumA, LumB, Her2, Basal, Normal
LUAD	3	Terminal respiratory unit, Proximal inflammatory, Proximal proliferative
BLCA	5	Luminal-papillary, Luminal-infilitrated, Luminal, Basal/Squamous, Neuronal
PAAD	2	Basal-like/Squamous, Classical/Progenitor
KIRC	4	KIRC-M1, KIRC-M2, KIRC-M3, KIRC-M4
STAD	3	Immunity-Deprived (ImD), Stroma-Enriched (StE), Immunity-Enriched (ImE)
UVM	4	Disomy 3 (D3)-UVM-1, D3-UVM-2, Monosomy 3 (M3)-UVM-3, M3-UVM-4
GBM	4	Proneural, Neural, Classical, Mesenchymal
SKCM	4	Mutant BRAF, Mutant RAS, Mutant NF1, Triple-WT (wild-type)
UCEC	4	POLE (ultramutated), MSI (hypermutated), Copy-number high (serous-like), Copy-number low (endometrioid)

TABLE 5 The –log10*p* values of survival analysis based on Cox log-rank model of ten cancers datasets on TCGA (bold indicates that this method performs best on the corresponding cancer dataset).

Cancer	SADLN	Subtype-GAN	AE	VAE	K-means	Spectral	LRA-cluster	SNF	NEMO	MCCA	iCluster
BLCA	2.4	**2.5**	0.1	0.1	0.6	1.8	0.1	1.2	2.3	1.1	1.0
BRCA	2.6	2.3	0.1	0.3	0.2	0.1	0.2	2.2	1.0	**2.7**	0.7
GBM	1.8	1.7	1.1	1.0	2.3	**2.6**	0.9	1.2	2.4	1.0	2.1
KIRC	6.6	5.7	2.6	6.0	4.2	4.6	**7.0**	4.4	4.3	7.0	3.9
LUAD	**3.1**	2.4	0.7	1.4	1.0	0.6	0.3	1.5	2.2	0.9	0.6
PAAD	**3.2**	1.7	0.1	2.5	2.3	2.0	2.2	2.1	2.0	2.1	1.0
SKCM	3.0	4.5	0.0	2.4	2.1	1.9	1.5	3.8	**4.7**	0.9	1.1
STAD	**1.4**	1.3	0.1	0.0	0.1	0.3	0.1	0.5	1.1	1.3	0.4
UCEC	4.0	5.4	0.4	5.4	5.7	0.8	4.2	5.0	**6.0**	5.0	1.3
UVM	**4.5**	4.2	2.7	2.1	1.6	1.9	2.3	2.5	2.3	2.4	1.1

been proved to be clinically informed (Berger et al., 2018; The Cancer Genome Atlas Research Network, 2013; Robertson et al., 2017a; The Cancer Genome Atlas Research Network, 2014; Levine, 2013; Akbani et al., 2015; Li and Wang, 2021; Verhaak et al., 2010; Raphael et al., 2017; Robertson et al., 2017b). Table 4 gives the cluster number and subtypes of ten cancers. For example, in a previous study, GBM was classified into Classical, Mesenchymal, Neural, and Proneural subtypes based on mRNA expression data (Verhaak et al., 2010).

Table 5 gives the –log10*p* values of survival analysis for eleven methods of ten cancer datasets on TCGA. The clustering results of the other ten compared methods come from Yang's literature (Yang et al., 2021a). Bold indicates that this method performs best on the corresponding cancer dataset.

As shown in Table 5, SADLN achieved the most significant results on PAAD, STAD, LUAD and UVM cancer datasets. Compared with Subtype-GAN, SADLN obtained better value on seven cancer datasets (BRCA, GBM, KIRC, LUAD, PAAD, STAD, and UVM). Compared with AE, SADLN obtained the best –log10p-value in ten cancer datasets. Compared with non-deep learning based methods, although same methods had best results in specific cancer datasets, the –log10*p*-value was highest on most cancer datasets.

Table 6 gives the clinical parameters enrichment analysis result of SADLN and other compared methods of ten cancer datasets.

From Table 6, we can see that SADLN obtained the best results on four datasets (KIRC, GBM, STAD, UCEC). Therefore,

TABLE 6 The clinical parameters enrichment analysis of SADLN and other methods of ten cancer datasets on TCGA (bold indicates that this method performs best on the corresponding cancer dataset).

Methods	BRCA	LUAD	BLCA	PAAD	KIRC	STAD	UVM	GBM	SKCM	UCEC
SADLN	5	3	5	2	**6**	**3**	0	**1**	1	**1**
Subtype-GAN	**6**	**5**	5	2	6	2	**2**	1	**4**	1
AE	0	1	0	1	5	1	0	1	0	0
VAE	5	2	**6**	1	6	2	1	0	1	1
K-means	5	1	3	0	6	2	0	1	1	1
Spectral	3	1	4	0	6	2	0	1	2	1
LRAcluster	5	1	3	1	6	1	0	0	0	1
SNF	5	3	6	2	4	1	0	0	4	1
NEMO	5	4	6	2	5	1	1	1	3	1
MCCA	5	4	3	**4**	3	2	1	1	0	1
iCluster	4	1	1	0	4	2	0	1	1	1

we believe that SADLN is competitive with other methods in cancer subtype recognition.

Friedman (1937) analysis was also used to evaluate the performance (Figure 2). From Figure 2, we can see that the performance of SADLN is better than the three methods iCluster, LRAcluster and AE ($p < 0.05$), but not better than other methods. We found that the performance of the methods is not exactly consistent under the two evaluation strategies.

3.5 Comparison of multiple omics data and single omics data

SADLN integrated four types of omics data. To demonstrate the necessity of integrating multiple omics data for subtype recognition, we compared multiple omics data and single omics data of SADLN (denoted as SADLN-single) on subtyping results. We use the random forest (RF) method to analyze the contribution of different omics data on the subtyping results of SADLN. The input of RF is the four original omics features and the subtype labels of SADLN. The output of RF was the Gini importance scores of the features. We perform RF using scikit-learn (1.0.1) package of python, where the key parameter max_depth is set to six and the other parameters are set to the default values. We summed all the Gini importance scores belonging to each type of omics data and quantified the contribution of different omics data to the final subtyping results. The results are shown in Figure 3.

From Figure 3 we can see that the greatest contribution of BRCA, BLCA, LUAD, SKCM, UCEC, and UVM datasets was mRNA data, the greatest contribution of GBM was CNV data and the greatest contribution of KIRC, PAAD, and STAD was DNA methylation data. For different cancers, we choose the greatest contribution of omics data as the input of SADLN-single. The settings of parameters remain the same as SADLN. We also use the metric of p-value of survival analysis in Cox log-rank model to compare the performance of SADLN and SADLN-single (Table 7).

From Table 7, we can see that the p-values of SADLN are all smaller than the values of SADLN-single on ten cancer datasets. These results demonstrated that the integration of multiple omics data can help improve the performance of subtyping.

3.6 Survival analysis and visualization of clustering results

Survival curves can also be used to express the heterogeneity of different subtypes. Figures 4A–J shows the ten cancers' Kaplan Meier survival alanalysis curves. From Figure 4, we can see that different clusters have significantly differences in survival curves (p-value < 0.05). Take BRCA cancer for example (Figure 4A), C1 has the longest average survival time, followed by C5, C2 and C3, C4 has a poor survival time.

To visualize the clustering results, we used the t-SNE embedding method to display the final integrated feature representation of the SADLN (Figure 5). From Figure 5, we can see that samples of the same cluster are almost grouped together, and samples of different clusters are almost departed.

3.7 Case study

In this section, BRCA data is used to analyze the cancer subtypes obtained by the proposed method SADLN. Firstly, we analyzed the overlaps of the identified subtype clusters with the

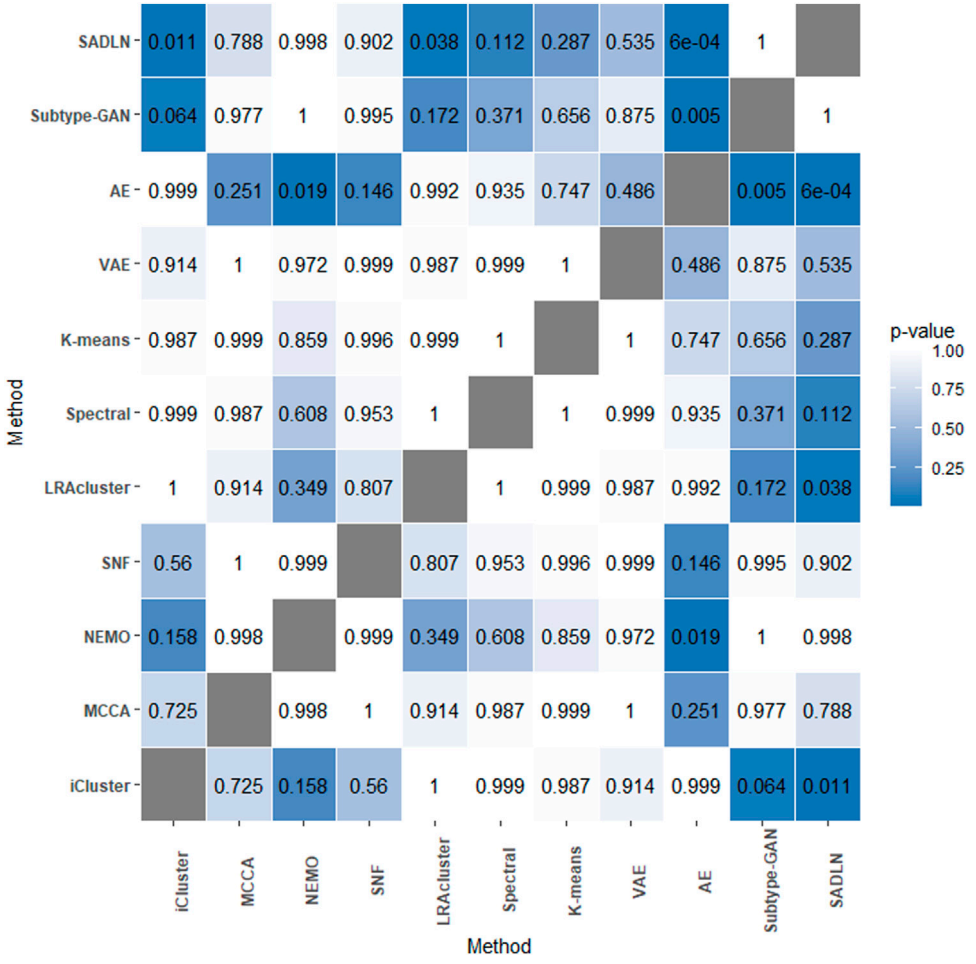


FIGURE 2
The *p*-values of the Friedman test on ten cancer datasets.

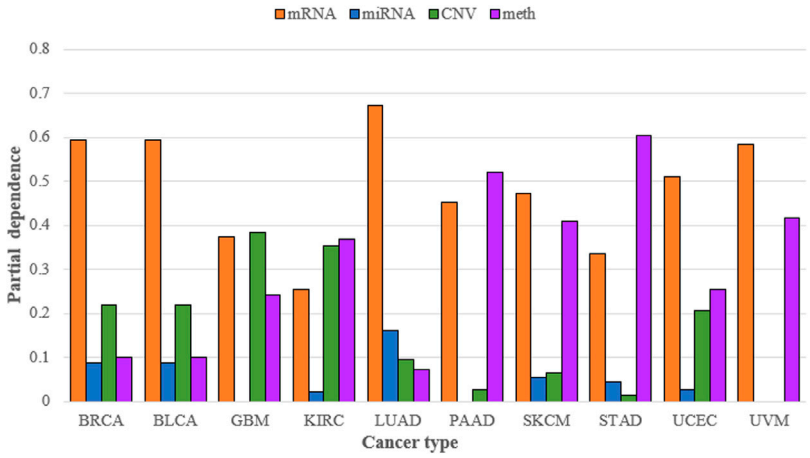


FIGURE 3
Contribution of mRNA, miRNA, CNV, and DNA methylation to the subtyping results of SADLN on ten cancer datasets.

TABLE 7 The *p* values of survival analysis in Cox log-rank model of SADLN based multiple omics data and single omics data (bold indicates that this method performs better on the corresponding cancer dataset).

Cancer	SADLN	SADLN-single
BRCA	**2.40e-03**	4.23e-01
BLCA	**4.39e-03**	2.30e-02
LUAD	**7.69e-04**	3.00e-02
SKCM	**9.22e-04**	2.37e-01
STAD	**4.30e-02**	2.37e-01
UVM	**3.38e-05**	6.74e-01
GBM	**1.77e-02**	1.33e-01
KIRC	**2.77e-07**	1.74e-01
UCEC	**9.52e-05**	1.24e-01
PAAD	**6.39e-04**	1.40e-02

PAM50 cancer subtypes (Parker et al., 2009). There are five PAM50 cancer subtypes (Normal, LumA, LumB, Basal, and Her2), among 1,031 BRCA samples, only 803 samples have PAM50 subtypes including 128 Basal, 66 Her2, 405 LumA, 182 LumB, and 22 Normal. Table 8 shows the results of the overlap. From Table 8, we can see that, cluster C3 is enriched with LumA, of the 252 samples, 223 samples (88.49%) are LumA. Cluster C2 is enriched with LumA and LumB, of the 170 samples, 109 samples (64.12%) are LumA and 54 samples (31.76%) are LumB. Cluster C4 is enriched with LumB and LumA, of the 101 samples, 61 samples (60.40%) are LumB and 30 samples (29.70%) are LumA. Her2 and Basal samples are centrally distributed in clusters C1 and C5.

In order to illustrate the difference between the identified subtype clusters of SADLN, we also analyzed the mutation profiles of BRCA using mutation data (the mutation data can be found at https://portal.gdc.cancer.gov). Among 1,031 samples in BRCA datasets, 820 samples have the mutation data. Figure 6 gives the 20 significantly mutated genes of the identified subtype clusters. From Figure 6, we can see that, clusters C2 and C3 have a significant difference in the frequency of PIK3CA and CDH1 genes, although clusters C2 and C3 are all dominated by LumA subtype. The C1 and C5 clusters have a high frequency of TP53 gene mutations, this also explains why clusters C1 and C5 are dominated by Basal and Her2 subtypes.

To illustrate the difference between clusters C1 and C5, we used RF method to analyzed the differential genes using mRNA expression data. Figure 7 gives the result.

Among these differential expression data, study has shown that the expression of ALDH3B2 was higher in SK-BR-3 cells compared with in other subtypes of breast cell lines, as determined by reverse transcription-polymerase chain reaction and western blot analysis. In addition, the expression levels of ALDH3B2 were higher in Her2 positive breast cancer compared with in other subtypes of breast cancer, as determined by immunohistochemistry, which may be used as a prognostic indicator for breast cancer (Feng et al., 2019). The expression level of CLEC10A to be positively associated with the level of different tumor-infiltrating immune cells in BRCA including CD8 T cells, B cells, macrophages, and NK cells. These results suggest that the relationship between lower CLEC10A expression level and poor prognosis in BRCA may be due to the role of

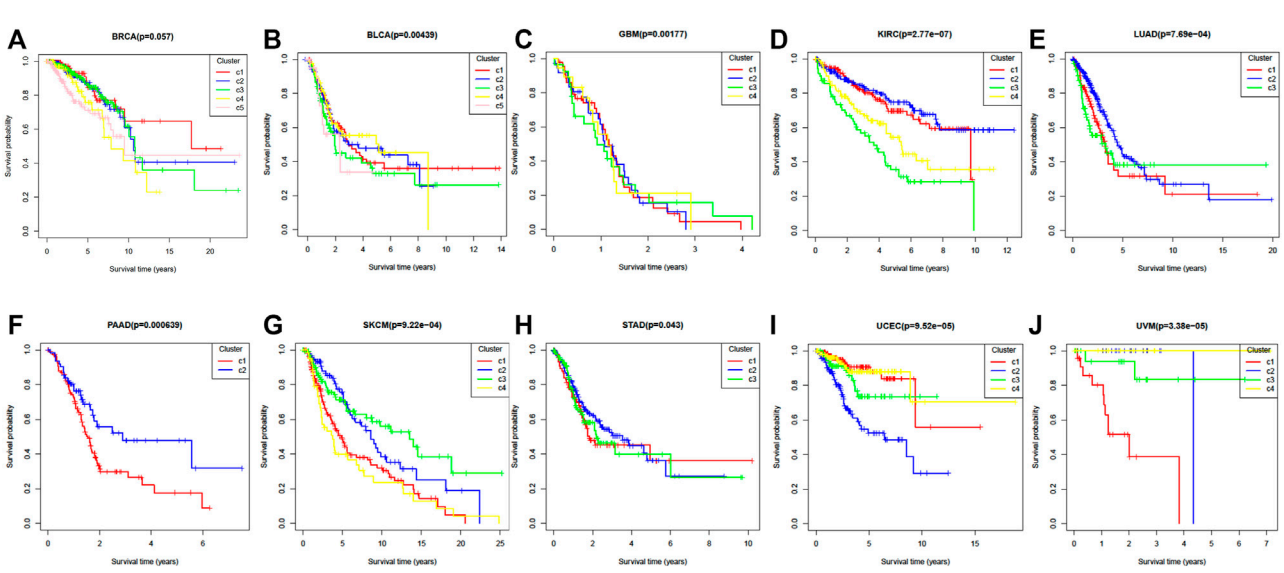


FIGURE 4
The Kaplan–Meier survival curves of ten cancer datasets. **(A)** BRCA, **(B)** BLCA, **(C)** GBM, **(D)** KIRC, **(E)** LUAD, **(F)** PAAD, **(G)** SKCM, **(H)** STAD, **(I)** UCEC, **(J)** UVM.

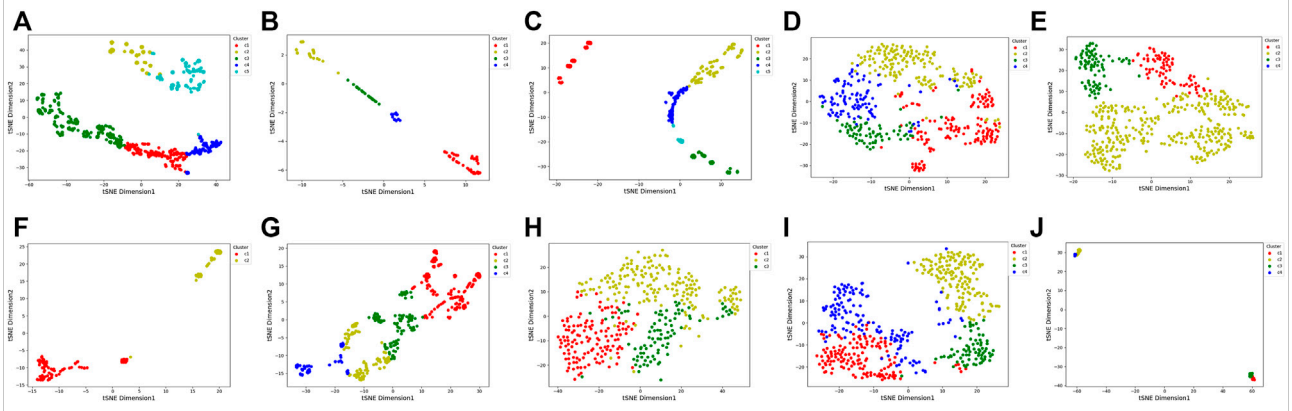


FIGURE 5
t-SNE visualization of the final integrated features by SADLN on ten cancer datasets. **(A)** BRCA, **(B)** BLCA, **(C)** GBM, **(D)** KIRC, **(E)** LUAD, **(F)** PAAD, **(G)** SKCM, **(H)** STAD, **(I)** UCEC, **(J)** UVM.

TABLE 8 The overlaps of the identified subtype clusters with PAM50 subtypes in BRCA cancer datasets.

Subtype ID	C1 (N = 134)	C2 (N = 170)	C3 (N = 252)	C4 (N = 101)	C5 (N = 146)
Basal (128)	60 (44.78%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	68 (46.58%)
Her2(66)	26 (19.40%)	5 (2.94%)	2 (0.79%)	9 (8.91%)	24 (16.44%)
LumA (405)	20 (14.93%)	109 (64.12%)	223 (88.49%)	30 (29.70%)	23 (15.75%)
LumB (182)	20 (14.93%)	54 (31.76%)	17 (6.75%)	61 (60.40%)	30 (20.55%)
Normal (22)	8 (5.97%)	2 (1.18%)	10 (3.97%)	1 (0.99%)	1 (0.68%)

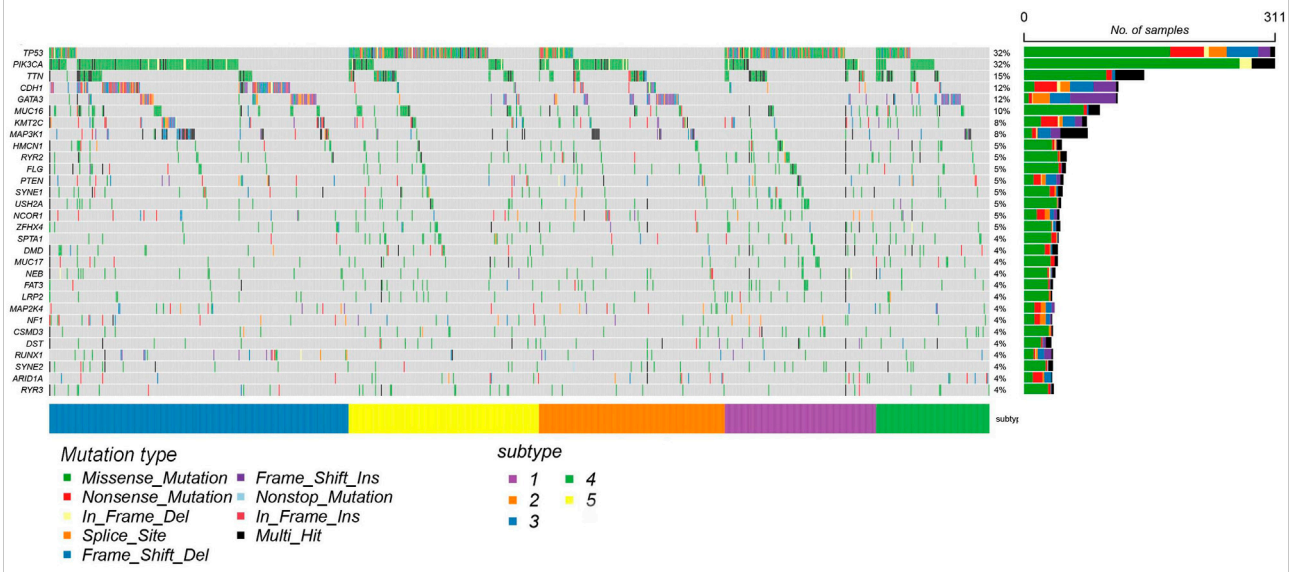


FIGURE 6
The mutation profiles of BRCA datasets with 20 significantly mutated genes using mRNA expression data.

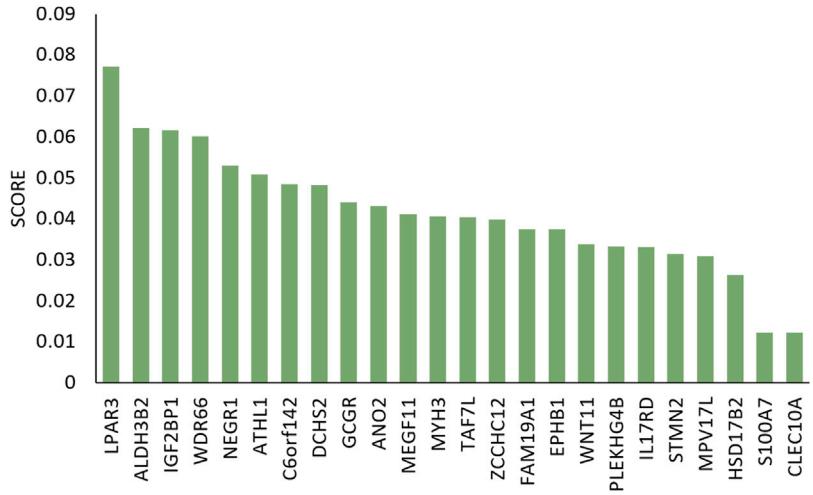


FIGURE 7
Gini importance scores of differential genes in C1 and C5 clusters.

TABLE 9 The five biomarkers most relevant to ten cancers.

Cancers	Biomarkers
BRCA	AGR3, GDF10, EEF1A2, ATP6V0A4, GIPC2
BLCA	GFPT2, SNX31, RASSF9, MUC4, CACNG3
GBM	SHROOM3, PLEKHG4B, CNTNAP4, KCP, PEG10
KIRC	ITPKA, PTPN3, SEMA3B, LRRC55, DNASE1L3
LUAD	HPGDS, UGT1A4, C1orf116, STAT4, ZG16
PAAD	RIMSI, HAL, PAX8, THEM5, EDN2
SKCM	PGLYRP3, TFAP2A, IGSF3, COL17A1, OGDHL
STAD	MEOX2, LIMS2, BEND7, TPSG1, APLN
UCEC	SPDEF, SORBS2, C1orf192, CD163L1, BCL2L14
UVM	PYGM, SCNN1A, SERPINA3, SLC47A22, PRPH

CLEC10A in the tumor immune microenvironment (Tang et al., 2022).

3.8 Identify the key biomarkers in each cancer

To identify the key biomarkers that determine the subtyping results in each cancer, we ranked the importance of mRNA features of each cancer dataset using the clustering labels of SADLN and RF method to achieve the five most essential biomarkers. For each cancer, Table 9 gives the five biomarkers most relevant to ten cancers.

For BRCA as an example, the five key biomarkers are AGR3, GDF10, EEF1A2, ATP6V0A4, and GIPC2. By literature review, we found that the AGR3 gene (de Moraes et al., 2022) affects the prognosis of luminal breast cancer patients. EEF1A2 gene (Hassan et al., 2020) and the GDF10 (Zhou et al., 2019) gene have influenced the prognosis of triple-negative breast cancer patients. The study has shown that the expression of the ATP6V0A4 gene (Savci-Heijink et al., 2019) is a signature of visceral organ metastasis in breast cancer. Although the GIPC2 gene (Dong et al., 2021) has not been found in BRCA but has been shown that it acts on the pathogenesis and development of a pheochromocytoma. All these literature reviews demonstrated the results of SADLN on the BRCA dataset are reliable.

4 Discussion

Recently, integrating multi-omics data for cancer subtyping is an important task in bioinformatics. In this paper, we proposed SADLN, a novel deep learning based integrated method for cancer subtyping. The method firstly introduced self-attention into the encoder-decoder based network architecture. It attempted to describe complex and diverse multi-omics data accurately and adaptively build the samples' relationship when learning a shared low-dimensional representation during molecular subtyping. Compared with three deep learning and seven non-deep learning based integration algorithms, SADLN has two characteristics: 1) Unlike the early integration methods such as AE and VAE, SADLN characterizes multi-omics data respectively which enables the model to effectively describe different omics data with distinct distributions, meanwhile, the output integrating representation fits the prior distribution. 2) The self-attention module in SADLN taking full use of the sample's multi-omics information, can

automatically learn the weight matrix between samples and make the results of feature integration more convincing.

We demonstrated the power of SADLN using ten datasets of TCGA. The experiments of survival analysis and Friedman analysis show that SADLN has a good clustering consequence. Meanwhile, the experiments of SADLN and SADLN-single show that integrating multiple omics data is a necessity and useful. The BRCA results indicated that SADLN can efficiently distinguish cancer subtypes.

SADLN found 50 biomarkers for all cancers. Some biomarkers have been verified in previous studies. In clinical research, researchers can conduct more subtype analysis studies on related cancers based on the biomarkers obtained by SADLN. For example, SADLN believes that MEOX2 is an important biomarker of STAD. The study (Wang et al., 2021) has shown that MEOX2 is a novel biomarker associated with macrophage infiltration in digestive system cancer.

Although SADLN has enhanced the performance of cancer subtyping recognition, it also has limitations. Firstly, it is unsuited to integrate binary data. Secondly, it could not find the genes modules that affect each subtype. Thirdly, the relationship between omics data was not considered. For the next research, we will continue our efforts to develop an attention based method to simultaneously learn the relationship between multi-omic and samples to explore cancer heterogeneity.

5 Conclusion

In this paper, we proposed Self-Attention Based Deep Learning Network (SADLN) for integrating multi-omics data for cancer subtype recognition. The novel method is based on recent advances in deep learning and self-attention. It can jointly learn different multi-omic data representations and relations between samples. In comparison to the state-of-the-art methods, experiments on ten datasets of TCGA have demonstrated the effectiveness of SADLN.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

PG conceived the project. QS and LC designed and implemented the algorithms and models. QS, LC, SG, and AM analyzed and interpreted the data. PG, QS, and LC drafted the manuscript. JC and LZ participated in study analysis. All authors approved the final article.

Funding

This work was supported in part by the Natural Science Foundation of China (No. 82001987), and in part by the Xuzhou Key Research and Development Program Project (No. KC20148).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adossa, N., Khan, S., Rytkönen, K., and Elo, L. (2021). Computational strategies for single-cell multi-omics integration. *Comput. Struct. Biotechnol. J.* 19, 2588–2596. doi:10.1016/j.csbj.2021.04.060

Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., et al. (2015). Genomic classification of cutaneous melanoma. *Cell* 161, 1681–1696. doi:10.1016/j.cell.2015.05.044

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Analysis Mach. Intell.* 39, 2481–2495. doi:10.1109/TPAMI.2016.2644615

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., et al. (2018). A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* 33, 690–705. doi:10.1016/j.ccell.2018.03.014

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clin.* 68, 394–424. doi:10.3322/caac.21492

Chai, H., Zhou, X., Zhang, Z., Rao, J., and Yang, Y. (2021). Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Comput. Biol. Med.* 134, 104481. doi:10.1016/j.compbiomed.2021.104481

Chen, C., Zha, Y., Zhu, D., Ning, K., and Cui, X. (2021). Hydrogen bonds meet self-attention: All you need for general-purpose protein structure embedding. bioRxiv

de Moraes, C. L., e Melo, N. C., Valoyes, M. A. V., and do Amaral, W. N. (2022). Agr2 and agr3 play an important role in the clinical characterization and prognosis of basal like breast cancer. *Clin. Breast Cancer* 22, 1–17. doi:10.1016/j.clbc.2021.07.008

Dong, Y., Huang, Y., Fan, C., Wang, L., Zhang, R., Li, W., et al. (2021). Gipc2 is an endocrine-specific tumor suppressor gene for both sporadic and hereditary tumors of ret-and sdhb-but not vhl-associated clusters of pheochromocytoma/paraganglioma. *Cell death Dis.* 12, 1–17. doi:10.1038/s41419-021-03731-7

Feng, L., Huang, S., An, G., Wang, G., Gu, S., and Zhao, X. (2019). Identification of new cancer stem cell markers and signaling pathways in her-2-positive breast cancer by transcriptome sequencing. *Int. J. Oncol.* 55, 1003–1018. doi:10.3892/ijo.2019.4876

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* 32, 675–701. doi:10.1080/01621459.1937.10503522

Gao, H., Li, Y., Wang, X., Han, J., and Li, R. (2019). "Ensemble attention for text recognition in natural images," in 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, July 14-19, 2019.

Gu, L., Zhang, X., Li, K., and Jia, G. (2020). "Using molecular fingerprints and unsupervised learning algorithms to find simulants of chemical warfare agents," in The 2020 International Seminar on Artificial Intelligence, Networking and Information Technology, Shanghai, China, 18-20 September 2020.

Guo, Y., Shang, X., and Li, Z. (2019). Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer. *Neurocomputing* 324, 20–30. doi:10.1016/j.neucom.2018.03.072

Hong Zhao, W., Luo, J., and Jiao, S. (2014). Comprehensive characterization of cancer subtype associated long non-coding rnas and their clinical implications. *Sci. Rep.* 4, 6591. doi:10.1038/srep06591

Hasin, Y., Seldin, M. M., and Lusis, A. J. (2017). Multi-omics approaches to disease. *Genome Biol.* 18, 83. doi:10.1186/s13059-017-1215-1

Hassan, M. K., Kumar, D., Patel, S. A., and Dixit, M. (2020). Eef1a2 triggers stronger erk mediated metastatic program in er negative breast cancer cells than in er positive cells. *Life Sci.* 262, 118553. doi:10.1016/j.lfs.2020.118553

Hawkins, R., Hon, G. C., and Ren, B. (2010). Next-generation genomics: An integrative approach. *Nat. Rev. Genet.* 11, 476–486. doi:10.1038/nrg2795

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10, 000 tumors from 33 types of cancer. *Cell* 173, 291–304. doi:10.1016/j.cell.2018.03.022

Hou, Y., Ma, Z., Liu, C., and Loy, C. C. (2019). "Learning lightweight lane detection cnns by self attention distillation," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, Oct. 27 2019 to Nov. 2 2019.

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi:10.1038/nrc3721

Le, K., Quan, T. T., Bui, T. H., and Petrucci, L. (2016). "Coca: Congestion-oriented clustering algorithm for wireless sensor networks," in 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN), Beijing, China, June 4-6, 2016, 450–454.

Levine, D. A., Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., et al. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature* 497, 67–73. doi:10.1038/nature12113

Li, L., and Wang, X. (2021). Identification of gastric cancer subtypes based on pathway clustering. *NPJ Precis. Oncol.* 5, 46–17. doi:10.1038/s41698-021-00186-z

Li, M., Wang, Y., Wang, Z., and Zheng, H. (2020a). A deep learning method based on an attention mechanism for wireless network traffic prediction. *Ad Hoc Netw.* 107, 102258. doi:10.1016/j.adhoc.2020.102258

Li, Y., Long, G., Shen, T., Zhou, T., Yao, L., Huo, H., et al. (2020b). Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. *ArXiv* abs/1911.11899

Liu, C., Zhang, L., Niu, J., Yao, R., and Wu, C. (2020). Intelligent prognostics of machining tools based on adaptive variational mode decomposition and deep learning method with attention mechanism. *Neurocomputing* 417, 239–254. doi:10.1016/j.neucom.2020.06.116

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., et al. (2018). An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics* 34, 1381–1388. doi:10.1093/bioinformatics/btx761

Mercer, E., and Neufeld, R. (2021). *Advances in artificial intelligence and security*. Berlin, Germany: Springer Science and Business Media LLC.

Mustafa Abualsaud, M. D. (2019). "Proceedings of the 28th acm international conference on information and knowledge management," in Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing China, November 3 - 7, 2019.

Nguyen, T., Tagett, R., Diaz, D., and Drăghici, S. (2017). A novel approach for data integration and disease subtyping. *Genome Res.* 27 (12), 2025–2039. doi:10.1101/gr.215129.116

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* 27, 1160–1167. doi:10.1200/JCO.2008.18.1370

Parodi, S., Filiberti, R. A., Marroni, P., Libener, R., Ivaldi, G., Mussap, M., et al. (2015). Differential diagnosis of pleural mesothelioma using logic learning machine. *BMC Bioinforma.* 16, S3. S3 – S3. doi:10.1186/1471-2105-16-S9-S3

Peng, J., Li, J., and Shang, X. (2020). A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. *BMC Bioinforma.* 21, 394. doi:10.1186/s12859-020-03677-1

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* 19, 3735–3746. doi:10.1016/j.csbj.2021.06.030

Poirion, O. B., Chaudhary, K., and Garmire, L. (2018). Deep learning data integration for better risk stratification models of bladder cancer. *AMIA Summits Transl. Sci. Proc.* 2018, 197206.

Raphael, B. J., Hruban, R. H., Aguirre, A. J., Moffitt, R. A., Yeh, J. J., Stewart, C., et al. (2017). Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* 32, 185–203. doi:10.1016/j.ccell.2017.07.007

Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Res.* 46, 10546–10562. doi:10.1093/nar/gky889

Rappoport, N., and Shamir, R. (2019). Nemo: Cancer subtyping by integration of partial multi-omic data. *Bioinformatics* 35, 3348–3356. doi:10.1093/bioinformatics/btz058

Robertson, A. G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A. D., et al. (2017a). Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell* 171, 540–556. doi:10.1016/j.cell.2017.09.007

Robertson, A. G., Shih, J., Yau, C., Gibb, E. A., Oba, J., Mungall, K. L., et al. (2017b). Integrative analysis identifies four molecular and clinical subsets in uveal melanoma. *Cancer Cell* 32, 204–220.e15. doi:10.1016/j.ccell.2017.07.003

Rainer and Muche (2001). "Applied survival analysis: Regression modeling of time to event data," in *lemeshow*. Editor Dw hosmer (new york: John wiley).

Savci-Heijink, C., Halfwerk, H., Koster, J., Horlings, H., and Van De Vijver, M. (2019). A specific gene expression signature for visceral organ metastasis in breast cancer. *BMC cancer* 19, 333–338. doi:10.1186/s12885-019-5554-z

Sayáns, M. P., Petronacci, C. C. C., Pouso, A. L. L., Iruegas, E. P., Carrión, A. B., Peñaranda, J. M. S., et al. (2019). Comprehensive genomic review of tcga head and neck squamous cell carcinomas (hnscc). *J. Clin. Med.* 8, 1896. doi:10.3390/jcm8111896

Sharifi-Noghabi, H., Zolotareva, O., Collins, C., and Ester, M. (2019). Moli: Multi-omics late integration with deep neural networks for drug response prediction. bioRxiv

Shaw, P., Uszkoreit, J., and Vaswani, A. (2018). Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25 22, 2906–2912. doi:10.1093/bioinformatics/btp543

Siegel, R., Miller, K., and Jemal, A. (2020). Cancer statistics, 2020. *CA A Cancer J. Clin.* 70, 7–30. doi:10.3322/caac.21590

Simidjievski, N., Bodnar, C., Tariq, I., Scherer, P., Andrés-Terré, H., Shams, Z., et al. (2019). Variational autoencoders for cancer data integration: Design principles and computational practice. *bioRxiv* 10, 1205. doi:10.3389/fgene.2019.01205

Song, H., Ruan, C., Xu, Y., Xu, T., Fan, R., Jiang, T., et al. (2021). Survival stratification for colorectal cancer via multi-omics integration using an autoencoder-based model. *Exp. Biol. Med.* 247, 898–909. doi:10.1177/15353702211065010

Song, M., Greenbaum, J., Luttrell, J., Zhou, W., Wu, C., Shen, H., et al. (2020). A review of integrative imputation for multi-omics datasets. *Front. Genet.* 11, 570255. doi:10.3389/fgene.2020.570255

Tang, S., Zhang, Y., Lin, X., Wang, H., Yong, L., Zhang, H., et al. (2022). Clec10a can serve as a potential therapeutic target and its level correlates with immune infiltration in breast cancer. *Oncol. Lett.* 24, 285–311. doi:10.3892/ol.2022.13405

The Cancer Genome Atlas Research Network (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499, 43–49. doi:10.1038/nature12222

The Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511, 543–550. doi:10.1038/nature13385

Tong, L., Mitchel, J., Chatlin, K., and Wang, M. D. (2020). Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Med. Inf. Decis. Mak.* 20, 225. doi:10.1186/s12911-020-01225-8

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell* 17, 98–110. doi:10.1016/j.ccr.2009.12.020

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi:10.1038/nmeth.2810

Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., et al. (2020). Moronet: Multi-omics integration via graph convolutional networks for biomedical data classification. bioRxiv.

Wang, Z., Yang, H., Zhang, R., Luo, B., Xu, B., Zhu, Z., et al. (2021). Meox2 serves as a novel biomarker associated with macrophage infiltration in oesophageal squamous cell carcinoma and other digestive system carcinomas. *Autoimmunity* 54, 373–383. doi:10.1080/08916934.2021.1919880

Witten, D. M., and Tibshirani, R. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.* 8, 28–27. doi:10.2202/1544-6115.1470

Xu, J., Wu, P., Chen, Y., Meng, Q., Dawood, H., and Dawood, H. (2019). A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. *BMC Bioinforma.* 20, 527. doi:10.1186/s12859-019-3116-7

Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., et al. (2016). Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE Trans. Med. Imaging* 35, 119–130. doi:10.1109/TMI.2015.2458702

Yang, H., Chen, R., Li, D., and Wang, Z. (2021a). Subtype-gan: A deep learning approach for integrative cancer subtyping of multi-omics data. *Bioinformatics* 37, 2231–2237. doi:10.1093/bioinformatics/btab109

Yang, H., Wang, M., Liu, X., Zhao, X., and Li, A. (2021b). Phosidn: An integrated deep neural network for improving protein phosphorylation site prediction by combining sequence and protein–protein interaction information. *Bioinformatics* 37, 4668–4676. doi:10.1093/bioinformatics/btab551

Yu, H., Li, J., Zhang, L., Cao, Y., Yu, X., and Sun, J. (2021). Design of lung nodules segmentation and recognition algorithm based on deep learning. *BMC Bioinforma.* 22, 314. doi:10.1186/s12859-021-04234-0

Yuan, S., Zhang, Y., Tang, J., Shen, H., and Wei, X. (2018). Modeling and predicting popularity dynamics via deep learning attention mechanism. *ArXiv* abs/1811.02117

Zeng, Z., Mao, C., Vo, A. H., Li, X., Nugent, J. O., Khan, S. A., et al. (2021). Deep learning for cancer type classification and driver gene identification. *BMC Bioinforma.* 22, 491. doi:10.1186/s12859-021-04400-4

Zhang, L., Yang, X., Li, S., Liao, T., and Pan, G. (2022). Answering medical questions in Chinese using automatically mined knowledge and deep neural networks: An end-to-end solution. *BMC Bioinforma.* 23, 136. doi:10.1186/s12859-022-04658-2

Zhang, Z., Wu, S., Chen, G., and Jiang, D. (2020). Self-attention and dynamic convolution hybrid model for neural machine translation. *IEEE Int. Conf. Knowl. Graph (ICKG)* 2020, 352–359. doi:10.1109/ICBK50248.2020.00057

Zhou, T., Yu, L., Huang, J., Zhao, X., Li, Y., Hu, Y., et al. (2019). Gdf10 inhibits proliferation and epithelial-mesenchymal transition in triple-negative breast cancer via upregulation of smad7. *Aging (Albany NY)* 11, 3298–3314. doi:10.18632/aging.101983

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/fgene.2022.1090394

OPEN ACCESS

EDITED BY
Dominik Heider,
University of Marburg, Germany

REVIEWED BY
Hans A. Kestler,
University of Ulm, Germany
Andreas Holzinger,
Medical University Graz, Austria

*CORRESPONDENCE
Jili Hu,
hujili@ahtcm.edu.cn

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 05 November 2022
ACCEPTED 09 December 2022
PUBLISHED 04 January 2023

CITATION
Liu C, Duan Y, Zhou Q, Wang Y, Gao Y, Kan H and Hu J (2023), A classification method of gastric cancer subtype based on residual graph convolution network. *Front. Genet.* 13:1090394. doi: 10.3389/fgene.2022.1090394

A classification method of gastric cancer subtype based on residual graph convolution network

Can Liu[1,2], Yuchen Duan[1], Qingqing Zhou[1], Yongkang Wang[1,2], Yong Gao[1,2], Hongxing Kan[1,2] and Jili Hu[1,2]*

[1]School of Medical Informatics Engineering, Anhui University of Chinese Medicine, Hefei, Anhui, China, [2]Anhui Computer Application Research Institute of Chinese Medicine, China Academy of Chinese Medical Sciences, Hefei, Anhui, China

Background: Clinical diagnosis and treatment of tumors are greatly complicated by their heterogeneity, and the subtype classification of cancer frequently plays a significant role in the subsequent treatment of tumors. Presently, the majority of studies rely far too heavily on gene expression data, omitting the enormous power of multi-omics fusion data and the potential for patient similarities.

Method: In this study, we created a gastric cancer subtype classification model called RRGCN based on residual graph convolutional network (GCN) using multi-omics fusion data and patient similarity network. Given the multi-omics data's high dimensionality, we built an artificial neural network Autoencoder (AE) to reduce the dimensionality of the data and extract hidden layer features. The model is then built using the feature data. In addition, we computed the correlation between patients using the Pearson correlation coefficient, and this relationship between patients forms the edge of the graph structure. Four graph convolutional network layers and two residual networks with skip connections make up RRGCN, which reduces the amount of information lost during transmission between layers and prevents model degradation.

Results: The results show that RRGCN significantly outperforms other classification methods with an accuracy as high as 0.87 when compared to four other traditional machine learning methods and deep learning models.

Conclusion: In terms of subtype classification, RRGCN excels in all areas and has the potential to offer fresh perspectives on disease mechanisms and disease progression. It has the potential to be used for a broader range of disorders and to aid in clinical diagnosis.

KEYWORDS

multi-omics, autoencoder, patient similarity network, residual graph convolutional network, classification

1 Introduction

Gastric cancer (GC) is a highly aggressive cancer with significant heterogeneity in terms of cell types, states, and subpopulation distribution in the immune microenvironment (Shao et al., 2021; Kim et al., 2022). According to the epidemiological survey (Ferlay et al., 2021), the incidence of GC is the fifth highest among tumor diseases worldwide, and the mortality rate is the third highest among tumor deaths (Wang et al., 2021; Dong et al., 2021). Studies have shown that several variables, including genetics, the immune system, lifestyle choices, and psychological factors, can affect the development and occurrence of tumors (Shin et al., 2022). Multiple pathological processes at various levels and dimensions, including the genome, transcriptome, and proteome, are involved in complex diseases like cancer (Menyhárt and Győrffy, 2021).

With the advancement of high-throughput sequencing and omics technology, researchers progressively understood the limits of employing a single omics (Sun et al., 2019; Jia et al., 2022). To better understand the essence of the disease, it is required to undertake a joint analysis of various types of data, get more comprehensive information, construct a perfect body regulatory network, and thoroughly investigate the regulation and causal relationships between molecules (Tao et al., 2020). Consequently, one of the areas of research that is now quite active is the integration of multi-omics data for cancer subtyping (Lindskrog et al., 2021; Sivadas et al., 2022). The biological information contained in multi-omics data is critical for disease diagnosis and treatment. However, due to its huge scale, high dimension, high noise, and strong heterogeneity, data is difficult to handle and analyze, posing significant obstacles to cancer typing (Duan et al., 2021; Picard et al., 2021).

The Graph Convolutional Network (GCN) (Kipf and Welling, 2017) is a convolutional neural network that was built in recent years that can directly act on graphs and use their structural information, and it is gaining popularity in the field of bioinformatics (Zhang et al., 2021). It can identify unlabeled nodes and categorize them using both the node's feature vector and network topology data (Li et al., 2022).

Kim et al., (2021) proposes an analytical framework named DrugGCN based on gene expression data for predicting drug responses using graph convolutional networks (GCNs). Baul et al., (2022) offers omicsGAT, a graph attention network (GAT) model that blends graph learning with attention processes for cancer subtype identification based on RNA-seq data. By allocating various attention coefficients to nearby samples, the multi-head attention mechanism can more successfully protect the connection between them. However, such experimental results are neither applicable nor interpretable when only one set of omic data is considered. According to studies (Sun and Hu, 2016; Xu et al., 2019), different forms of data have complementarities, and multi-omics can fuse the rich information in each type of data to facilitate categorization. Li et al., (2022) developed a multi-omics ensemble model, MoGCN, with two-layer graph convolutional networks for the classification and analysis of cancer subtypes. Ramirez et al., (2020) constructed a graph convolutional neural network for classifying tumor and non-tumor samples based on unstructured gene expression data. Unfortunately, as depth increases, graph convolutional networks suffer from vanishing gradients and over-smoothing, which significantly reduces model accuracy. Zhang et al., (2022) proposes a new method for detecting liver cancer using a fusion similarity network, denoising autoencoder, and dense graph convolutional neural network. Liang et al., (2021) proposed a Consensus Guided Graph Autoencoder (CGGA) to identify cancer subtypes and bring fresh insights into the treatment of patients with diverse subtypes. Wang et al., (2021) introduces a unique multi-omics integrative approach called the Multi-Omics Graph Convolutional Networks (MOGONET), which is utilized for biomedical classification and can find key biomarkers from various omics data sources. Finally, Dai et al., (2021) combined GCN with a residual network to build a cancer subtype classification model, named ERGCN, which performed well on three different TCGA cancer types, presenting a new method for precision cancer treatment.

Therefore, we integrated multi-omics data and designed a model RRGCN based on graph convolution for GC subtype classification. High-dimensional multi-omics data is integrated into low-dimensional space using an artificial neural network autoencoder (AE) to extract hidden layer characteristics. The Patient Similarity Network (PSN) combines the network topology generated by each data type and analyzes the links between patients using the Pearson correlation coefficient (Benesty et al., 2009). The fused network can collect information from multiple data sources that are both shared and complementary. Two residual networks with skip connections are merged with four GCN layers to collect feature matrices and patient similarity correlations to discover and classify GC subtypes, and the classification results are finally output by softmax. The results of the comparison with random forest (RF), support vector machine (SVM), MoGCN, and ERGCN reveal that RRGCN has the best performance. The classification accuracy of the GC subtype is 0.87, the AUC value is 0.98, and the values of other indicators of RRGCN are also the highest when compared to other methods. We believe that RRGCN can provide new and unique insights into the identification, classification, and clinical diagnosis of GC subtypes.

2 Materials and methods

Proposed method

We designed a GC subtype classification model, namely RRGCN, which is based on the residual graph convolutional

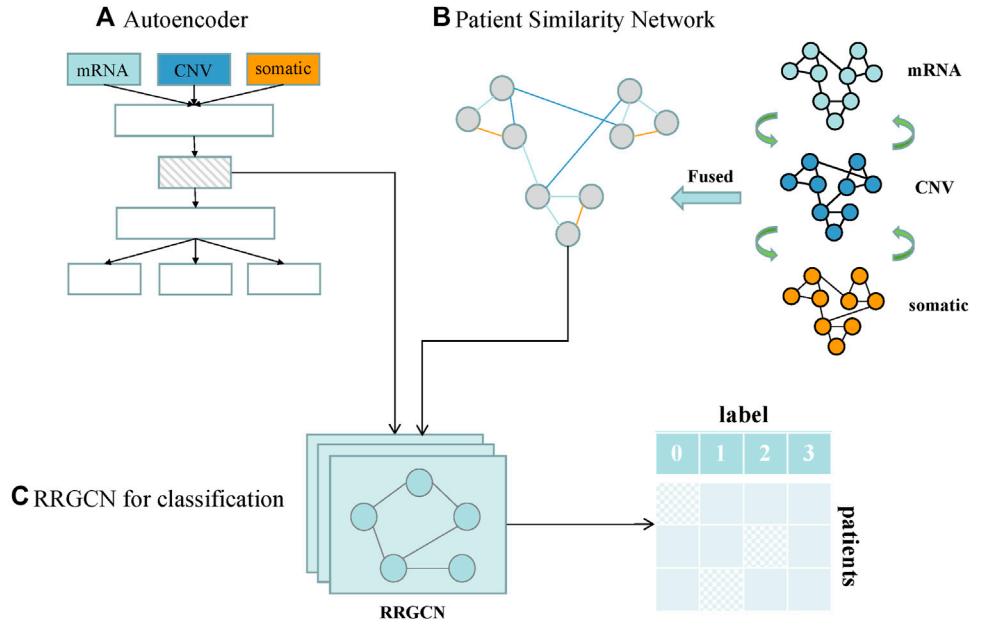


FIGURE 1
Workflow of RRGCN. **(A)** Features extracted by AE network. **(B)** Patient similarity network constructed by SNF algorithm. **(C)** RRGCN for GC subtype classification.

network. The input consists of the multi-omics fusion data and the patient similarity network following AE dimensionality reduction. The graph nodes are then embedded through two residual networks with skip connections and a 4-layer GCN, and the classification results are then output using a softmax layer. We compared and assessed RRGCN's performance with several traditional machine learning models and deep learning methods in the third chapter of the paper. Figure 1 shows the workflow of RRGCN.

Datasets and data preprocessing

To train the model, we used information on GC from the TCGA (https://tcga-data.nci.nih.gov/tcga/). Transcriptomic data, copy number variations (CNV), and somatic mutation data are all included in our study. We got 272 labeled samples and four subtypes of data from the R tool "TCGAbiolinks" (Colaprico et al., 2016). We download the experimental data using the R package "TCGA-assembler 2" (Wei et al., 2018). The transcriptome data is from the Illumina HiSeq_RNASeqV2 sequencing platform, the CNV data is from the cna_cnv.hg19 sequencing platform and the somatic mutation data is from the somaticMutation_DNAseq sequencing platform. In addition, to make it easier for the model to categorize the input data, we define four GC subtypes as numbers, Epstein-Barr virus type (EBV) as 0, Microsatellite instability type (MSI) as 1, Genetically stable type (GS) as 2, and Chromosome instability type (CIN) as 3.

The dataset in TCGA has to be preprocessed because it contains a large amount of zero and missing value data. The preprocessing step helps to reduce the redundancy and inconsistency of the dataset, thus improving the accuracy and speed of the subsequent mining process. From the phenotypic data, sample information with labels for the various cancer subtypes was first retrieved, and the features that were absent from all samples or had a zero-expression level were subsequently eliminated. So we ended up with 272 samples. Second, among the genes that have been duplicated, we

TABLE 1 Overview of the STAD dataset.

Multi-omics	Number of features	Subtypes	Samples
mRNA	20,468	EBV	25
CNV	22,434	MSI	60
Somatic	19,600	GS	51
—	—	CIN	136
Total	62,502	Total	272

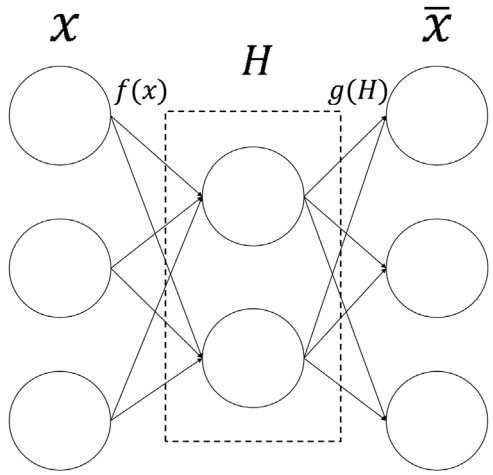


FIGURE 2
The structure of Autoencoder. The input data is first encoded by the encoder, then embedded in the hidden layer, and finally output by the decoder.

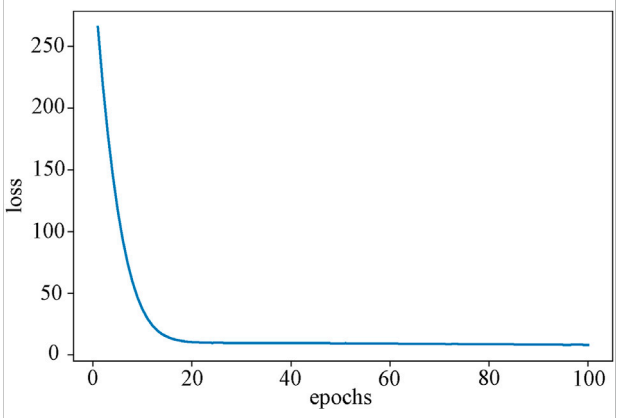


FIGURE 3
The loss curve of the AE training process. As the epoch grows, training tends to converge around epoch 20.

choose the one whose mean expression across all samples has the least absolute value. Finally, for the transcriptome data, we expressed expression levels in units of log2 (FPKM + .1), where FPKM stands for Fragments Per Kilobase of Exon Model per Million mapped Fragment. In this study, we removed the number of zero values and missing values in mRNA, CNV and somatic cells to be 62, 2,481 and 2 respectively, resulting in 20,468, 22,434 and 19,600 features for subsequent model construction. Table 1 shows the details of the dataset.

Autoencoder architecture

The autoencoder (AE) (Hinton and Salakhutdinov, 2006) is an unsupervised artificial neural network model that belongs to the deep learning category. AE can extract latent embedding representations from multi-omics datasets to reduce dimensionality and computational cost. It can first learn the hidden features of the input data through encoding, then output to the next hidden layer, and then decode and rebuild the original input data with the learned new features (Binbusayyis and Vaiyapuri, 2021). Figure 2 is the basic framework of Autoencoder. The formula is:

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{\delta}(\boldsymbol{\omega x} + \boldsymbol{b}) = \boldsymbol{H} \tag{1}$$

$$\boldsymbol{g}(\boldsymbol{H}) = \boldsymbol{\delta}(\boldsymbol{\omega}'\boldsymbol{H} + \boldsymbol{b}') = \bar{\boldsymbol{x}} \tag{2}$$

Where x is the input feature in the AE, which is encoded and decoded to $\bar{x}$. $f(x)$ represents the encoder function, H represents the hidden unit, $g(H)$ represents the decoder function, $\bar{x}$ represents the output, δ represents the activation function, ω represents the weight matrix, b represents the bias. We used the mean square error (MSE) (Sammut and Webb, 2010) as the loss function to calculate the loss between the predicted value and the true value, where the predicted value is $\bar{x}$ and the true value is x. The formula is:

$$\boldsymbol{mseloss}(\boldsymbol{x}, \bar{\boldsymbol{x}}) = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{i}})^{\boldsymbol{2}} \tag{3}$$

Since RRGCN uses three different forms of data, we gave each omics data a varied weight based on prior knowledge (Li et al., 2022) to emphasize their contributions to the model, and all weights sum up to 1. In light of this, the loss function is described as:

$$\boldsymbol{L_{AE}} = \boldsymbol{a}*\boldsymbol{mseloss}(\boldsymbol{x_1}, \bar{\boldsymbol{x}}_{\boldsymbol{1}}) + \boldsymbol{b}*\boldsymbol{mseloss}(\boldsymbol{x_2}, \bar{\boldsymbol{x}}_{\boldsymbol{2}}) + \boldsymbol{c}*\boldsymbol{mseloss}(\boldsymbol{x_3}, \bar{\boldsymbol{x}}_{\boldsymbol{3}}) \tag{4}$$

L_{AE} represents the MSE loss function, and a, b, and c represent the weights of the input data, respectively for 0.4, 0.3, and 0.3. As the input data are characterized by multi-omics data types and represented by multiple matrices x_1, x_2, and x_3, corresponding to the mRNA, CNV, and somatic matrices, $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$ correspond to the output of three types of data.

In this study, we took into account high-dimensional multi-omics data using an AE with three hidden layers. The three hidden layers were (500, 200, 500), and the training epoch was set to 100, which ultimately converged after 20 epochs (Figure 3). All layers employ the sigmoid function as their activation function. AE is trained by back-propagation through the Adam (Kingma and Ba, 2015) optimizer. Additionally, we used grid search to select the batch size from (32, 64, 128) and the learning rate (LR) from (0.01, 0.001, 0.0001). The final batch size is 32, and

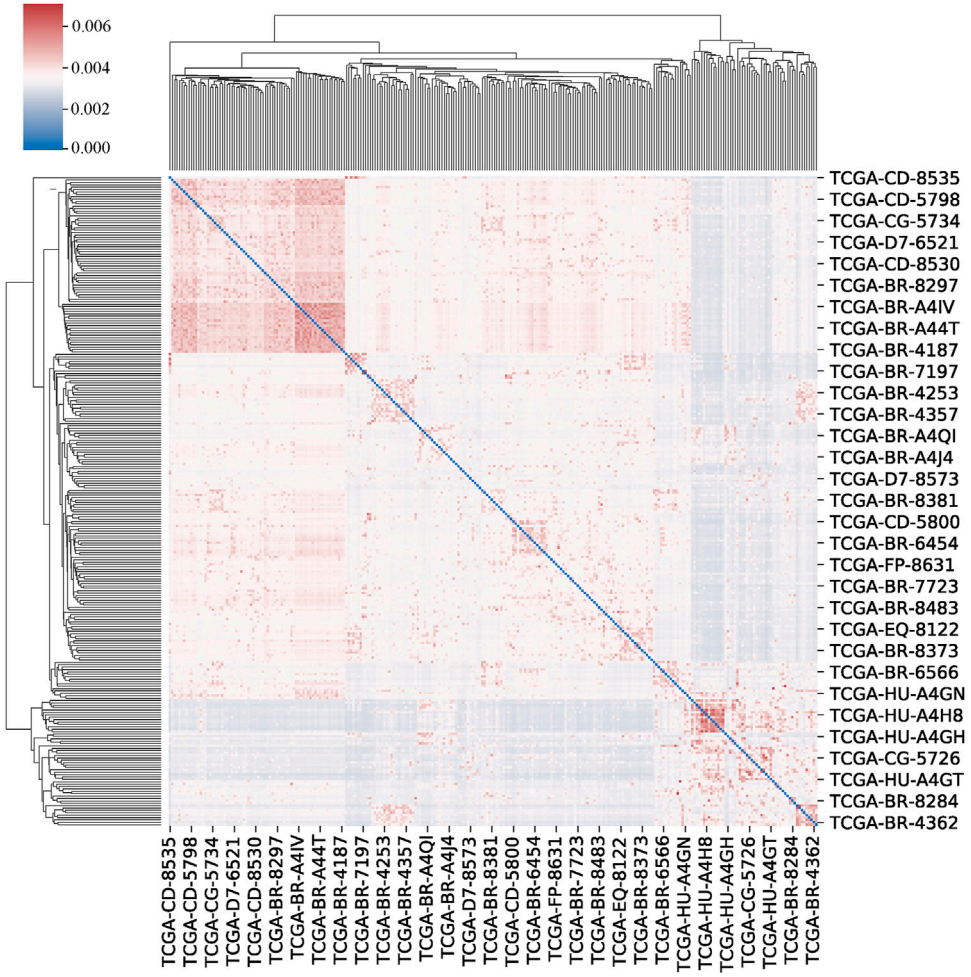


FIGURE 4
The clustermap of patients. The clustermap shows the clustering relationship among multiple samples, with red representing high correlation and blue representing low correlation.

LR is 0.001. Every model used in our study is built by PyTorch (v1.8.0) (Paszke et al., 2019). The feature matrix extracted by the AE hidden layer will be used as the input of the RRGCN.

Patient similarity network

The Similarity Network Fusion (SNF) (Wang et al., 2014) algorithm is a computational approach that creates a network of similarities across patients for each type of data to provide a holistic perspective of a certain disease or biological process. We used the SNF algorithm to compute and fuse patient similarity networks from each data type in the GC dataset to create an overall view of GC patients. The advantage of PSN is that it enables RRGCN to seek and obtain important information from the neighbour nodes of the patient, rather than relying solely on the level of gene expression. This improves the accuracy and applicability of the model. The SNF algorithm creates patient-patient similarity matrices for each data type and construct the patient adjacency matrix, then builds a network through the matrix, and lastly fuses various forms of patient-patient similarity networks to create a fusion network. SNF can fully exploit the complementarity of various source data (El-Manzalawy et al., 2018; Picard et al., 2021), which is far superior to the comprehensive analysis approach established by employing a single dataset and has significant advantages in the detection and classification of cancer subtypes (Wang T.-H. et al., 2021; Franco et al., 2021).

Assume there are n samples and m various categories of data (in this study, the data types include mRNA, CNV, and

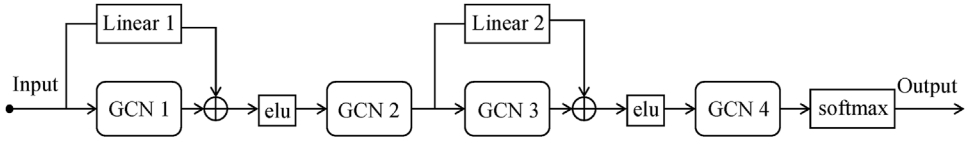


FIGURE 5
The structure of RRGCN.

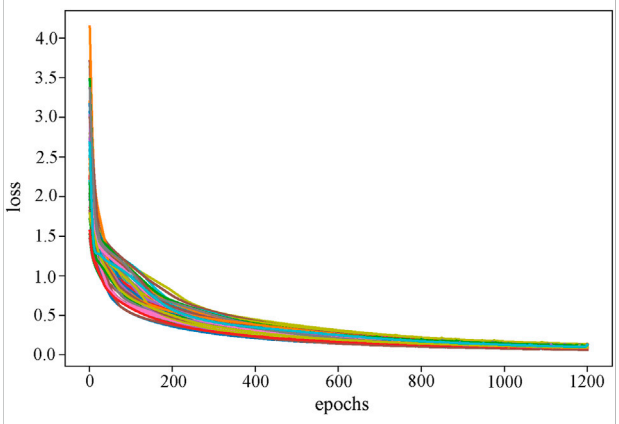


FIGURE 6
The loss curve of the RRGCN training process. Different colors reflect the loss curves of different cross-validation times.

TABLE 2 Confusion matrix.

Predicted	Actual	
	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

somatic data). We refer to a PSN as a graph $G = (V, E)$, where the vertex V is a collection of samples made up of $(x_1, x_2, \cdots x_n)$, and E makes up the edges of the graph. A similarity matrix defined by the scaled exponential similarity kernel was computed:

$$\boldsymbol{w(i,j) = \exp\left(-\frac{\rho^2(x_i, x_j)}{\mu \varepsilon_{i,j}}\right)} \tag{5}$$

Among them, w represents the similarity matrix between samples, $\rho(x_i, x_j)$ represents the Euclidean distance between the patient x_i and patient x_j, μ is a hyperparameter set by experience, and the commonly used range is (0.3, 0.8), and $\varepsilon_{i,j}$ is a parameter used to eliminate the scaling problem, which is defined as:

$$\boldsymbol{\varepsilon_{i,j} = \frac{mean(\rho(x_i, N_i)) + mean(\rho(x_j, N_j)) + \rho(x_i, x_j)}{3}} \tag{6}$$

where N_i is the set of x_i's neighbors and $mean(\rho(x_i, N_i))$ is the mean distance from x_i to each neighbor. Thus, to compute fusion matrices from multiple data types, the similarity matrix is defined as:

$$\boldsymbol{P_{i,j} = \begin{cases} \frac{W_{i,j}}{2\sum_{k \neq i} W_{i,k}}, & j \neq i \\ \frac{1}{2}, & j = i \end{cases}} \tag{7}$$

Then, the affinity matrix S is calculated:

$$\boldsymbol{S_{i,j} = \begin{cases} \frac{W_{i,j}}{\sum_{k \in N_i} W_{i,k}}, & j \in N_i \\ 0, & otherwise \end{cases}} \tag{8}$$

In the case of various data types:

$$\boldsymbol{P^{(v)} = S^{(v)} \left(\frac{\sum_{k \neq v} P^{(k)}}{m-1}\right) \left(S^{(v)}\right)^T, v = 1, 2, \cdots, m} \tag{9}$$

where the $S^{(v)}$ represents the affinity matrix of vth type of data, the $P^{(v)}$ represents the similarity matrix of vth type data. The Pearson correlation coefficient is used to calculate the correlation (linear correlation) between two variables and has a value between −1 and 1. We determined how similar patients were to one another using the Pearson correlation coefficient, and if their similarity exceeded a predetermined threshold, we categorized this as a correlation between patients. The patient similarity network established by the merging of many types of data (multi-omics) is finally obtained by the continual update and iteration of the preceding algorithm.

We set the number of neighbours to consider when creating the affinity matrix to 20 and the scaling factor to 0.5. The clustermap of patients is shown in Figure 4.

Construction of RRGCN

We use GCN to process non-Euclidean data computed using the SNF algorithm. The purpose of GCN is to learn latent

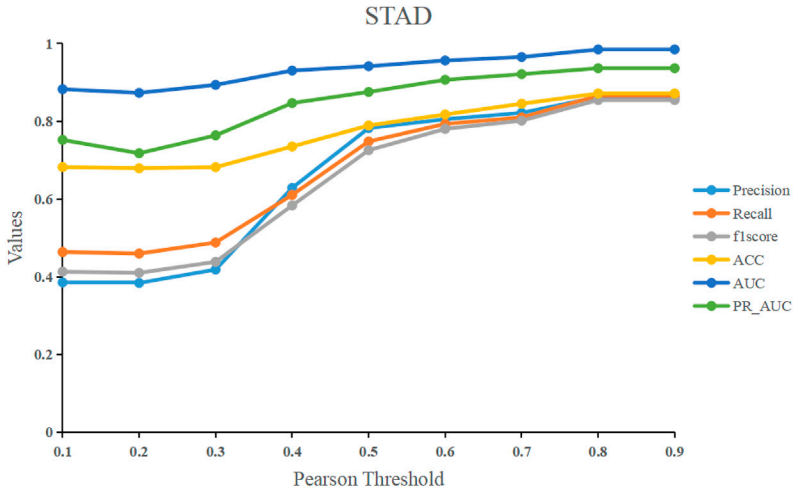


FIGURE 7
The effect of different Pearson correlation thresholds on model performance.

representations based on the node feature matrix X (input, graph nodes) and the similarity matrix A (similarities between nodes). Mathematically, the propagation formula between GCN layers is:

$$H^L = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{L-1}W^{L-1}\right) \quad (10)$$

H^L represents the output of the Lth layer, that is, the node features learned by the Lth layer, W^{L-1} represents the weight matrix of the L-1-th layer, and σ represents the non-linear activation function in the GCN. D is the degree matrix of A, $\tilde{A} = A + E$, E represents the identity matrix.

We use the ResNet (He et al., 2016) concept and add skip connections between GCN layers to overcome the problem of model degradation in deep neural network training. The insertion of skip connections can compensate for the loss of features between the data of the previous layer and the data of the following layer, reducing information loss and improving model performance (Yamanaka et al., 2017). At the same time, to avoid the inconsistency between the output of GCN and the dimension of the input data, we add an independent linear layer to the skip connection. The formula for skip connection can be defined as:

$$H^{L+1} = elu\left(H^L + linear\left(X\right)\right) \quad (11)$$

H^L represents the output of the previous GCN layer. The input feature matrix is sent to the linear layer, and the result is added to the GCN layer and then passed to the non-linear activation function Exponential Linear Units (ELU) (Clevert et al., 2016) to generate the output H^{L+1}, which is utilized as the input of the next skip connection.

RRGCN, which has more skip connections than ERGCN, which only has one, improves model performance by increasing the information flow between layers, making up for information loss, increasing the connectivity between the upper and lower information, and improving the flow of information between layers. The RRGCN as a whole consists of 2 residual networks with skip connections, 4 GCN layers, and 1 softmax layer for generating classification results. To compute the difference between the classification results and the true labels, we utilize the cross-entropy loss function:

$$L = -\left[y \cdot log\left(\bar{y}\right) + \left(1 - y\right) \cdot \log\left(1 - \bar{y}\right)\right] \quad (12)$$

where y represents the true label corresponding to the sample, and $\bar{y}$ is the probability value output by the softmax layer. The structure of RRGCN is shown in Figure 5.

We set the dimensions of the four graph convolutional layers to 64, 32, 16, and the number of subtypes, respectively. By performing grid search on the LR and weight decay in (0.1, 0.01, 0.001, 0.0001) and (0.1, 0.01, 0.001), respectively, the optimal LR and weight decay are determined to be 0.0001 and 0.01. We use the Adam optimizer function and set the epoch to 1,200, the training process finally converges at 600 (Figure 6). RRGCN employs ELU as the non-linear activation function, and the

TABLE 3 Results of multi-omics data compared with single-dimensional data.

Omics	Accuracy	AUC
mRNA	0.7384	0.9339
CNV	0.4766	0.7809
Somatic	0.5190	0.8272
Multi-omics	**0.8713**	**0.9848**

Bold values emphasize that the experimental results of multi-omics are better than other groupings.

TABLE 4 Results in comparison to other methods.

Model	Accuracy	F1 score	Precision	Recall
RF	0.8363	0.7665	0.8172	0.7471
SVM	0.7455	0.7340	0.7792	0.7460
MoGCN	0.7944	0.8078	0.8034	0.7407
ERGCN	0.7901	0.6826	0.7160	0.6120
RRGCN	**0.8713**	**0.8544**	**0.8621**	**0.8654**

Bold values are to highlight the performance of our model over other classical models.

classification results are finally output *via* the softmax layer. We used 80% of the multi-omics fusion data as the training set and reserved 20% for validation. Model performance was evaluated using 5-fold cross-validation on the training set. Furthermore, to eliminate the bias introduced by a single trial, we took the average outcome of ten iterations of the 5-fold cross-validation test set as the evaluation metric.

Model evaluation metrics

In the classification task, the model produces four main prediction results: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). The confusion matrix in Table 2 can be constructed based on the four different prediction outcomes.

Precision refers to the probability that the prediction is correct in the sample that is predicted to be true. It is defined as:

$$\boldsymbol{precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{13}$$

Recall, also known as sensitivity, is the measure of how many samples are selected as being true. It is defined as:

$$\boldsymbol{recall} = \boldsymbol{sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{P}} \tag{14}$$

The F1 score is a weighted harmonic average of precision and recall that is unaffected by imbalanced samples. The F1 score has a maximum value of one and a minimum value of zero. The higher the value, the higher the model quality. In most circumstances, the f1 score can be used directly to evaluate and pick the model, and some well-known machine learning competitions do as well. It is defined as:

$$\boldsymbol{F1\ score} = \frac{\boldsymbol{2{*}precision{*}recall}}{\boldsymbol{precision} + \boldsymbol{recall}} \tag{15}$$

Accuracy is defined as the ratio of accurately predicted samples to total samples. It is defined as:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}} \tag{16}$$

The area contained by the curve with the false positive rate (FPR) on the abscissa and the true positive rate (TPR) on the ordinate is known as the area under the receiver operating characteristic curve (ROC) curve (AUC). The categorization skill given by the ROC curve is intuitively reflected by AUC. The AUC value ranges between 0 and 1, and the higher the value, the better the classifier's performance. FPR is the likelihood that the prediction is a positive sample but the prediction is incorrect. It is defined as:

$$\boldsymbol{FPR} = \frac{\boldsymbol{FP}}{\boldsymbol{TN} + \boldsymbol{FP}} \tag{17}$$

TPR reflects the likelihood that the forecast is a positive sample and that the prediction is right. It is defined as:

$$\boldsymbol{TPR} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{18}$$

The area contained by the curve with recall on the abscissa and precision on the ordinate is known as PR-AUC, and it is the mean value of precision calculated for each recall threshold (Géron, 2017). All model evaluation metrics are based on Scikit-learn (Pedregosa et al., 2011).

3 Results

Determination of pearson correlation threshold

We used the Pearson correlation threshold to see if there was a link between samples. If the Pearson correlation coefficient between samples is larger than the threshold, we connect the two samples with an edge and set the corresponding value in the adjacency matrix to 1. In contrast, there is no edge connecting the two samples, and the corresponding values in the adjacency matrix are 0. To examine the performance of the models, we fixed the threshold to a value ranging from 0.1 to 0.9. Figure 7 shows that before 0.5, the model's performance improves significantly as the threshold is raised. After 0.5, it tends to be flat, and the model's performance peaks at the final threshold of 0.8. Our model RRGCN performed best when Pearson correlation threshold was 0.8, where the Precision, Recall, F1score, ACC, AUC and PR_AUC reached 0.862, 0.865, 0.854, 0.871, 0.984, and 0.936, respectively.

Performance of RRGCN in multi-omics

To verify the superiority of multi-omics data, as well as the validity and contribution of each type of data to the model, we conduct experiments on different types of data separately. From the experimental results (Table 3), it can be seen that using a single omics data training model, the highest performance is the

mRNA group with an accuracy of 0.7384, followed by the somatic group with an accuracy of 0.5190. The CNV group has the lowest accuracy, only 0.4766. It can be seen that although RNA-seq data has good performance and is indeed used in most studies, its effect is still inferior to multi-omics data. Of course, this also reflects from a certain level that RNA-seq data contains extremely important biological information, has good performance in the classification of cancer subtypes, and is extremely important for cancer diagnosis and treatment (Yang et al., 2021). It can be verified that there is complementary information between different omics, which can explain the nature of cancer from different perspectives and improve the diagnostic efficiency of cancer.

Comparison with other classical methods

To validate RRGCN's classification performance, we compare it to two other classical machine learning methods and two graph convolution-based classification approaches and evaluate it using four standard external evaluation measures. We employ four classification methods: Random Forest (Breiman, 2001), Support Vector Machine (Cortes and Vapnik, 1995), MoGCN (Li X. et al., 2022), and ERGCN (Dai et al., 2021).

- ·Random Forest (RF) is essentially a bagging algorithm, which randomly selects a feature from the most important features for branching, creates multiple decision trees, and finally votes on which category the data finally belongs to.
- ·Support Vector Machine (SVM), a binary classification model whose basic model is defined as a linear classifier with the biggest margin on the feature space. The goal is to build an objective function based on the structural risk reduction principle that distinguishes between the two types as much as possible.
- ·MoGCN is a multi-omics integration model based on GCN. The model utilizes feature extraction and network visualization for further biological knowledge discovery and subtype classification.
- ·ERGCN is a cancer subtype classification method based on residual graph convolutional networks and sample similarity networks for gene co-expression patterns.

To begin, we unified the AE latent layer feature matrix as input data for each model to ensure the rigor of the compared tests. Then, we utilized scikit-learn to construct these algorithms and grid search to optimize the RF and SVM parameters. The best number of sub-decision trees (n_esimators) for RF is between 1 and 101, with a step size of 10. Finally, 5-fold cross-validation yielded an optimal n_esimators of 26. The maximum number of features (max_features) should ideally be between 1 and 21, with a stride of 1. Finally, the optimal max_features is selected as 20 through 5-fold cross-validation. We also use grid search for SVM, choosing the penalty coefficient (C) from (0.1, 1, 100, 1,000) and the kernel function coefficient (gamma) from (0.0001, 0.001, 0.005, 0.1, 1, 3, 5), as well as the kernel function (kernel) from ("linear," "rbf"). The final optimized C is 1,000, the gamma is 0.001, and the kernel is "rbf." For MoGCN and ERGCN, we use the optimal parameters already set by their authors. The model comparison results are shown in Table 4.

From the results, we can see that RRGCN has an excellent performance in the classification of GC subtypes. The classification accuracy of RRGCN is as high as 0.8713, which is 5.49% higher than the best RF among the other four methods, and 11.47%, 8.83%, and 9.32% higher than the other three methods, respectively. The F1 score, Precision, and Recall of RRGCN are 0.8544, 0.8621, and 0.8654 respectively, and the values are also much higher than other methods. Most crucially, as compared to ERGCN, RRGCN performs better on each of the four evaluation metrics by 10.28%, 25.17%, 20.41%, and 41.41%, respectively. In defining the various subtypes of GC, RRGCN has more advantages. In the future, it might be used to treat more diseases, offering novel perspectives on how to diagnose and treat clinical illnesses.

4 Discussion

Heterogeneity causes cancer to differentiate into different subtypes, and subtypes with different degrees of differentiation and malignancy have different sensitivities to clinical therapeutic drugs, which brings great challenges to the diagnosis and treatment of the disease (Lin et al., 2021; Yuan et al., 2022). GC is a highly heterogeneous tumor, and its average somatic gene copy number changes are much higher than those of other tumor types (Joshi and Badgwell, 2021). Therefore, in clinical studies, the progression of GC is the slowest (Li et al., 2021).

Therefore, by integrating multi-omics data, we propose a graph convolutional network based on residual networks to realize the subtype classification of GC. Multi-omics datasets are dimensionally reduced by AE to extract representative latent layer features. The SNF algorithm is used to find the associations existing between patients. Finally, PSN combined with the feature matrix was input into RRGCN, and the classification results were output through the softmax layer. The results show that the accuracy of RRGCN reaches 0.8713.

The improvement of RRGCN over previous models is that multi-omics data is used as the basis of research, and the neglected similarity between patients is combined as the input of the model. For model selection, we introduce two skip connections to alleviate the loss of information during training and solve the model degradation problem.

To explain the advantage of multi-omics data, we retrain three different types of data separately and compare the results with multi-omics data. The results show that the performance of the model trained with multi-omics data is much higher than that of the single-omics data, and the accuracy is improved by about 18.00%. To prove the superiority of RRGCN, we compare RRGCN with classical machine learning methods and well-performing deep learning models, respectively. The results show that the performance of RRGCN is higher than other methods in all aspects. Most importantly, the accuracy of RRGCN is 10.28% higher than that of ERGCN.

The model is sensitive to the selection of the Pearson threshold, and the supervised learning method also brings inconvenience to the selection of data. In the future, we will focus on studying the application of graph convolution combined with other classical convolutional neural networks, considering the development of new unsupervised learning methods for cancer subtype recognition and classification.

5 Conclusion

In summary, we proposed a new classification method for gastric cancer subtypes called RRGCN by borrowing skip connections in residual networks. Through the deep mining of GC multi-omics data and the consideration of the relationship between patients, and comparing RRGCN with other classical machine learning methods and deep learning models, we verify the excellent performance of RRGCN in various aspects and improve the cancer subtype classification method to a higher level. The development of new models opens up new avenues for precise treatment. Li J. et al., (2022), Yang et al., (2022), and Hu et al., (2021). have tried to combine GCN with spatial transcriptomics for cell clustering and the identification of cancer subtypes. In the future, we will look into the spatial coordinate information of gastric cancer cells and employ unsupervised learning algorithms to provide more robust support for clinical diagnosis and treatment of gastric cancer.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CL designed the study and wrote the manuscript. YW collected data. YG and HK validated the findings of the experiment. QZ and YD analyzed the data. JH supervised the study, revised the manuscript and gave the final approval of the version to be published. All authors reviewed and approved this paper.

Funding

This work was supported by the University Excellent Talent Funding Project of Anhui Province (Grant no. gxgnfx2020088); the Natural Science Project of Anhui University of Chinese Medicine (Grant no. 2020wtzx02); and the Industry-University Cooperation Collaborative Education Project of the Ministry of Education of the People's Republic of China (Grant no. 202101123001).

Acknowledgments

I'd like to thank all of the participants for their contributions to the study, especially my mentor JH instruction and assistance, as well as the fund's assistance.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Baul, S., Ahmed, K. T., Filipek, J., and Zhang, W. (2022). omicsGAT: Graph attention network for cancer subtype analyses. *IJMS* 23, 10220. doi:10.3390/ijms231810220

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). "Pearson correlation coefficient," in *Noise reduction in speech processing springer topics in signal processing* (Berlin, Heidelberg: Springer Berlin Heidelberg), 1–4. doi:10.1007/978-3-642-00296-0_5

Binbusayyis, A., and Vaiyapuri, T. (2021). Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM. *Appl. Intell.* 51, 7094–7108. doi:10.1007/s10489-021-02205-9

Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi:10.1023/a:1010933404324

Clevert, D. A., Unterthiner, T., and Hochreiter, S. (2016). *Fast and accurate deep network learning by exponential linear units (elus)*. International Conference on Learning Representations (ICLR).

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: An R/bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44, e71. doi:10.1093/nar/gkv1507

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20 (3), 273–297. doi:10.1007/bf00994018

Dai, W., Yue, W., Peng, W., Fu, X., Liu, L., and Liu, L. (2021). Identifying cancer subtypes using a residual graph convolution model on a sample similarity network. *Genes* 13, 65. doi:10.3390/genes13010065

Dong, X., Chen, C., Deng, X., Liu, Y., Duan, Q., Peng, Z., et al. (2021). A novel mechanism for C1GALT1 in the regulation of gastric cancer progression. *Cell Biosci.* 11, 166. doi:10.1186/s13578-021-00678-2

Duan, R., Gao, L., Gao, Y., Hu, Y., Xu, H., Huang, M., et al. (2021). Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLoS Comput. Biol.* 17, e1009224. doi:10.1371/journal.pcbi.1009224

El-Manzalawy, Y., Hsieh, T.-Y., Shivakumar, M., Kim, D., and Honavar, V. (2018). Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data. *BMC Med. Genomics* 11, 71. doi:10.1186/s12920-018-0388-0

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., et al. (2021). Cancer statistics for the year 2020: An overview. *Int. J. Cancer* 149, 778–789. doi:10.1002/ijc.33588

Franco, E. F., Rana, P., Cruz, A., Calderón, V. V., Azevedo, V., Ramos, R. T. J., et al. (2021). Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data. *Cancers* 13, 2013. doi:10.3390/cancers13092013

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA, United States: O'Reilly Media, Inc.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27-30 June 2016 (IEEE), 770–778. doi:10.1109/CVPR.2016.90

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi:10.1126/science.1127647

Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., et al. (2021). SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* 18, 1342–1351. doi:10.1038/s41592-021-01255-8

Jia, Q., Chu, H., Jin, Z., Long, H., and Zhu, B. (2022). High-throughput single-cell sequencing in cancer research. *Sig Transduct. Target Ther.* 7, 145. doi:10.1038/s41392-022-00990-4

Joshi, S. S., and Badgwell, B. D. (2021). Current treatment and recent progress in gastric cancer. *CA A Cancer J. Clin.* 71, 264–279. doi:10.3322/caac.21657

Kim, J., Park, C., Kim, K. H., Kim, E. H., Kim, H., Woo, J. K., et al. (2022). Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage diversity and intratumoral heterogeneity. *npj Precis. Onc.* 6, 9. doi:10.1038/s41698-022-00251-1

Kim, S., Bae, S., Piao, Y., and Jo, K. (2021). Graph convolutional network for drug response prediction using gene expression data. *Mathematics* 9, 772. doi:10.3390/math9070772

Kingma, D. P., and Ba, J. A. (2015). *A method for stochastic optimization*. ICLR. arXiv preprint arXiv:1412.6980. doi:10.48550/arXiv.1412.6980

Kipf, T. N., and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. Available at: http://arxiv.org/abs/1609.02907 (Accessed November 3, 2022).

Li, J., Chen, S., Pan, X., Yuan, Y., and Shen, H.-B. (2022). Cell clustering for spatial transcriptomics data with graph neural networks. *Nat. Comput. Sci.* 2, 399–408. doi:10.1038/s43588-022-00266-5

Li T., T., Liu, Y., Liu, Q., Xu, W., Xiao, Y., and Liu, H. (2022). A malware propagation prediction model based on representation learning and graph convolutional networks. *Digital Commun. Netw.*, S2352-8648(22)00106-7. doi:10.1016/j.dcan.2022.05.015

Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F., et al. (2022). MoGCN: A multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Front. Genet.* 13, 806842. doi:10.3389/fgene.2022.806842

Li, Y.-K., Zhu, X.-R., Zhan, Y., Yuan, W.-Z., and Jin, W.-L. (2021). NEK7 promotes gastric cancer progression as a cell proliferation regulator. *Cancer Cell Int.* 21, 438. doi:10.1186/s12935-021-02148-8

Liang, C., Shang, M., and Luo, J. (2021). Cancer subtype identification by consensus guided graph autoencoders. *Bioinformatics* 37, 4779–4786. doi:10.1093/bioinformatics/btab535

Lin, Y., Pan, X., Zhao, L., Yang, C., Zhang, Z., Wang, B., et al. (2021). Immune cell infiltration signatures identified molecular subtypes and underlying mechanisms in gastric cancer. *npj Genom. Med.* 6, 83. doi:10.1038/s41525-021-00249-x

Lindskrog, S. V., Prip, F., Lamy, P., Taber, A., Groeneveld, C. S., Birkenkamp-Demtröder, K., et al. (2021). An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. *Nat. Commun.* 12, 2301. doi:10.1038/s41467-021-22465-w

Menyhárt, O., and Győrffy, B. (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput. Struct. Biotechnol. J.* 19, 949–960. doi:10.1016/j.csbj.2021.01.009

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Adv. neural Inf. Process. Syst.* 32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* 19, 3735–3746. doi:10.1016/j.csbj.2021.06.030

Ramirez, R., Chiu, Y.-C., Hererra, A., Mostavi, M., Ramirez, J., Chen, Y., et al. (2020). Classification of cancer types using graph convolutional neural networks. *Front. Phys.* 8, 203. doi:10.3389/fphy.2020.00203

Sammut, C., and Webb, G. I. (2010). "Mean squared error," in *Encyclopedia of machine learning* (Boston: Springer), 653.

Shao, W., Yang, Z., Fu, Y., Zheng, L., Liu, F., Chai, L., and Jia, J. (2021). The pyroptosis-related signature predicts prognosis and indicates immune microenvironment infiltration in gastric cancer. *Front. Cell Dev. Biol.* 9, 676485. doi:10.3389/fcell.2021.676485

Shin, J., Shin, D. W., Lee, J., Hwang, J., Lee, J. E., Cho, B., et al. (2022). Exploring socio-demographic, physical, psychological, and quality of life-related factors related with fear of cancer recurrence in stomach cancer survivors: A cross-sectional study. *BMC Cancer* 22, 414. doi:10.1186/s12885-022-09507-2

Sivadas, A., Kok, V. C., and Ng, K.-L. (2022). Multi-omics analyses provide novel biological insights to distinguish lobular ductal types of invasive breast cancers. *Breast Cancer Res. Treat.* 193, 361–379. doi:10.1007/s10549-022-06567-7

Sun, J., Zhou, Q., and Hu, X. (2019). Integrating multi-omics and regular analyses identifies the molecular responses of zebrafish brains to graphene oxide: Perspectives in environmental criteria. *Ecotoxicol. Environ. Saf.* 180, 269–279. doi:10.1016/j.ecoenv.2019.05.011

Sun, Y. V., and Hu, Y.-J. (2016). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv. Genet.* 93, 147–190. doi:10.1016/bs.adgen.2015.11.004

Tao, G.-Y., Ramakrishnan, M., Vinod, K. K., Yrjälä, K., Satheesh, V., Cho, J., et al. (2020). Multi-omics analysis of cellular pathways involved in different rapid growth stages of moso bamboo. *Tree Physiol.* 40, 1487–1508. doi:10.1093/treephys/tpaa090

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi:10.1038/nmeth.2810

Wang, B., Zhang, Y., Qing, T., Xing, K., Li, J., Zhen, T., et al. (2021). Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq. *Sci. Rep.* 11, 1141. doi:10.1038/s41598-020-80881-2

Wang, T.-H., Lee, C.-Y., Lee, T.-Y., Huang, H.-D., Hsu, J. B.-K., and Chang, T.-H. (2021). Biomarker identification through multiomics data analysis of prostate cancer prognostication using a deep learning model and similarity network fusion. *Cancers* 13, 2528. doi:10.3390/cancers13112528

Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., et al. (2021). MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* 12, 3445. doi:10.1038/s41467-021-23774-w

Wei, L., Jin, Z., Yang, S., Xu, Y., Zhu, Y., and Ji, Y. (2018). TCGA-Assembler 2: Software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics* 34, 1615–1617. doi:10.1093/bioinformatics/btx812

Xu, J., Wu, P., Chen, Y., Meng, Q., Dawood, H., and Dawood, H. (2019). A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. *BMC Bioinforma.* 20, 527. doi:10.1186/s12859-019-3116-7

Yamanaka, J., Kuwashima, S., and Kurita, T. (2017). "Fast and accurate image super resolution by deep CNN with skip connection and network in network," in International Conference on Neural Information Processing. Bangkok, Thailand, November, 2017 (Cham: Springer), 217–225. doi:10.1007/978-3-319-70096-0_23

Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., et al. (2022). scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* 4, 852–866. doi:10.1038/s42256-022-00534-z

Yang, H., Chen, R., Li, D., and Wang, Z. (2021). Subtype-GAN: A deep learning approach for integrative cancer subtyping of multi-omics data. *Bioinformatics* 37, 2231–2237. doi:10.1093/bioinformatics/btab109

Yuan, Q., Deng, D., Pan, C., Ren, J., Wei, T., Wu, Z., et al. (2022). Integration of transcriptomics, proteomics, and metabolomics data to reveal HER2-associated metabolic heterogeneity in gastric cancer with response to immunotherapy and neoadjuvant chemotherapy. *Front. Immunol.* 13, 951137. doi:10.3389/fimmu.2022.951137

Zhang, G., Peng, Z., Yan, C., Wang, J., Luo, J., and Luo, H. (2022). A novel liver cancer diagnosis method based on patient similarity network and DenseGCN. *Sci. Rep.* 12, 6797. doi:10.1038/s41598-022-10441-3

Zhang, X.-M., Liang, L., Liu, L., and Tang, M.-J. (2021). Graph neural networks and their current applications in bioinformatics. *Front. Genet.* 12, 690049. doi:10.3389/fgene.2021.690049

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 05 January 2023
DOI 10.3389/fgene.2022.1078609

OPEN ACCESS

EDITED BY
Dominik Heider,
University of Marburg, Germany

REVIEWED BY
Paolo Martini,
University of Brescia, Italy
Markus List,
Technical University of Munich, Germany

*CORRESPONDENCE
Enrique Hernández-Lemus,
ehernandez@inmegen.gob.mx

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 24 October 2022
ACCEPTED 15 December 2022
PUBLISHED 05 January 2023

CITATION
Ochoa S and Hernández-Lemus E (2023), Functional impact of multi-omic interactions in breast cancer subtypes. *Front. Genet.* 13:1078609. doi: 10.3389/fgene.2022.1078609

Functional impact of multi-omic interactions in breast cancer subtypes

Soledad Ochoa[1,2] and Enrique Hernández-Lemus[1,3]*

[1]Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2]Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3]Center for Complexity Sciences, Universidad Nacional Autónoma de México, Mexico City, Mexico

Multi-omic approaches are expected to deliver a broader molecular view of cancer. However, the promised mechanistic explanations have not quite settled yet. Here, we propose a theoretical and computational analysis framework to semi-automatically produce network models of the regulatory constraints influencing a biological function. This way, we identified functions significantly enriched on the analyzed omics and described associated features, for each of the four breast cancer molecular subtypes. For instance, we identified functions sustaining over-representation of invasion-related processes in the basal subtype and DNA modification processes in the normal tissue. We found limited overlap on the omics-associated functions between subtypes; however, a startling feature intersection within subtype functions also emerged. The examples presented highlight new, potentially regulatory features, with sound biological reasons to expect a connection with the functions. Multi-omic regulatory networks thus constitute reliable models of the way omics are connected, demonstrating a capability for systematic generation of mechanistic hypothesis.

KEYWORDS

multi-omics, breast cancer, network biology, HIF, RAS, WNT, SOX9, DNA methylation

1 Introduction

The establishment of high-throughput technologies has made possible a systems biology approach to cancer through multi-omics integration (Kristensen et al., 2014). The multi-omics perspective takes advantage of the complementarity between different molecular levels of description. However, the promise of attaining mechanistic explanations (Bersanelli et al., 2016) has not settled yet.

Although there is a plethora of statistical approximations (Huang et al., 2017), sparse multivariate methods are arguably nearer to the mechanistic explanation goal, given their capacity to pinpoint potential regulators (Li et al., 2012; Sohn et al., 2013; Bose et al., 2022). These approaches have even identified potential key regulators for each breast cancer subtype (Huang et al., 2019), and for the subgroups of the triple-negative breast cancer subtype (Chappell et al., 2021). The networks shown in some of these works (Li et al., 2012; Sohn et al., 2013; Huang et al., 2019) constitute hypothesized models of the way regulators are connected, demonstrating a capability for systematic production of testable regulatory mechanisms.

Here, we applied the sparse generalized canonical correlation analysis (SGCCA) to data on DNA methylation and gene and miRNA expression from TCGA. The SGCCA is a statistical method that outputs correlated features among a large collection by the use of LASSO penalization (Tenenhaus et al., 2014). The SGCCA has been successfully used for biomarker discovery from cancer (Fan et al., 2020) and non-cancer contexts (Garali et al., 2018). In order to find not just the features but the connections between them, SGCCA was

coupled with ARACNE (Margolin et al., 2006), a method for inference of transcriptional networks, that has allowed our group to find transcriptional master regulators (Tapia-Carrillo et al., 2019), to document a loss of long-distance co-expression (García-Cortés et al., 2020; Dorantes-Gilardi et al., 2021; García-Cortés et al., 2021), and to evaluate the role that relevant miRNAs play in some oncogenic processes (Drago-García et al., 2017; Zamora-Fuentes et al., 2022), among other applications in the large-scale molecular study of cancer. As an outcome, we describe some of the reconstructed networks and their implications, highlighting their relevance to understand cancer biology and potentially impact treatment. The general pipeline is described in Figure 1.

2 Methods

All the analyses described hereafter were performed with R programming language version 4.1.1 (R Core Team, 2021) and can be found at http://csbig.inmegen.gob.mx/SGCCA/. Release 105 of `biomaRt` was used all along and plots were produced with `ggplot2` (Wickham, 2016).

2.1 Data acquisition

TCGA data were obtained through the `TCGAbiolinks` R package. We only used samples with Illumina Human Methylation 450, RNA-seq, and miRNA-seq data from unique patients. This constraints the number of samples to 128 from the basal subtype, 46 from Her2-enriched, 416 from luminal A, 140 from luminal B, and 75 samples from normal adjacent tissue.

Pre-processing has been described before (Ochoa et al., 2021) and follows published guidelines (Aryee et al., 2014; Tam et al., 2015; Tarazona et al., 2015). As a first step, only protein-coding transcripts were kept since for our purposes, these were considered the main functional effectors. This restriction toward the study of non-coding features was chosen in order to focus on the expression regulatory layers of DNA methylation, miRNA expression and, hidden among the transcripts, the layer of transcription factors. Length and GC content biases were checked with the `NOISeq` package (Tarazona et al., 2015) and alleviated using `EDASeq` (Risso et al., 2011) full normalization. Genes with zero counts were (the only ones) discarded at the low count filter, TMM normalization was applied between samples, and the batch effect was corrected. Since batch effects can be induced by *a priori*-unknown factors, `ARSyNseq` was used to remove all systematic noise not associated with the subtypes (Nueda et al., 2012). Preprocessing of microRNAs is the same, except there is no length or GC bias and the normalization used between samples is the median method.

Finally, CpG probes with over 25% missing values and non-mapped or overlapping SNPs were discarded. The remaining missing values were imputed *via* nearest neighbors and transformed into M-value matrices. This way, datasets account for 393,132 methylation probes, 17,077 coding transcripts, and 604 miRNA precursors.

2.2 Sparse generalized canonical correlation analysis

Once pre-processing was performed, we normalized each omic by the square root of the first eigenvalue and concatenated them patient-wise, obtaining one matrix per breast cancer subtype and one for the normal tissue. Using this normalization ensures the influence of each omic over upcoming analysis depends on its variance (De Tayrac et al., 2009).

Afterward, we approached the SGCCA as implemented in the `mixOmics` package (Rohart et al., 2017) and largely followed the Garali et al. guidance (Garali et al., 2018). The analysis takes as input the different blocks of data and a sparsity parameter per block, the number of components to recover (`ncomp`), a design matrix, and a function to maximize the covariance. Sparsity parameters were chosen for each omic from the sequence [0.01, 0.02, ..., 0.09, 0.1, ..., 0.9], by cross-validation. With this purpose, a balanced dataset, composed of 10 samples per tumor subtype and 10 samples from normal tissue, was randomly taken from the original data, 10 times per each sparsity parameter value. Each time, a simple SGCCA was run, recovering only one component and taking note of the selected number of features and the average variance explained (AVE). Summing the different combinations, in total, every value was tested 11,340 times per omic. Sparsity parameters were chosen in order to obtain the largest AVE with the lowest number of features (Supplementary Figure S1), namely, 0.02 for CpG sites and transcripts and 0.05 for microRNAs.

Data analytics included several stages: independent pre-processing to deal with factors specific to the platforms, while normalization and penalization concern appropriate data integration. Eigenvalue normalization was further performed to equilibrate the still disparate rank of the different values. Separate penalization considers the different signal sizes the distinct omics may have. Shrinking the same CpG coefficients and miRNA coefficients may over-penalize relevant associations yet with effects smaller than those coming from other omics Liu et al. (2018). After the fitting process, we noticed that miRNAs are slightly less penalized than the other omics.

The definite SGCCA for each subtype and the normal tissue was run using the fitted values. The smaller the sparsity value, the fewer features get selected. For each subtype, we used the number of samples minus 1 as ncomp, the default design matrix, and the `centroid` function, which enables negative correlation.

Feature selection attained by SGCCA is expected to be a bit unstable due to the LASSO penalization. Mimicking the filter used in `miRDriver` (Bose and Bozdag, 2019), we re-run SGCCA 100 times per subtype, or the normal tissue, using a random subset of half the samples each time. We only kept those features selected at least 70% of the time.

2.3 Functional enrichment analysis

SGCCA results include a matrix of the loadings a feature has in each component. The said matrix is quite sparse, except for the features summarizing the relevant information between and within omics. These non-zero loadings indicate co-selected features that can be tested for functional enrichment.

With the idea of exploiting the full set of co-selected features, and not just the transcripts, all the features, being CpG probes, miRNA precursors, or transcripts, were mapped to Entrez gene IDs. Both transcripts and miRNAs have a direct annotation at Entrez, (e.g., hsa-mir-34b becomes MIR34B). To translate CpG probes to Entrez IDs, we recovered the genes affected by each probe from the microarray annotation file. This results in an amplification of CpG representation

since one site can be associated with a whole cluster of genes and assumes a methylation effect on overlapping genes, which is not necessarily true. Both are cons of this mapping that need to be considered.

Then, the group of features with non-zero loading in every SGCCA component was submitted to a separate over-representation analysis, taking Entrez IDs as input. Enrichment was run using the `clusterProfiler` package (Wu et al., 2021) against the pathways from the KEGG database (Kanehisa and Goto, 2000) and against the biological process gene ontology (Consortium, 2021). A significance threshold of FDR-corrected p-values < 0.01 was set. The intersection between sets of enriched functions was plotted with the `UpSetR` package (Gehlenborg, 2019). Functions exclusively enriched in one dataset were tested for over-representation. With this purpose, exclusively enriched functions were grouped according to GOslim and KEGG classes. Dependence between grouped categories and the subtypes was assessed with Fisher's test, and p-values were adjusted for multiple testing using the Bonferroni method.

In an independent manner, we ran a gene set enrichment analysis (GSEA), only with transcript data, to check for functions affected by differential expression. GSEA was also performed with the `clusterProfiler` package, in this case, without a p-value cutoff. The idea is to recover a GSEA enrichment score for every one of the functions over-represented in the SGCCA results. Such scores would answer if functions over-represented among the features related through different omics are also enriched among genes with altered expression. We must stress, however, that all discussed functions are significantly over-represented (p-value < 0.01), but only the specified ones also have a significant GSEA score.

2.4 Network reconstruction

Chosen functions were represented as networks to draw potentially regulatory models. To achieve this, we estimated mutual information (MI) between every pair of nodes using ARACNE software (Margolin et al., 2006) and then filtered out all the pairs with lower MI than the median value observed for known regulatory interactions. Thus, for each chosen function, we recovered all the features co-selected (co-varying) with the features responsible for the functional enrichment and focused on this set.

1. We extracted a sub-matrix from the original dataset and run ARACNE.
2. We retrieved regulatory interactions involving the selected features. Again, this was performed with the microarray annotation file for the CpGs, assuming position overlap is enough to affect gene expression. The `multiMiR` package (Ru et al., 2014) was used in the case of miRNAs and `TFtargets` (github.com/slowkow) for the transcript coding for transcription factors. This latter package queries several resources, namely, TRED, ITFP, ENCODE, TRRUST, and the databases from Neph et al., 2012; Marbach et al., 2016 (Jiang et al., 2007; Zheng et al., 2008; Consortium et al., 2012; Neph et al., 2012; Han et al., 2015; Marbach et al., 2016), which include validated and predicted interactions. We considered those hits coming from ChIP-seq, DNaseI footprinting, and small-scale experiments as validated.
3. We obtained MI values for such regulatory interactions, using the `infotheo` package (Meyer, 2014) (the use of this specific tool obeys the need to focus on a reduced set of given pairs, instead of estimating all the pairs with a feature of interest in the adjacency matrix, as ARACNE would perform).
4. We took the median MI value for the regulatory interactions as the threshold. Since MI is expected to differ between the distinct kinds of pairs, different thresholds were obtained for the different types of edges: CpG–transcript, CpG–miRNAs, TF transcript–transcript, and miRNA–transcript. The median was preferred over the mean to avoid outliers dominating the threshold.
5. The MI value distribution obtained with ARACNE was contrasted between types of edges, *via* Kolmorogov–Smirnov tests. If distributions were not significantly different, the lowest median MI from regulatory interactions—obtained with infotheo—was chosen as the unique threshold to pass, no matter the edge-type, relaxing the threshold and increasing the MI interactions accepted in the final network.

The output of these items is a table with predicted interactions and weights that illustrate the largest statistical dependencies between the features selected by the SGCCA.

2.5 Network analysis

Mutual information networks were analyzed with the `igraph` package (Csardi and Nepusz, 2006) and represented with `Cytoscape` (Shannon et al., 2003), making use of the RCy3 package (Gustavsen et al., 2019).

Node colors represent logFC values between every subtype and the normal tissue. MiRNA differential expression went through *voom* normalization and *eBayes* `limma` function. Since the batch effect was not corrected in methylation data, we used the `missMethyl` package for the differential analysis. This tool removes systematic errors of unknown origin, bypassing the lack of batch-effect correction (Maksimovic et al., 2015).

The node degree was calculated for the whole network; however, only those network components with features annotated as players of a function are shown in the corresponding figures. Since Her2+ and luminal B subtypes produce large networks, we further zoomed in the graph by selecting only the first neighbors of functional features. Such subnetworks may serve as a model of the regulatory pressures influencing the function.

Every neighbor of a functional node was searched in PubMed, together with the associated functions, to find out if some biological role has already been suggested. PubMed was also queried with every pair of interacting nodes, as well as the databases for predicted regulatory links accessible through `multiMir`. Transcription factor-related features are identified according to the list from humantfs (Lambert et al., 2018). This achieves a fairly automated way to build a regulatory model for the functions enriched in the SGCCA.

3 Results and discussion

By applying SGCCA, we have identified, for each one of the breast cancer subtypes, transcripts whose expression patterns better reflect the variance in its own block, while also co-varying with the other

blocks of data. The pattern of selected features by omics and subtype is provided in Supplementary Figure S2.

SGCCA uses a LASSO penalization, which may select inconsistent sets of features. Since this could affect the reliability of functional enrichment, identifying functions dependent on unstable features, we just proceeded with the features most consistently selected, whose proportion is shown in Supplementary Figure S3. There are no individual transcripts or miRNAs selected simultaneously across all five datasets, but there are six CpG sites in this situation which potentially affect MAPK8IP3, AFAP1, LFNG, and VSTM2B.

The transcripts repeatedly selected in the same subtype have known associations with breast cancer. The top three transcripts selected more often for the basal subtype are MCL1, CTNNA1, and NOTCH3. MCL1 is an anti-apoptotic member of the BCL2 family that is required for mammary stem cell function (Fu et al., 2015), and it is expected to be overexpressed in tumors of this subtype (Farrugia et al., 2015). Meanwhile, catenin alpha 1 is postulated to act as a tumor suppressor in E-cadherin-negative basal-like breast cancer cells (Piao et al., 2014), and NOTCH3 seems to function as a promoter of the epithelial–mesenchymal transition (Liang et al., 2018).

Her2 enriched has also been clearly associated with its most selected transcripts: CEACAM5, ACACA, and PGK1. Though heterogeneously expressed, Her2-enriched tumors tend to be positive for CEACAM5 (Bechmann et al., 2020) and so this adhesion molecule has been suggested as a target for T-cell bi-specific antibodies (Messaoudene et al., 2019). Inhibitors of acetyl-CoA carboxylase, ACACA, work over MCF-7 cells overexpressing Her2 by interfering with cancer stem cell lipid biosynthesis and the Warburg effect (Corominas-Faja et al., 2014). At last, PGK1 protein has been found overexpressed in these tumors (Schulz et al., 2009), while being linked to macrophages and stratifying patients at higher risk (Li et al., 2021).

Interestingly, microRNAs from the let-7 family were among the top selected for basal, Her2+, and luminal B subtypes, as well as for normal breast tissue. These miRNAs regulate JAK-STAT3 and Myc signaling pathways, thus affecting stemness and metastasis (Thammaiah and Jayaram, 2016).

3.1 Functions enriched on SGCCA output differ between datasets

After inspecting the overall output of SGCCA, we wanted to know if there are functions involving the co-varying features. Enrichment against GO biological processes and KEGG pathways allows us to identify functions affected by the specific regulatory mechanisms identified.

A total of 683 GO biological processes and 69 KEGG pathways were found significantly over-represented (FDR adjusted p. value < 0.01) among the SGCCA co-selected features. Figure 2 shows the intersections between subtypes. Few functions were found enriched across all subtypes, and most of them are either exclusive or shared only by a pair of subtypes. That is, functions associated with DNA methylation and miRNA expression are not the same for all subtypes.

There are three biological processes significantly enriched (FDR-corrected p-value ≤.0099, for the specific values, see Supplementary Table S1) in the four subtypes and the normal tissue. These are the developmental processes: metanephric nephron development (GO:0072210), metanephros development (GO:0001656), and pattern specification process (GO:0007389). Since GO:0072210 is a part of GO:0001656, they may be considered the same.

Then, we wondered if functions linked with DNA methylation and miRNA expression in cancer and normal tissue maintain an intact circuitry connecting CpGs, transcripts, and miRNAs. In more general terms, does a function enriched twice involve identical features and interactions?

3.2 Features responsible for the same functional enrichment differ across subtypes

The first step toward a shared circuitry connecting CpGs, transcripts, and miRNAs in different phenotypes would be to have the same (or similar) features behind the functional enrichment. To verify if this happens, we calculated the Jaccard index for every pair of functions enriched more than once. The Jaccard index divides the size of intersection between two sets by their union, measuring similarity with a normalized value between 0 (fully disjoint sets) and 1 (the same set). Distributions for the Jaccard index are shown in Figure 3A.

The obtained distributions are enough to state that, for most functions, the CpG–transcript–miRNA circuitry is not the same across datasets since the features involved are not the same. Only seven biological processes enriched in a given pair of SGCCA results share more than 50% of the involved features. Five of them are related to development, while the other two are related to cell adhesion. These are the functions that may share the interactions between CpG sites, transcripts, and miRNAs.

If this index hints at the similarity between subtypes pertaining to CpG–transcript–miRNA co-variation, the distance with Her2-enriched subtype results are intriguing. This may be caused by a bias induced by the low number of samples. Or perhaps this is associated with the lower correlation with DNA methylation patterns (Network et al., 2012). Not surprisingly, the pair with the most similarly enriched functions corresponds to the two luminal subtypes.

3.3 Exclusive category over-representation

To answer if functions exclusively found in one dataset bring to light subtype-specific properties, we analyzed over-representation of GOslim categories and KEGG classes. The proportion of biological processes found for each dataset in every one of the categories is given in Figure 3B, while the equivalent plot for KEGG pathways is found in Supplementary Figure S4.

None of the KEGG classes is biased toward a given subtype, but there is an enrichment for the categories: *cellular component organization* in the basal SGCCA components, *establishment of localization* in luminal A, and *DNA metabolic process* in the normal tissue. There are seven biological processes behind the *cellular component organization* over-representation, comprising five processes related to axon extension, which are clustered with regulation of the extent of cell growth. Collagen fibril organization is not in the cluster and is the seventh process, suggesting a potential bond between the basal subtype and invasiveness.

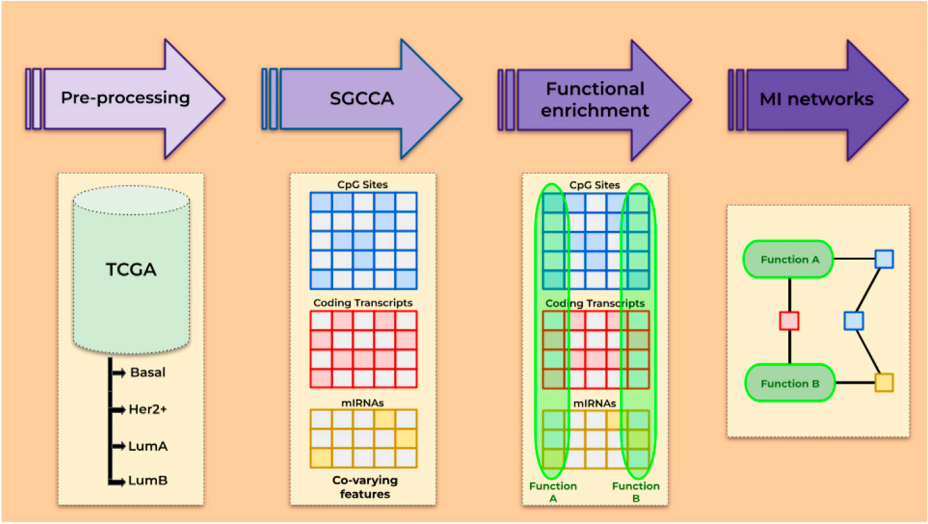


FIGURE 1
Overview of the steps followed.

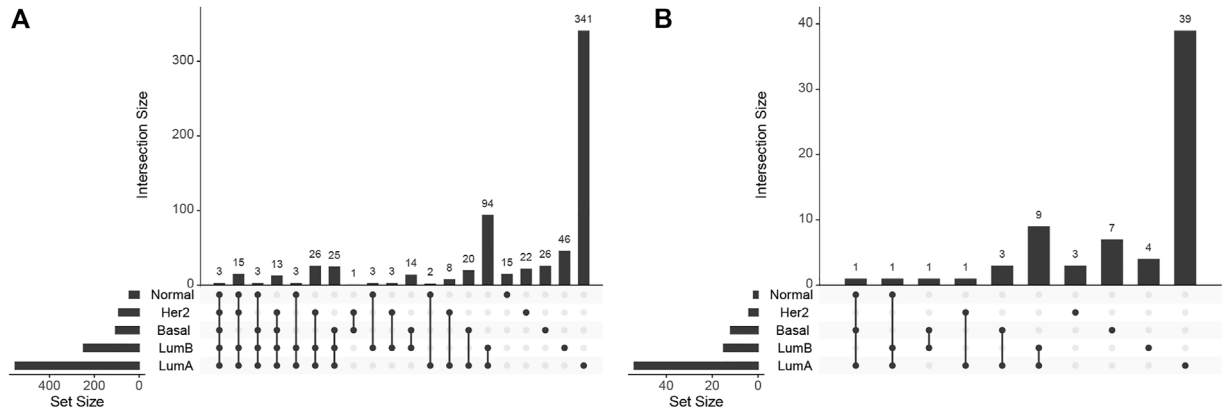


FIGURE 2
UpSet plot for **(A)** biological processes and **(B)** KEGG pathways enrichment.

In the case of luminal A, there are 62 biological processes behind the over-representation of *establishment of localization*. These processes affect transport and secretion and conform to 32 different clusters. Regarding over-representation in the normal tissue, it is interesting that it is related to DNA alkylation and methylation processes, perhaps implying that these processes are somehow disarranged on the tumor subtypes.

3.4 Within subtypes, different functions can be connected through correlated features

When checking the features responsible for the enrichment of a given function, we discovered that several functions are enriched in the exact same set of co-varying features, that is, the same set of SGCCA components. This suggests some level of crosstalk between functions that can be connected through correlated features. This observation has been made subtype-wise and implies that a single network of correlated features may actually span several functions.

Going through each subtype separately, we clustered functions by the proportion of SGCCA components shared. Figure 4 shows Her2 clusters. There are 11 clusters and six functions that cannot be grouped since they involve features that are not related with the clusters. Taking the bigger labels as a guide, purple, orange, and fuchsia clusters are related with development of kidney structures. Green and blue clusters at the bottom are linked with connective tissue development. Pale pink nodes refer to distinct processes of morphogenesis, while the nodes in yellow allude development of reproductive structures. The small brown and pale green clusters are related to cardiac muscle and neural cells, respectively. Finally, the small clusters in the center, in bright green and pale orange, are linked with metabolism and loaded with functions exclusively found in this subtype, a fact that may be

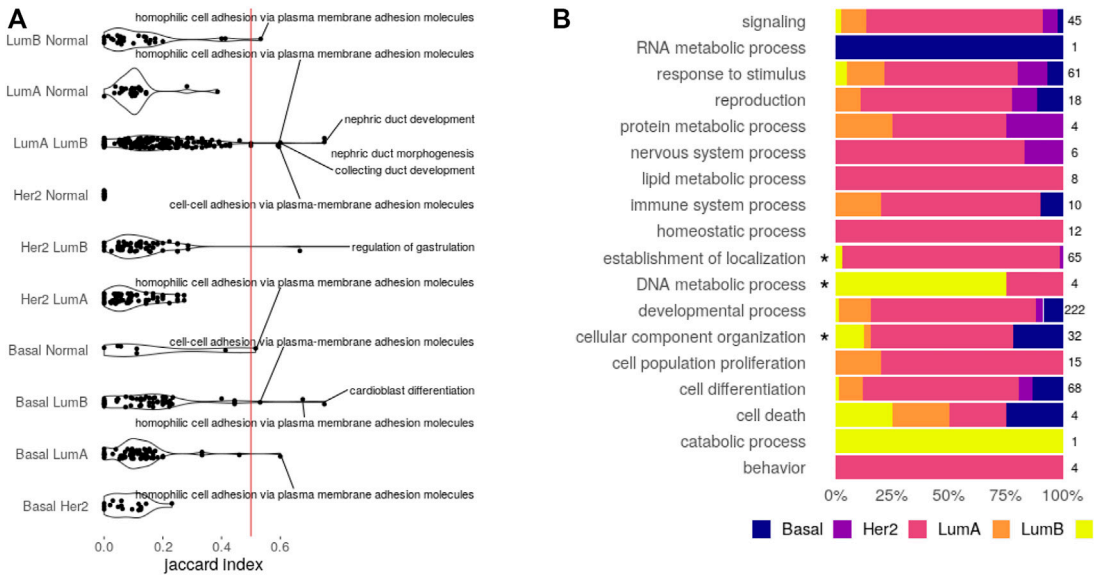


FIGURE 3
Enriched functions. **(A)** Feature similarity between functions shared by the pair of datasets indicated. Functions with similarity over 0.5 are displayed. **(B)** Bias of exclusive functions. An asterisk marks categories with significant over-representation (Fisher's test, Bonferroni adjusted p-value < 0.05).

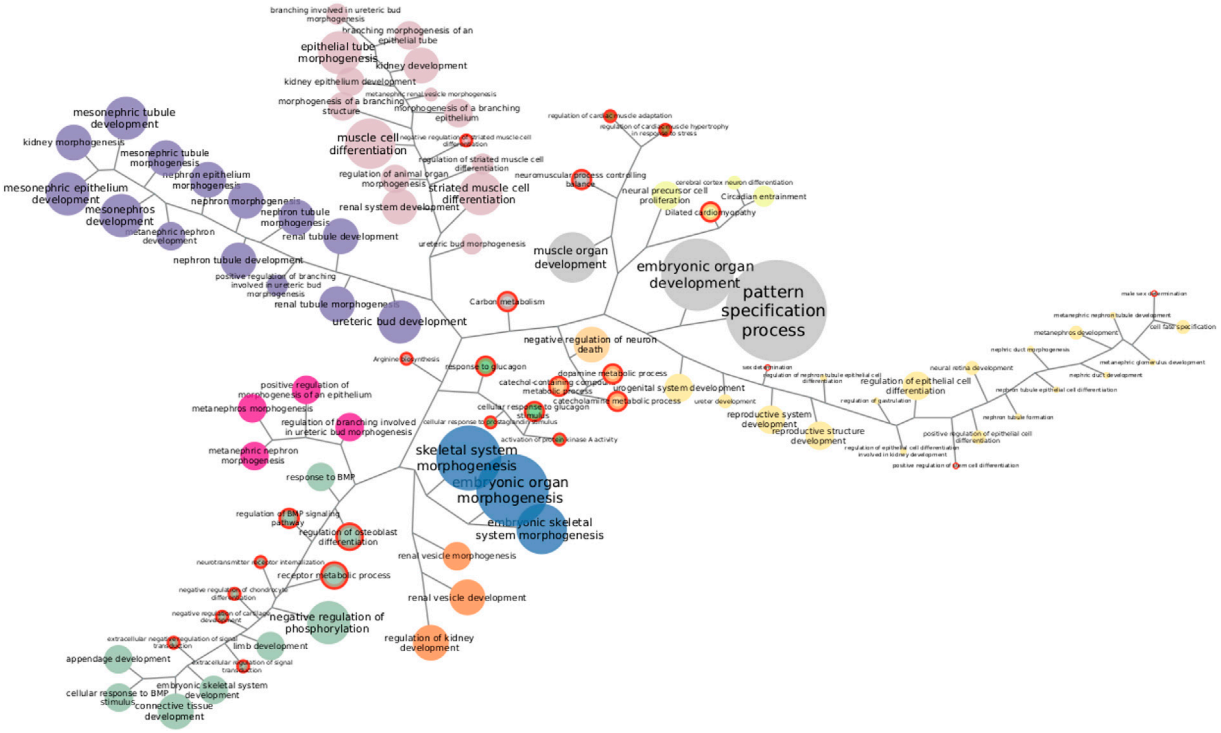


FIGURE 4
Functions enriched in Her2 SGCCA components. Both KEGG pathways and GO biological processes are represented together. Same-color clustered nodes are enriched in the same components. Nodes in gray do not belong to any cluster. The size of nodes and labels reflect the number of features behind the enrichment. Functions exclusively found in this subtype are highlighted with a red border.

interesting to explore further. The functions enriched with the most genes do not form a part of any cluster.

Clustering exposes information that needs to be accounted when discussing one particular enrichment. Functions exclusively found in one subtype may reveal mechanistic explanations of subtype-specific alterations, but, if exclusive functions are clustered with others that are non-exclusive and better represented, relevance may be debatable. Similarly, clusters may help explain some odd enrichments, like the one found in the luminal A dataset for morphine addiction. Morphine addiction has been found enriched on methylation-driven genes (Xu et al., 2019) but depends on features correlated with those responsible for ECM–receptor interaction, suggesting co-variation may be pulling up the enrichment for this addiction. Even after considering clusters, there are enrichments hard to figure out fully; however, some specific features can be actually tracked (Supplementary Tables S1, S2).

In order to select functions to explore further, we repeated the analysis described with Her2+ for each SGCCA result. While not all clustering are displayed here, full groups and enrichment results are supplied as Supplementary Files. A filtering step was necessary because, even with the clustering, there are almost 500 sets of functionally related features. It is interesting that the two cell adhesion processes with the Jaccard index over 0.5 appear consistently out of any cluster in the subtypes with such enrichment.

3.5 Network examples

In our path to answer if a function enriched twice involves identical features and interactions, we found that a given function is commonly enriched through distinct sets of features in two different datasets. At the same time, we observed several functions over-represented among the same sets of co-selected features and wondered how functions were connected. Functions involving the same features are already identifiable in the annotation databases, but by means of this multi-omic integration strategy, we have been able to find cross-linking paths across single layers and maybe even connect seemingly independent functions through multi-omics pattern co-variation. To check how this appears, we built mutual information (MI) networks. The networks went through a stringent threshold to keep just the interactions that are most likely regulatory. To this end, we obtained the MI values accompanying true regulatory interactions and took the median value as the minimum MI required to consider an edge as possibly regulatory. Within these reduced sets of interactions, the following figures show the network components that contain those features annotated as participating in the functions, though some of the obtained networks extend further.

The intuition is that co-selected features, whose patterns are correlated with those of functional features, may also be participating in a given function. Beyond that, nodes for miRNAs, CpGs, and transcripts that ultimately code for transcription factors may be playing regulatory roles. The stringent threshold attempts to filter out the interactions owed to simple co-variation. Two broad possible scenarios are expected, 1) disconnected components per function, each with its own potential regulators, or 2) functions that crosstalk through common features, whose potential regulators could be of medical interest. The different scenarios are exemplified through the four subtypes and the normal tissue in the coming sections.

3.5.1 HIF-1 signaling in the basal subtype

Hypoxia-inducible factor 1 (HIF-1) signaling is one of the KEGG pathways enriched exclusively in the basal SGCCA results. HIF-1 is the master regulator of oxygen homeostasis since it induces transcription from at least 100 hypoxia-responsive elements (Corrado and Fontana, 2020). HIF-1 signaling is activated in tumors not only under hypoxic conditions but also by oxygen-independent factors, like TP53 and BRCA mutations (de Heer et al., 2020), which have been associated with the basal subtype (Network et al., 2012).

The network we identified for this function is given in Figure 5. AMPK signaling is enriched in a subset of the same SGCCA components such as HIF-1 signaling, which is consistent with the idea that these two pathways interplay in cancer metabolism re-programming (Moldogazieva et al., 2020). However, after applying the MI threshold, each pathway occupies disconnected components.

Only two functional features, that is, annotated as participants of the function, pass the MI threshold, NOS3 and RPS6. It is important to clarify that the enrichment does not rest only on these two features, but we only find interactions over the threshold for them. There are also two nodes that have been linked with the signaling pathway without being participants as such. PEG3 gets upregulated after hypoxia in mouse lungs (Wollen et al., 2013), while SUMO3 would be one of the modifiers affecting HIF-1 stability (Matic et al., 2008). Thus, nodes seem to be associated with the function.

On the other hand, the complete network is formed by CpG–transcript interactions, more specifically, by edges linking a CpG with a transcript coding for a ribosomal protein. Since CpG sites are not in the same chromosome as the transcript, a direct regulatory influence can be discarded. To account for indirect relations, we estimated the mutual information between the corresponding transcripts, even when these were not originally in the SGCCA set of co-selected features. Obtained MI values are smaller than the global threshold and smaller than the edges between CpGs and ribosomal protein-coding transcripts. Hence, indirect effects going through the transcript linked with the CpG do not seem to fully explain the phenomenon.

Most nodes are not significantly different from the normal tissue, either regarding expression or methylation values. This is consistent with the lack of significance of the pathway GSEA score (NES = 0.9252, adjusted p-value: 0.7937). HIF-1 signaling in the basal subtype is transcriptionally comparable with that of the normal tissue. Nevertheless, the pathway is not found enriched in the normal tissue SGCCA output, suggesting a change in the correlation between omics.

3.5.2 Positive regulation of stem cell differentiation in the Her2-enriched subtype

Cancer stem cells are largely responsible for relapse and metastasis. Her2 variants, observed in Her2+ patients with poor clinical outcomes, have been reported to drive maintenance and enrichment of breast cancer stem cells (Pupa et al., 2021). *Positive regulation of stem cell differentiation* was found enriched exclusively in Her2+ data, but related processes also appear in the other three subtypes. The process is clustered with several other functions, as shown in Figure 6, where we have focused on the first neighbors of the functional features.The transcription factor SOX9 is the only feature

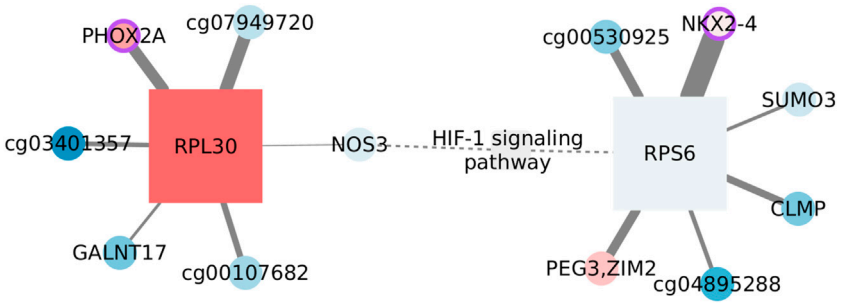


FIGURE 5
Features connected with HIF-1 signaling in the basal subtype. Circles represent CpGs, and squares are transcripts. When possible, CpGs are identified with the symbol of the gene they affect; otherwise, the ID of the probe is used. The shades of red indicate the level of overexpression/methylation against the normal tissue, while blue tones represent values under what is expected. The node size reflects its degree. A purple border identifies nodes whose protein plays a transcription factor role. The weight of the link is the extent of mutual information between connected nodes. Dashed edges link MI components with prior information.

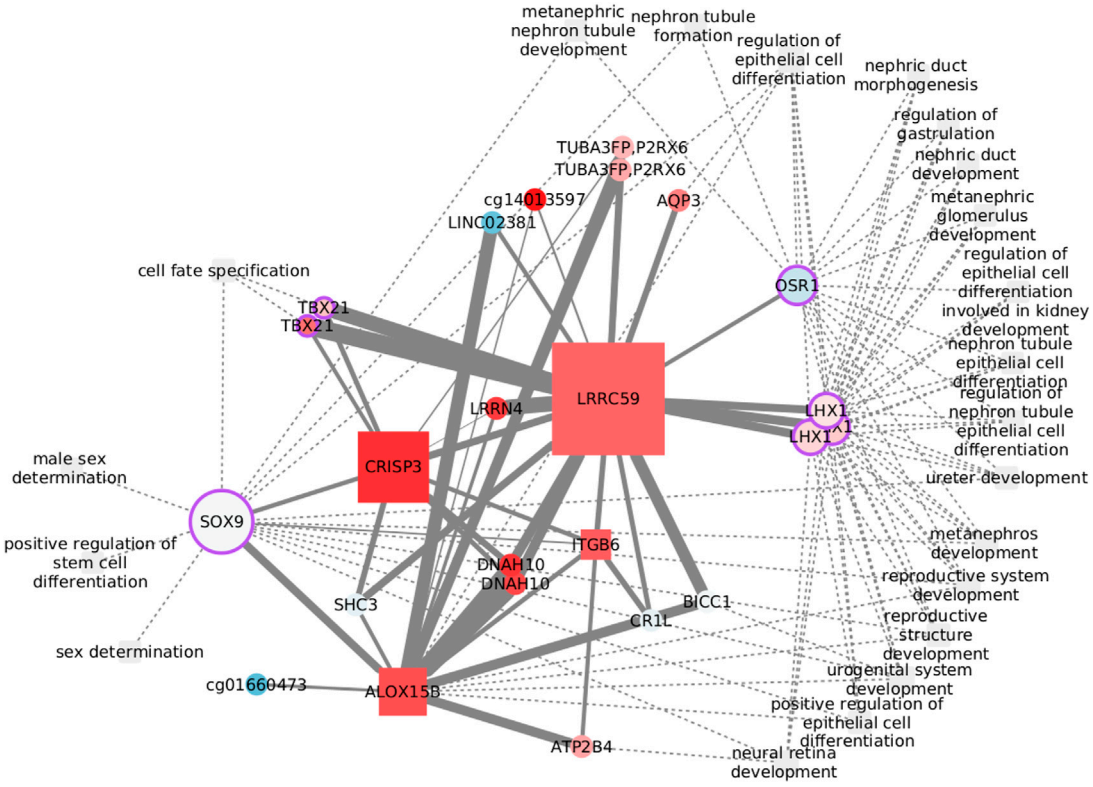


FIGURE 6
Features connected with the regulation of stem cell differentiation in the Her2-enriched subtype. Node size reflects the betweenness centrality.

from *positive regulation of stem cell differentiation* with edges passing the MI threshold. SOX9 binds functions related with cell fate and sex determination, while LHX1 and OSR1 are at the crossroads of most functions. None of the edges has been previously reported, but several nodes have known links with these functions. The relation between CRISP3 and sex determination, for instance, may be explained by the role of the protein in sperm function (Weigel Muñoz et al., 2019) and its up regulation in prostate cancer (Pathak et al., 2016). DNAH10 is another feature with a known bond with sex determination, specifically with sperm flagella morphological abnormalities (Li et al., 2022). The connection with ITGB6 is perhaps weaker since it rests only on differential expression analysis of prostate cancer (Li et al., 2013). CR1L is involved in B lymphocyte activation (Fernández-Centeno et al., 2000) and may have a role in renal injury (He et al., 2005). Finally, the somehow unexpected *neural retina development* is related with the function of SHC3 (Nakazawa et al., 2002).

The functional implications of some of these nodes are specifically dependent on DNA methylation. Although epigenetically altered

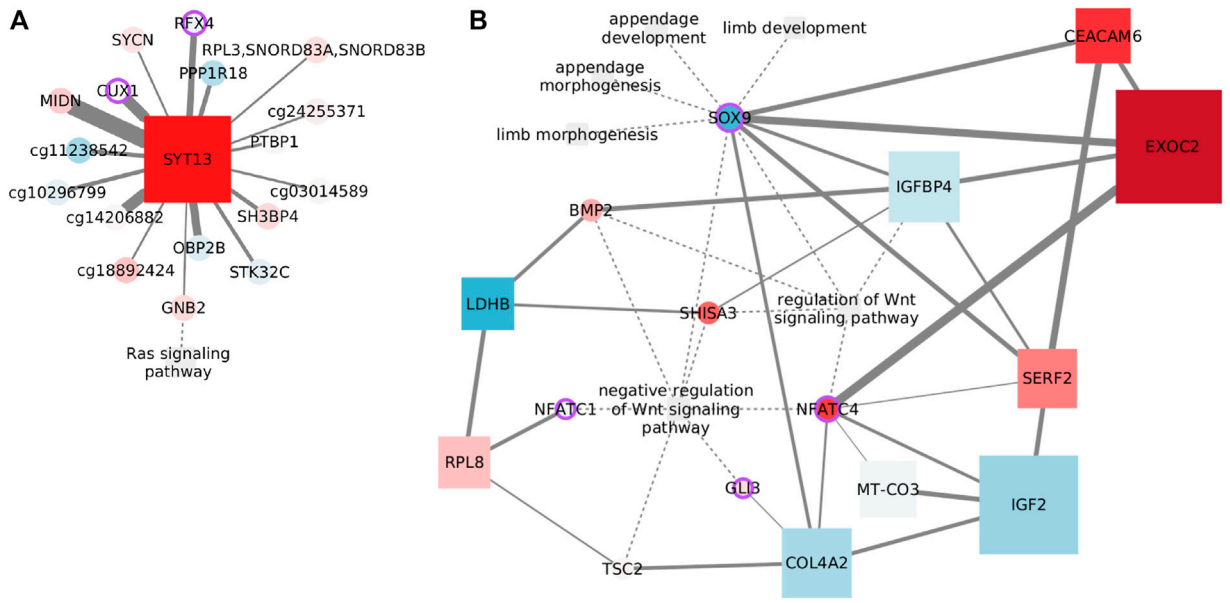


FIGURE 7
Example networks for the luminal subtypes. **(A)** Features connected with the Ras signaling pathway in luminal A data. Node size indicates degree. **(B)** Features connected with Wnt signaling in the luminal B subtype. Node size reflects betweenness centrality.

CR1L is linked with Alzheimer's and dementia (Bahado-Singh et al., 2021), DNAH10 has emerged when studying renal carcinomas with a CpG-island methylator phenotype (Arai et al., 2015). Finally, CpG methylation of the lncRNA LINC02381 functions as a tumor suppressor in colorectal cancer (Jafarzadeh et al., 2020). While all of these features are represented by CpG sites in the network, LINC02381 appearance highlights the complexity of transcription regulation and the need to widen multi-omic analysis to include more data layers.

Despite that transcription factors may be the obvious option to explore the crosstalk between biological processes, less explored options, like ALOX15B, CRISP3, and LRRC59, with elevated graph betweenness, may result of interest.

3.5.3 Ras signaling pathway in the luminal A subtype

Ras signaling is one of the many pathways exclusively found enriched in the luminal A subtype. It is a well-documented pathway influencing cancer aspects like cell proliferation, survival, migration, and differentiation. Although the pathway is more frequently activated in the other subtypes, it has been reported as an indicator of poor prognosis in luminal tumors (Wright et al., 2015). Not surprisingly, Ras signaling components are under-expressed relative to the normal tissue (NES: 1.5796, adjusted *p*-value: 0.0084) in this analysis.

Only one functional feature endures the MI threshold, GNB2. The subunit beta 2 of G protein links the signaling pathway with a set of CpGs associated with cell communication and brain function, through the calcium sensor SYT13. Genes affected by the CpGs include the brain active kinase, STK32C; MIDN, that is predicted to enable kinase binding; OBP2B, which is supposed to enable binding of small volatile molecules; a TF from early brain development, RFX4; and SYCN, which is predicted to be active in exocytosis.

Among the remaining nodes, the connection with the CUX1 CpG site agrees with the cooperation observed between this transcription factor and Kras-G12V mutant in lung cancer (Ramdzan et al., 2014). In a similar way, PTBP1 overexpression is known to co-occur with oncogenic KRAS mutations in colon cancer (Hollander et al., 2016). Finally, a connection with the transferrin receptor internalization protein, SH3BP4, has been predicted before by a random forest classifier (Xin et al., 2021).

Again, the network shown in Figure 7A links a transcript with CpG sites all over the genome. Although it has been proposed that Ras signaling controls aberrant DNA methylation (Patra, 2008), the specific influence nodes may have over the signaling pathway remains unclear.

3.5.4 Negative regulation of the Wnt signaling pathway in the luminal B subtype

Wnt signaling normally controls organ development. In breast cancer, Wnt signaling is involved in tumor proliferation and metastasis, immune microenvironment regulation, stemness maintenance, and therapeutic resistance (Xu et al., 2020). The relevance of this function does not end here, but it has also been associated explicitly with the luminal B subtype. Though generalized DNA hypomethylation is common in cancer (Vidal Ocabo et al., 2017), a fraction of luminal B tumors exhibit hypermethylation, specifically affecting Wnt signaling (Network et al., 2012).

In our results, *negative regulation of* the *Wnt signaling pathway* is exclusively found enriched in this subtype, but related Wnt pathways were also found for luminal A. The cross-talking functions shown in Figure 7B are not in the same cluster but are found in a subset of the SGCCA components, where negative regulation of Wnt appears. Since these related functions makeup the largest network—after the threshold—we have, and this network consists of a large single

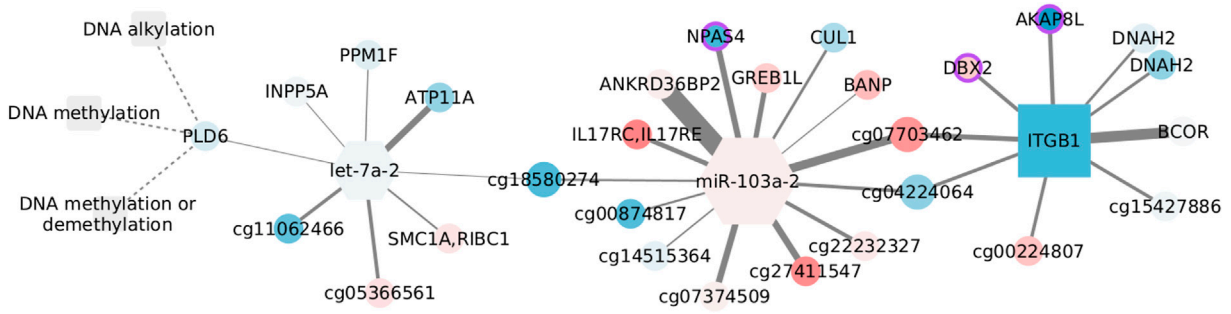


FIGURE 8
Features connected with DNA methylation in the normal adjacent tissue. Color corresponds to log fold changes in the basal subtype.

component, we decided to focus on the first neighbors of the functional features.

As expected, *negative regulation of Wnt signaling* and *regulation of Wnt* share functional features. The transcription factor for skeletal development, Sox9, is represented by its CpG at the crossroad between Wnt signaling with the developmental processes, but there are also multiple indirect paths. Since the genes coding collagen subunit Col4a2 and cell adhesion molecule Ceacam6 are targeted by Sox9 (Sumi et al., 2007), and Sox9 acts in cooperation with Gli3 (Tan et al., 2018), that pair of edges are easy to justify. Similarly, the link between COL4A2 and NFATC4 could be explained by the inhibition of the nuclear translocation of NFATc4 by Col4a2 in cardiomyocytes (Sugiyama et al., 2020), while both COL4A2 and IGF2 code for extracellular proteins deregulated under diseases with EMT (Bueno et al., 2011). Additionally, bone marrow stromal cells induced with IGFBP4, among other factors, overexpress SOX9 (Liu et al., 2012). Insulin-like growth factor-binding protein 4 is also connected with BMP2, as IGFBP4 overexpression impairs BMP2-induced osteogenic differentiation (Wu et al., 2017).

In summary, there are sound biological reasons to expect co-variation of the connected features. The question to solve is how such connections affect Wnt signaling and luminal B cancer progression, specifically what is the role of the node with the highest betweenness. Exocyst complex component 2 is related with the Wnt pathway as an effector of Hedgehog signaling (Arraf et al., 2020) and has been associated with metastasis and different cancer types (Cerhan et al., 2014; Hazelett and Yeaman, 2012; D'Aloia et al., 2018), but not with breast cancer.

3.5.5 DNA methylation in the normal adjacent tissue

DNA methylation is exclusively enriched in the normal tissue, but we choose to discuss it because of its relevance for cancer (Baylin and Jones, 2016). In addition, unlike the other examples, this network does contain microRNAs, including the top selected let-7a-2.

For consistency, we colored the nodes in Figure 8. However, since the normal tissue is our reference value, we used the log fold changes obtained by contrasting basal and normal tissue expression. This subtype has significant overexpression of related genes (NES = 1.9251, adjusted p-values = 0.0031) and has been linked with hypomethylation (Network et al., 2012). Yet, we have to emphasize that DNA methylation is not enriched in the basal data, and so, the relation between CpGs, miRNAs, and transcripts may not follow what is suggested in this figure.

Despite none of the interactions has been reported, a couple of nodes are somehow connected with the DNA methylation machinery. AKAP8L interacts with core subunits of the H3K4 histone methyltransferase complexes (Bieluszewska et al., 2018), whose action is interrelated with DNA modification (Rose and Klose, 2014). BCOR is part of the non-canonical polycomb repressive complex 1 and is altered in distinct cancer types (Astolfi et al., 2019). It has been observed that BANP can open the chromatin at unmethylated CpG-island promoters, thus activating essential genes in pluripotent stem and differentiated neuronal cells (Grand et al., 2021). Finally, *de novo* DNA methyltransferase, DNMT3b, can interact with CUL1, involving this node in aberrant methylation (Shamay et al., 2010).

In contrast, another set of nodes hinges on epigenetic silencing, as is the case of INPP5A in lung adenocarcinoma (Ke et al., 2020). Together with ATP11A, INPP5A CpG methylation has shown discriminatory capacity for colorectal cancer (Izquierdo et al., 2021). In the same manner, ATP11A methylation distinguishes several diseases including metastatic-lethal prostate cancer (Zhao et al., 2017), while a methylation signature including the growth regulation by estrogen in breast cancer 1 like GREB1L separates gastric adenocarcinoma cases by overall survival, and DBX2 methylation marks the serum from hepatocellular cancer patients (Zhang et al., 2013). Similar to its paralog DNAH10, DNAH2 aberrations are frequent in renal carcinomas with a CpG-island methylator phenotype (Arai et al., 2015). Although unexpected, the brain-specific transcription factor NPAS4, present in the form of a CpG site, is known to be regulated by DNA methylation (Furukawa-Hibi et al., 2015) and has been linked with colon adenocarcinoma survival (Luo et al., 2021). Last, though ITGB1 methylation is expected to be constant both in cancer and normal tissue (Strelnikov et al., 2021), alteration of the gene expression has been observed in basal-like tumors and cells with BRCA mutation, highlighting the relevance of migration and mesenchymal properties for this subtype (Privat et al., 2018).

Interestingly, the two miRNAs in the network are associated with migration and invasion, although in opposite ways. The let-7 family works as a tumor suppressor and is inhibited by DNA methylation and several regulators (Thammaiah and Jayaram, 2016). Contrastingly, miR-103 acts as an oncogene in triple-negative tumors, and its over-expression is linked with poor prognosis (Xiong et al., 2017). In spite of the low fold changes, the expression of both miRNAs is coherent with what would be expected in the basal subtype.

4 Conclusion

Here, we have described the kind of multi-omic network models that can be obtained through the sequential application of SGCCA and ARACNE. The collection of interactions shown in any of these networks suggests a multi-omic model that may or may not have regulatory implications. To asseverate regulation, wet laboratory testing would be needed. However, the nature of nodes as CpG sites, microRNAs, or transcript coding for functional proteins must be considered, as shown in the examples. Although further testing is required, the examples embody the level of details we can get in the way toward targeted experimental validation of multi-omic regulatory phenomena.

Though the interactions encountered seem to be subtype-specific, given the low values of the Jaccard index, there is no restriction to believe these same associations could not be repeated in other contexts, with somehow equivalent patterns of methylation and expression. Instead, an interesting question arises about the traceability of tissue and disease signals. A fair attempt to carry out would be to compare cancer and tissue networks with the same nodes, even if the edge weights are disparate, which were not produced here. Also, it has to be noticed that the normal adjacent tissue may not be the best control since it carries detected alterations across tissues (Aran et al., 2017).

The use of SGCCA allowed us to identify the functions enriched in features co-varying across DNA methylation, transcript, and miRNA expression. This does not mean such functions may not be influenced by other regulatory mechanisms: this simply indicates the functions, like HIF signaling in the basal subtype, depending the furthest on features whose methylation and expression co-vary. The con of the method is the instability of the LASSO, which forced us to keep just the features identified in over 70% of subsamples. Even when other tools (Hernández-de Diego et al., 2018; Meng et al., 2019) could achieve the multi-omic functional enrichment without the instability issue, we prefer the sparse method exactly because of the stable portion of the feature set. Then, possible improvements include the elastic network penalization, which overcomes the stability problem.

mixOmics output for the SGCCA includes a complete graph connecting all the features selected in a component. However, having found the same functions over-represented in different components, we wanted to further explore the relations among all the features co-varying with those associated with a given function. The mutual information statistical dependency measure has desirable properties for multi-omic integration, such as being able to capture non-linear relations and being a parameterization invariant. Moreover, we wanted to discern likely regulatory interactions, a task that has been successfully achieved with ARACNE for transcriptomics. With edges linking different types of nodes, such discerning becomes harder because ARACNE's data processing inequality (DPI) cannot be used in a straightforward manner. Thus, the setting of varying thresholds based on regulatory interactions is established. In this case, MI ability to recover non-linear relations may not be fully profited, being posterior to the lineal filter of SGCCA. MI is, however, used as a way to bring together all the results concerning a function and highlight some potentially interesting pairs of nodes.

The DPI posed with ARACNE discards the lowest weighted edge from a triad, as a likely indirect interaction driven by the other pair of nodes. The difficulty of using it comes from the observation that mutual information distributions change with the different omics (Drago-García et al., 2017). While maintaining the treatment of lower weighted interactions as indirect, the threshold we applied accounts for the difference between omics by estimating MI values from known regulatory interactions.

It is worth considering that MI has a dependency on the number of observations, which varies between subtypes and the normal tissue. Her2 enriched has a smaller number of samples than recommended, and so special care must be taken with it. Given that MI is rank-invariant, it is expected that, even with the stringent threshold, only a subset of the interactions in Figure 6 keep relevance when increasing dataset size. By progressing from a set where every feature is correlated with one another to highly significant interactions (Pethel and Hahs, 2014; Mukherjee et al., 2020), we pursue an automatic assembly of regulatory models. Tools better suited to find regulatory interactions (Kuijjer et al., 2020; Sonawane et al., 2021) require prior information not always available or heavier calculations (Weighill et al., 2021), making the approach described here an accessible solution.

To end with the pros and cons' discussion, here, we have overlooked interactions between CpG sites because those are beyond described regulatory mechanisms. Nevertheless, links between CpG sites are accompanied by large MI values that would surpass our threshold and may become of relevance in the cancer context (Akulenko and Helms, 2013; Zhang and Huang, 2017). On the other hand, links with miRNAs were expected but only appeared in the normal tissue example. Drago-García et al. had already reported lower MI values for these types of links (Drago-García et al., 2017). Despite the threshold attempted to incorporate this difference on the MI, our multi-omic pipeline does not recover miRNA interactions as well as other dedicated methods (Bose et al., 2022).

The networks produced in this way capture statistical dependencies that may guide further work. However, such a hypothetical future work depends on a user being able to find these kinds of networks and research the reasons behind a statistical dependency. Article databases can serve this purpose, as we have done here, but may become unspecific. Instead, network databases (Arif et al., 2021; Ben Guebila et al., 2022) may offer a smoother connection between wet and dry laboratories, in order to transcend statistical description toward actual knowledge acquisition.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

SO organized data, performed calculations, analyzed data, discussed results, and drafted the manuscript; EH-L designed the study, contributed to the methodological approach, discussed results, reviewed the manuscript, and supervised the project. Both authors read and approved the final manuscript.

Funding

SO is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 615847 from CONACYT. This work was partially performed at cluster INMEGEN and received technical support from Israel Aguilar-Ordoñez. The results published here are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1078609/full#supplementary-material

References

Akulenko, R., and Helms, V. (2013). Dna co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Hum. Mol. Genet.* 22, 3016. doi:10.1093/hmg/ddt158

Arai, E., Gotoh, M., Tian, Y., Sakamoto, H., Ono, M., Matsuda, A., et al. (2015). Alterations of the spindle checkpoint pathway in clinicopathologically aggressive c p g island methylator phenotype clear cell renal cell carcinomas. *Int. J. cancer* 137, 2589–2606. doi:10.1002/ijc.29630

Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., et al. (2017). Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat. Commun.* 8, 1077–1114. doi:10.1038/s41467-017-01027-z

Arif, M., Zhang, C., Li, X., Güngör, C., Çakmak, B., Arslantürk, M., et al. (2021). Inetmodels 2.0: an interactive visualization and database of multi-omics data. *Nucleic acids Res.* 49, W271–W276. doi:10.1093/nar/gkab254

Arraf, A. A., Yelin, R., Reshef, I., Jadon, J., Abboud, M., Zaher, M., et al. (2020). Hedgehog signaling regulates epithelial morphogenesis to position the ventral embryonic midline. *Dev. Cell.* 53, 589–602. doi:10.1016/j.devcel.2020.04.016

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014). Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinforma. Oxf. Engl.* 30, 1363–1369. doi:10.1093/bioinformatics/btu049

Astolfi, A., Fiore, M., Melchionda, F., Indio, V., Bertuccio, S. N., and Pession, A. (2019). Bcor involvement in cancer. *Epigenomics* 11, 835–855. doi:10.2217/epi-2018-0195

Bahado-Singh, R. O., Vishweswaraiah, S., Aydas, B., Yilmaz, A., Metpally, R. P., Carey, D. J., et al. (2021). Artificial intelligence and leukocyte epigenomics: Evaluation and prediction of late-onset alzheimer's disease. *PloS one* 16, e0248375. doi:10.1371/journal.pone.0248375

Baylin, S. B., and Jones, P. A. (2016). Epigenetic determinants of cancer. *Cold Spring Harb. Perspect. Biol.* 8, a019505. doi:10.1101/cshperspect.a019505

Bechmann, M. B., Brydholm, A. V., Codony, V. L., Kim, J., and Villadsen, R. (2020). Heterogeneity of ceacam5 in breast cancer. *Oncotarget* 11, 3886–3899. doi:10.18632/oncotarget.27778

Ben Guebila, M., Lopes-Ramos, C. M., Weighill, D., Sonawane, A. R., Burkholz, R., Shamsaei, B., et al. (2022). Grand: A database of gene regulatory network models across human conditions. *Nucleic acids Res.* 50, D610–D621. doi:10.1093/nar/gkab778

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., et al. (2016). Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinforma.* 17, S15. doi:10.1186/s12859-015-0857-9

Bieluszewska, A., Weglewska, M., Bieluszewski, T., Lesniewicz, K., and Poreba, E. (2018). Pka-binding domain of akap 8 is essential for direct interaction with dpy 30 protein. *FEBS J.* 285, 947–964. doi:10.1111/febs.14378

Bose, B., and Bozdag, S. (2019). "mirdriver: A tool to infer copy number derived mirna-gene networks in cancer," in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 366.

Bose, B., Moravec, M., and Bozdag, S. (2022). Computing microrna-gene interaction networks in pan-cancer using mirdriver. *Sci. Rep.* 12, 3717–17. doi:10.1038/s41598-022-07628-z

Bueno, D. F., Sunaga, D. Y., Kobayashi, G. S., Aguena, M., Raposo-Amaral, C. E., Masotti, C., et al. (2011). Human stem cell cultures from cleft lip/palate patients show enrichment of transcripts involved in extracellular matrix modeling by comparison to controls. *Stem Cell. Rev. Rep.* 7, 446–457. doi:10.1007/s12015-010-9197-3

Cerhan, J. R., Berndt, S. I., Vijai, J., Ghesquières, H., McKay, J., Wang, S. S., et al. (2014). Genome-wide association study identifies multiple susceptibility loci for diffuse large b cell lymphoma. *Nat. Genet.* 46, 1233–1238. doi:10.1038/ng.3105

Chappell, K., Manna, K., Washam, C. L., Graw, S., Alkam, D., Thompson, M. D., et al. (2021). Multi-omics data integration reveals correlated regulatory features of triple negative breast cancer. *Mol. Omics* 17, 677–691. doi:10.1039/d1mo00117e

Consortium, E. P., et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature* 489, 57–74. doi:10.1038/nature11247

Consortium, G. O. (2021). The gene ontology resource: Enriching a gold mine. *Nucleic Acids Res.* 49, D325–D334. doi:10.1093/nar/gkaa1113

Corominas-Faja, B., Cuyàs, E., Gumuzio, J., Bosch-Barrera, J., Leis, O., Martin, Á. G., et al. (2014). Chemical inhibition of acetyl-coa carboxylase suppresses self-renewal growth of cancer stem cells. *Oncotarget* 5, 8306–8316. doi:10.18632/oncotarget.2059

Corrado, C., and Fontana, S. (2020). Hypoxia and hif signaling: One axis with divergent effects. *Int. J. Mol. Sci.* 21, 5611. doi:10.3390/ijms21165611

Csardi, G., and Nepusz, T. (2006). *The igraph software package for complex network research*. Cambridge, MA: NECSI, 1695.

D'Aloia, A., Berruti, G., Costa, B., Schiller, C., Ambrosini, R., Pastori, V., et al. (2018). Ralgps2 is involved in tunneling nanotubes formation in 5637 bladder cancer cells. *Exp. Cell. Res.* 362, 349–361. doi:10.1016/j.yexcr.2017.11.036

de Heer, E. C., Jalving, M., Harris, A. L., et al. (2020). Hifs, angiogenesis, and metabolism: Elusive enemies in breast cancer. *J. Clin. investigation* 130, 5074–5087. doi:10.1172/JCI137552

De Tayrac, M., Lê, S., Aubry, M., Mosser, J., and Husson, F. (2009). Simultaneous analysis of distinct omics data sets with integration of biological knowledge: Multiple factor analysis approach. *BMC genomics* 10, 32. doi:10.1186/1471-2164-10-32

Dorantes-Gilardi, R., García-Cortés, D., Hernández-Lemus, E., and Espinal-Enríquez, J. (2021). k-core genes underpin structural features of breast cancer. *Sci. Rep.* 11, 16284–16317. doi:10.1038/s41598-021-95313-y

Drago-García, D., Espinal-Enríquez, J., and Hernández-Lemus, E. (2017). Network analysis of emt and met micro-rna regulation in breast cancer. *Sci. Rep.* 7, 13534. doi:10.1038/s41598-017-13903-1

Fan, Z., Zhou, Y., and Ressom, H. W. (2020). Mota: Network-based multi-omic data integration for biomarker discovery. *Metabolites* 10, 144. doi:10.3390/metabo10040144

Farrugia, M., Sharma, S., Lin, C., McLaughlin, S., Vanderbilt, D., Ammer, A., et al. (2015). Regulation of anti-apoptotic signaling by kruppel-like factors 4 and 5 mediates lapatinib resistance in breast cancer. *Cell. death Dis.* 6, e1699. doi:10.1038/cddis.2015.65

Fernández-Centeno, E., de Ojeda, G., Rojo, J. M., and Portolés, P. (2000). Crry/p65, a membrane complement regulatory protein, has costimulatory properties on mouse t cells. *J. Immunol.* 164, 4533–4542. doi:10.4049/jimmunol.164.9.4533

Fu, N. Y., Rios, A. C., Pal, B., Soetanto, R., Lun, A. T., Liu, K., et al. (2015). Egf-mediated induction of mcl-1 at the switch to lactation is essential for alveolar cell survival. *Nat. Cell. Biol.* 17, 365–375. doi:10.1038/ncb3117

Furukawa-Hibi, Y., Nagai, T., Yun, J., and Yamada, K. (2015). Stress increases dna methylation of the neuronal pas domain 4 (npas4) gene. *Neuroreport* 26, 827–832. doi:10.1097/WNR.0000000000000430

Garali, I., Adanyeguh, I. M., Ichou, F., Perlbarg, V., Seyer, A., Colsch, B., et al. (2018). A strategy for multimodal data integration: Application to biomarkers identification in spinocerebellar ataxia. *Briefings Bioinforma.* 19, 1356–1369. doi:10.1093/bib/bbx060

García-Cortés, D., de Anda-Jáuregui, G., Fresno, C., Hernández-Lemus, E., and Espinal-Enríquez, J. (2020). Gene co-expression is distance-dependent in breast cancer. *Front. Oncol.* 10, 1232. doi:10.3389/fonc.2020.01232

García-Cortés, D., Hernández-Lemus, E., and Espinal-Enríquez, J. (2021). Luminal a breast cancer co-expression network: Structural and functional alterations. *Front. Genet.* 12, 629475. doi:10.3389/fgene.2021.629475

Gehlenborg, N. (2019). *UpSetR: A more scalable alternative to venn and euler diagrams for visualizing intersecting sets*. Oxford, England: Oxford Academic.

Grand, R. S., Burger, L., Gräwe, C., Michael, A. K., Isbel, L., Hess, D., et al. (2021). Banp opens chromatin and activates cpg-island-regulated genes. *Nature* 596, 133–137. doi:10.1038/s41586-021-03689-8

Gustavsen, A., J., Pai, S., Isserlin, R., et al. (2019). Rcy3: Network biology using cytoscape from within r. *F1000Research* doi:10.12688/f1000research.20887.3

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., et al. (2015). Trrust: A reference database of human transcriptional regulatory interactions. *Sci. Rep.* 5, 11432–11511. doi:10.1038/srep11432

Hazelett, C. C., and Yeaman, C. (2012). Sec5 and exo84 mediate distinct aspects of rala-dependent cell polarization. *PLoS One* 7, e39602. doi:10.1371/journal.pone.0039602

He, C., Imai, M., Song, H., Quigg, R. J., and Tomlinson, S. (2005). Complement inhibitors targeted to the proximal tubule prevent injury in experimental nephrotic syndrome and demonstrate a key role for c5b-9. *J. Immunol.* 174, 5750–5757. doi:10.4049/jimmunol.174.9.5750

Hernández-de Diego, R., Tarazona, S., Martínez-Mira, C., Balzano-Nogueira, L., Furió-Tarí, P., Pappas, G. J., et al. (2018). Paintomics 3: A web resource for the pathway analysis and visualization of multi-omics data. *Nucleic acids Res.* 46, W503-W509. doi:10.1093/nar/gky466

Hollander, D., Donyo, M., Atias, N., Mekahel, K., Melamed, Z., Yannai, S., et al. (2016). A network-based analysis of colon cancer splicing changes reveals a tumorigenesis-favoring regulatory pathway emanating from elk1. *Genome Res.* 26, 541–553. doi:10.1101/gr.193169.115

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: Recent progress in multi-omics data integration methods. *Front. Genet.* 8, 84. doi:10.3389/fgene.2017.00084

Huang, S., Xu, W., Hu, P., and Lakowski, T. M. (2019). Integrative analysis reveals subtype-specific regulatory determinants in triple negative breast cancer. *Cancers* 11, 507. doi:10.3390/cancers11040507

Izquierdo, A. G., Boughanem, H., Diaz-Lagares, A., Arranz-Salas, I., Esteller, M., Tinahones, F. J., et al. (2021). Dna methylome in visceral adipose tissue can discriminate patients with and without colorectal cancer. *Epigenetics* 1–12, 665–676. doi:10.1080/15592294.2021.1950991

Jafarzadeh, M., Soltani, B. M., Soleimani, M., and Hosseinkhani, S. (2020). Epigenetically silenced linc02381 functions as a tumor suppressor by regulating pi3k-akt signaling pathway. *Biochimie* 171, 63–71. doi:10.1016/j.biochi.2020.02.009

Jiang, C., Xuan, Z., Zhao, F., and Zhang, M. Q. (2007). Tred: A transcriptional regulatory element database, new entries and other development. *Nucleic acids Res.* 35, D137–D140. doi:10.1093/nar/gkl1041

Kanehisa, M., and Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Ke, H., Wu, Y., Wang, R., and Wu, X. (2020). Creation of a prognostic risk prediction model for lung adenocarcinoma based on gene expression, methylation, and clinical characteristics. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.* 26, 9258333–e925841. doi:10.12659/MSM.925833

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi:10.1038/nrc3721

Kuijjer, M. L., Fagny, M., Marin, A., Quackenbush, J., and Glass, K. (2020). Puma: Panda using microrna associations. *Bioinformatics* 36, 4765–4773. doi:10.1093/bioinformatics/btaa571

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell.* 172, 650–665. doi:10.1016/j.cell.2018.01.029

Li, J., Xu, Y.-H., Lu, Y., Ma, X.-P., Chen, P., Luo, S.-W., et al. (2013). Identifying differentially expressed genes and small molecule drugs for prostate cancer by a bioinformatics strategy. *Asian Pac. J. cancer Prev.* 14, 5281–5286. doi:10.7314/apjcp.2013.14.9.5281

Li, W., Zhang, S., Liu, C.-C., and Zhou, X. J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics* 28, 2458–2466. doi:10.1093/bioinformatics/bts476

Li, Y., Wang, Y., Wen, Y., Zhang, T., Wang, X., Jiang, C., et al. (2022). Whole-exome sequencing of a cohort of infertile men reveals novel causative genes in teratozoospermia that are chiefly related to sperm head defects. *Hum. Reprod.* 37, 152–177. doi:10.1093/humrep/deab229

Li, Y., Zhao, X., Liu, Q., and Liu, Y. (2021). Bioinformatics reveal macrophages marker genes signature in breast cancer to predict prognosis. *Ann. Med.* 53, 1019–1031. doi:10.1080/07853890.2021.1914343

Liang, Y.-K., Lin, H.-Y., Dou, X.-W., Chen, M., Wei, X.-L., Zhang, Y.-Q., et al. (2018). Mir-221/222 promote epithelial-mesenchymal transition by targeting notch3 in breast cancer cell lines. *NPJ breast cancer* 4, 20–29. doi:10.1038/s41523-018-0073-7

Liu, J., Liang, G., Siegmund, K. D., and Lewinger, J. P. (2018). Data integration by multi-tuning parameter elastic net regression. *BMC Bioinforma.* 19, 369. doi:10.1186/s12859-018-2401-1

Liu, J., Liu, X., Zhou, G., Xiao, R., and Cao, Y. (2012). Conditioned medium from chondrocyte/scaffold constructs induced chondrogenic differentiation of bone marrow stromal cells. *Anatomical Rec. Adv. Integr. Anat. Evol. Biol.* 295, 1109–1116. doi:10.1002/ar.22500

Luo, Y., Sun, F., Peng, X., Dong, D., Ou, W., Xie, Y., et al. (2021). Integrated bioinformatics analysis to identify abnormal methylated differentially expressed genes for predicting prognosis of human colon cancer. *Int. J. General Med.* 14, 4745–4756. doi:10.2147/IJGM.S324483

Maksimovic, J., Gagnon-Bartsch, J. A., Speed, T. P., and Oshlack, A. (2015). Removing unwanted variation in a differential methylation analysis of illumina humanmethylation450 array data. *Nucleic acids Res.* 43, e106. doi:10.1093/nar/gkv526

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. methods* 13, 366–370. doi:10.1038/nmeth.3799

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinforma. Biomed. Cent.* 7, S7. doi:10.1186/1471-2105-7-S1-S7

Matic, I., van Hagen, M., Schimmel, J., Macek, B., Ogg, S. C., Tatham, M. H., et al. (2008). *In vivo* identification of human small ubiquitin-like modifier polymerization sites by high accuracy mass spectrometry and an *in vitro* to *in vivo* strategy. *Mol. Cell. proteomics* 7, 132–144. doi:10.1074/mcp.M700173-MCP200

Meng, C., Basunia, A., Peters, B., Gholami, A. M., Kuster, B., and Culhane, A. C. (2019). Mogsa: Integrative single sample gene-set analysis of multiple omics data. *Mol. Cell. Proteomics* 18, S153-S168–S168. doi:10.1074/mcp.TIR118.001251

Messaoudene, M., Mourikis, T., Michels, J., Fu, Y., Bonvalet, M., Lacroix-Trikki, M., et al. (2019). T-Cell bispecific antibodies in node-positive breast cancer: Novel therapeutic avenue for mhc class i loss variants. *Ann. Oncol.* 30, 934–944. doi:10.1093/annonc/mdz112

Meyer, P. E. (2014). *Infotheo: Information-Theoretic measures*. Princeton, NJ: R. package.

Moldogazieva, N. T., Mokhosoev, I. M., and Terentiev, A. A. (2020). Metabolic heterogeneity of cancer cells: An interplay between hif-1, gluts, and ampk. *Cancers* 12, 862. doi:10.3390/cancers12040862

Mukherjee, S., Asnani, H., and Kannan, S. (2020). "Ccmi: Classifier based conditional mutual information estimation," in Proceedings of Machine Learning Research.

Nakazawa, T., Nakano, I., Sato, M., Nakamura, T., Tamai, M., and Mori, N. (2002). Comparative expression profiles of trk receptors and shc-related phosphotyrosine adapters during retinal development: Potential roles of n-shc/shcc in brain-derived neurotrophic factor signal transduction and modulation. *J. Neurosci. Res.* 68, 668–680. doi:10.1002/jnr.10259

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell.* 150, 1274–1286. doi:10.1016/j.cell.2012.04.040

Network, C. G. A., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi:10.1038/nature11412

Nueda, M. J., Ferrer, A., and Conesa, A. (2012). Arsyn: A method for the identification and removal of systematic noise in multifactorial time course microarray experiments. *Biostatistics* 13, 553–566. doi:10.1093/biostatistics/kxr042

Ochoa, S., de Anda-Jáuregui, G., and Hernández-Lemus, E. (2021). An information theoretical multilayer network approach to breast cancer transcriptional regulation. *Front. Genet.* 12, 617512. doi:10.3389/fgene.2021.617512

Pathak, B. R., Breed, A. A., Apte, S., Acharya, K., and Mahale, S. D. (2016). Cysteine-rich secretory protein 3 plays a role in prostate cancer cell invasion and affects expression of psa and anxa1. *Mol. Cell. Biochem.* 411, 11–21. doi:10.1007/s11010-015-2564-2

Patra, S. K. (2008). Ras regulation of dna-methylation and cancer. *Exp. Cell. Res.* 314, 1193–1201. doi:10.1016/j.yexcr.2008.01.012

Pethel, S. D., and Hahs, D. W. (2014). Exact test of independence using mutual information. *Entropy* 16, 2839–2849. doi:10.3390/e16052839

Piao, H.-l., Yuan, Y., Wang, M., Sun, Y., Liang, H., and Ma, L. (2014). α-catenin acts as a tumour suppressor in e-cadherin-negative basal-like breast cancer by inhibiting nf-κb signalling. *Nat. Cell. Biol.* 16, 245–254. doi:10.1038/ncb2909

Privat, M., Rudewicz, J., Sonnier, N., Tamisier, C., Ponelle-Chachuat, F., and Bignon, Y.-J. (2018). Antioxydation and cell migration genes are identified as potential therapeutic targets in basal-like and brca1 mutated breast cancer cell lines. *Int. J. Med. Sci.* 15, 46–58. doi:10.7150/ijms.20508

Pupa, S. M., Ligorio, F., Cancila, V., Franceschini, A., Tripodo, C., Vernieri, C., et al. (2021). Her2 signaling and breast cancer stem cells: The bridge behind her2-positive breast cancer aggressiveness and therapy refractoriness. *Cancers* 13, 4778. doi:10.3390/cancers13194778

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramdzan, Z. M., Vadnais, C., Pal, R., Vandal, G., Cadieux, C., Leduy, L., et al. (2014). Ras transformation requires cux1-dependent repair of oxidative dna damage. *PLoS Biol.* 12, e1001807. doi:10.1371/journal.pbio.1001807

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-seq data. *BMC Bioinforma.* 12, 480. doi:10.1186/1471-2105-12-480

Rohart, F., Gautier, B., Singh, A., and Le Cao, K.-A. (2017). mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* 13, e1005752. doi:10.1371/journal.pcbi.1005752

Rose, N. R., and Klose, R. J. (2014). Understanding the relationship between dna methylation and histone lysine methylation. *Biochimica Biophysica Acta (BBA)-Gene Regul. Mech.* 1839, 1362–1372. doi:10.1016/j.bbagrm.2014.02.007

Ru, Y., Kechris, K. J., Tabakoff, B., Hoffman, P., Radcliffe, R. A., Bowler, R., et al. (2014). The multimir r package and database: Integration of microrna–target interactions along with their disease and drug associations. *Nucleic acids Res.* 42, e133. doi:10.1093/nar/gku631

Schulz, D. M., Bollner, C., Thomas, G., Atkinson, M., Esposito, I., Hofler, H., et al. (2009). Identification of differentially expressed proteins in triple-negative breast carcinomas using dige and mass spectrometry. *J. proteome Res.* 8, 3430–3438. doi:10.1021/pr900071h

Shamay, M., Greenway, M., Liao, G., Ambinder, R. F., and Hayward, S. D. (2010). De novo dna methyltransferase dnmt3b interacts with nedd8-modified proteins. *J. Biol. Chem.* 285, 36377–36386. doi:10.1074/jbc.M110.155721

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Sohn, K.-A., Kim, D., Lim, J., and Kim, J. H. (2013). Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors. *BMC Syst. Biol.* 7, S9. doi:10.1186/1752-0509-7-S6-S9

Sonawane, A. R., DeMeo, D. L., Quackenbush, J., and Glass, K. (2021). Constructing gene regulatory networks using epigenetic data. *npj Syst. Biol. Appl.* 7, 45–13. doi:10.1038/s41540-021-00208-3

Strelnikov, V. V., Kuznetsova, E. B., Tanas, A. S., Rudenko, V. V., Kalinkin, A. I., Poddubskaya, E. V., et al. (2021). Abnormal promoter dna hypermethylation of the integrin, nidogen, and dystroglycan genes in breast cancer. *Sci. Rep.* 11, 2264–2314. doi:10.1038/s41598-021-81851-y

Sugiyama, A., Okada, M., and Yamawaki, H. (2020). Canstatin suppresses isoproterenol-induced cardiac hypertrophy through inhibition of calcineurin/nuclear factor of activated t-cells pathway in rats. *Eur. J. Pharmacol.* 871, 172849. doi:10.1016/j.ejphar.2019.172849

Sumi, E., Iehara, N., Akiyama, H., Matsubara, T., Mima, A., Kanamori, H., et al. (2007). Sry-related hmg box 9 regulates the expression of col4a2 through transactivating its enhancer element in mesangial cells. *Am. J. pathology* 170, 1854–1864. doi:10.2353/ajpath.2007.060899

Tam, S., Tsao, M.-S., and McPherson, J. D. (2015). Optimization of mirna-seq data preprocessing. *Briefings Bioinforma.* 16, 950–963. doi:10.1093/bib/bbv019

Tan, Z., Niu, B., Tsang, K. Y., Melhado, I. G., Ohba, S., He, X., et al. (2018). Synergistic co-regulation and competition by a sox9-gli-foxa phasic transcriptional network coordinate chondrocyte differentiation transitions. *PLoS Genet.* 14, e1007346. doi:10.1371/journal.pgen.1007346

Tapia-Carrillo, D., Tovar, H., Velazquez-Caldelas, T. E., and Hernandez-Lemus, E. (2019). Master regulators of signaling pathways: An application to the analysis of gene regulation in breast cancer. *Front. Genet.* 10, 1180. doi:10.3389/fgene.2019.01180

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., et al. (2015). Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package. *Nucleic acids Res.* 43, e140. doi:10.1093/nar/gkv711

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics* 15, 569–583. doi:10.1093/biostatistics/kxu001

Thammaiah, C. K., and Jayaram, S. (2016). Role of let-7 family microrna in breast cancer. *Non-coding RNA Res.* 1, 77–82. doi:10.1016/j.ncrna.2016.10.003

Vidal Ocabo, E., Sayols, S., Moran, S., Guillaumet-Adkins, A., Schroeder, M. P., Royo, R., et al. (2017). A dna methylation map of human cancer at single base-pair resolution. *Oncogene* 36 (40), 5648–5657. doi:10.1038/onc.2017.176

Weigel Muñoz, M., Carvajal, G., Curci, L., Gonzalez, S. N., and Cuasnicu, P. S. (2019). Relevance of crisp proteins for epididymal physiology, fertilization, and fertility. *Andrology* 7, 610–617. doi:10.1111/andr.12638

Weighill, D., Burkholz, R., Guebila, M. B., Zacharias, H. U., Quackenbush, J., and Altenbuchinger, M. (2021). DRAGON: Determining regulatory associations using graphical models on multi-omic networks. *Oxford, England: Nucleic Acids Res.* [Epub ahead of print]. doi:10.1093/nar/gkac1157

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Wollen, E. J., Sejersted, Y., Wright, M. S., Bik-Multanowski, M., Madetko-Talowska, A., Günther, C.-C., et al. (2013). Transcriptome profiling of the newborn mouse lung after hypoxia and reoxygenation: Hyperoxic reoxygenation affects mtor signaling pathway, dna repair, and jnk-pathway regulation. *Pediatr. Res.* 74, 536–544. doi:10.1038/pr.2013.140

Wright, K. L., Adams, J. R., Liu, J. C., Loch, A. J., Wong, R. G., Jo, C. E., et al. (2015). Ras signaling is a key determinant for metastatic dissemination and poor survival of luminal breast cancer patients. *Cancer Res.* 75, 4960–4972. doi:10.1158/0008-5472.CAN-14-2992

Wu, J., Wang, C., Miao, X., Wu, Y., Yuan, J., Ding, M., et al. (2017). Age-related insulin-like growth factor binding protein-4 overexpression inhibits osteogenic differentiation of rat mesenchymal stem cells. *Cell. Physiology Biochem.* 42, 640–650. doi:10.1159/000477873

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). Clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* 2, 100141. doi:10.1016/j.xinn.2021.100141

Xin, S., Fang, W., Li, J., Li, D., Wang, C., Huang, Q., et al. (2021). Impact of stat1 polymorphisms on crizotinib-induced hepatotoxicity in alk-positive non-small cell lung cancer patients. *J. Cancer Res. Clin. Oncol.* 147, 725–737. doi:10.1007/s00432-020-03476-4

Xiong, B., Lei, X., Zhang, L., and Fu, J. (2017). mir-103 regulates triple negative breast cancer cells migration and invasion through targeting olfactomedin 4. *Biomed. Pharmacother.* 89, 1401–1408. doi:10.1016/j.biopha.2017.02.028

Xu, N., Wu, Y.-P., Ke, Z.-B., Liang, Y.-C., Cai, H., Su, W.-T., et al. (2019). Identification of key dna methylation-driven genes in prostate adenocarcinoma: An integrative analysis of tcga methylation data. *J. Transl. Med.* 17, 311–315. doi:10.1186/s12967-019-2065-2

Xu, X., Zhang, M., Xu, F., and Jiang, S. (2020). Wnt signaling in breast cancer: Biological mechanisms, challenges and opportunities. *Mol. cancer* 19, 165–235. doi:10.1186/s12943-020-01276-5

Zamora-Fuentes, J. M., Hernández-Lemus, E., and Espinal-Enríquez, J. (2022). Oncogenic role of mir-217 during clear cell renal carcinoma progression. *Front. Oncol.* 12, 934711. doi:10.3389/fonc.2022.934711

Zhang, J., and Huang, K. (2017). Pan-cancer analysis of frequent dna co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers. *Bmc Genomics* 18, 1045–1114. doi:10.1186/s12864-016-3259-0

Zhang, P., Wen, X., Gu, F., Deng, X., Li, J., Dong, J., et al. (2013). Methylation profiling of serum dna from hepatocellular carcinoma patients using an infinium human methylation 450 beadchip. *Hepatol. Int.* 7, 893–900. doi:10.1007/s12072-013-9437-0

Zhao, S., Geybels, M. S., Leonardson, A., Rubicz, R., Kolb, S., Yan, Q., et al. (2017). Epigenome-Wide tumor DNA methylation profiling identifies novel prognostic biomarkers of metastatic-lethal progression in men diagnosed with clinically localized prostate cancer. *Clin. Cancer Res.* 23, 311–319. doi:10.1158/1078-0432.CCR-16-0549

Zheng, G., Tu, K., Yang, Q., Xiong, Y., Wei, C., Xie, L., et al. (2008). Itfp: An integrated platform of mammalian transcription factors. *Bioinformatics* 24, 2416–2417. doi:10.1093/bioinformatics/btn439

frontiers | Frontiers in Genetics

TYPE Methods
PUBLISHED 10 January 2023
DOI 10.3389/fgene.2022.1076440

OPEN ACCESS

EDITED BY
Angelo Facchiano,
National Research Council (CNR), Italy

REVIEWED BY
Wenan Chen,
St. Jude Children's Research Hospital, United States
Tianyuan Lu,
McGill University, Canada

*CORRESPONDENCE
Hannah Klinkhammer,
klinkhammer@imbie.uni-bonn.de

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 21 October 2022
ACCEPTED 20 December 2022
PUBLISHED 10 January 2023

CITATION
Klinkhammer H, Staerk C, Maj C, Krawitz PM and Mayr A (2023), A statistical boosting framework for polygenic risk scores based on large-scale genotype data.
Front. Genet. 13:1076440.
doi: 10.3389/fgene.2022.1076440

A statistical boosting framework for polygenic risk scores based on large-scale genotype data

Hannah Klinkhammer[1,2]*, Christian Staerk[1], Carlo Maj[2,3], Peter Michael Krawitz[2] and Andreas Mayr[1]

[1]Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Bonn, Germany, [2]Institute for Genomic Statistics and Bioinformatics, Medical Faculty, University of Bonn, Bonn, Germany, [3]Center for Human Genetics, University of Marburg, Marburg, Germany

Polygenic risk scores (PRS) evaluate the individual genetic liability to a certain trait and are expected to play an increasingly important role in clinical risk stratification. Most often, PRS are estimated based on summary statistics of univariate effects derived from genome-wide association studies. To improve the predictive performance of PRS, it is desirable to fit multivariable models directly on the genetic data. Due to the large and high-dimensional data, a direct application of existing methods is often not feasible and new efficient algorithms are required to overcome the computational burden regarding efficiency and memory demands. We develop an adapted component-wise L_2-boosting algorithm to fit genotype data from large cohort studies to continuous outcomes using linear base-learners for the genetic variants. Similar to the snpnet approach implementing lasso regression, the proposed snpboost approach iteratively works on smaller batches of variants. By restricting the set of possible base-learners in each boosting step to variants most correlated with the residuals from previous iterations, the computational efficiency can be substantially increased without losing prediction accuracy. Furthermore, for large-scale data based on various traits from the UK Biobank we show that our method yields competitive prediction accuracy and computational efficiency compared to the snpnet approach and further commonly used methods. Due to the modular structure of boosting, our framework can be further extended to construct PRS for different outcome data and effect types—we illustrate this for the prediction of binary traits.

KEYWORDS

polygenic risk score (PRS), high-dimensional data, variable selection, boosting, GWAS—genome-wide association study, prediction

1 Introduction

In times of next-generation sequencing and decreasing costs for whole genome sequencing, the amount of available genotype data has increased dramatically in recent years, giving rise to new genetic insights (Beesley et al., 2020; National Human Genome Research Institute, 2021).

Polygenic risk scores (PRS) measure the individual genetic liability to a certain trait and can provide relevant information in the context of disease-risk stratification. In contrast to high-impact monogenic variants, which are mostly rare and have a high effect size, PRS are derived from common variants such as single-nucleotide polymorphisms (SNPs) with low or medium effect sizes. Polygenic effects could also explain part of the incomplete penetrance seen in many identified monogenic variants, as for example in the genes BRCA1 and BRCA2 both leading to a highly increased risk of breast cancer (Kuchenbaecker et al., 2017). Recent studies on the UK Biobank suggest that high-impact monogenic variants, PRS and family history could contribute

additively to the risk of developing breast and prostate cancer (Hassanin et al., 2021). Despite these findings, PRS still lack to explain relevant parts of the estimated heritability of many traits.

PRS are typically derived as a sum of risk allele counts weighted by univariate effect estimates of the measured variants based on summary statistics from genome-wide association studies (GWAS) (Choi et al., 2020). Despite several approaches to account for linkage disequilibrium (LD, referring to the correlation structure between variants) and for the selection of informative variants (Euesden et al., 2014; Vilhjálmsson et al., 2015; Mak et al., 2017; Privé et al., 2021), the univariate structure of the estimation cannot fully account for interdependencies between the variants. For example, lassosum (Mak et al., 2017) adopts an L_1 penalty term and solves a lasso-like problem while only using summary statistics and a LD reference panel. However, as published summary statistics and LD reference panels are most often based on different samples, lassosum can generally only approximate the full lasso path. A natural extension of using effect estimates from univariate models could hence be to fit a single multivariable model. While this approach seems natural from a methodological perspective, a direct application of existing methods is typically infeasible due to the high dimensionality of the genotype data, which can easily exceed the available computer memory. Recently, some approaches have been proposed to overcome this computational burden (Privé et al., 2018; Qian et al., 2020; Maj et al., 2022). In particular, Qian et al. proposed the so-called batch screening iterative lasso (BASIL) algorithm to fit the lasso on the complete original genotype data (Tibshirani, 1996; Qian et al., 2020; Li et al., 2022). The algorithm works on subsets of variants and computes the complete lasso path in an iterative fashion. Apart from the lasso, the algorithm can also be extended to other penalized regression methods such as the relaxed lasso (Meinshausen, 2007) or the elastic net (Zou and Hastie, 2005). In this context, Qian et al. were able to demonstrate that multivariable regularized PRS models fitted *via* the BASIL algorithm outperform the classical GWAS-based PRS for various traits such as height and high cholesterol.

While penalized regression models like the lasso and the elastic net impose explicit regularization, statistical boosting represents an alternative approach by introducing an implicit algorithmic regularization when combined with early stopping (Bühlmann and Hothorn, 2007; Mayr and Hofner, 2018). Boosting algorithms iteratively fit pre-defined base-learners to the gradient of the loss function, selecting the most influential base-learner in each step. The main tuning parameter of boosting algorithms is the number of iterations, which enables implicit variable selection and leads to sparse models. Due to its modular structure, boosting allows to combine possible base-learners with any convex loss function. These algorithms hence offer a great flexibility for statistical modelling, including various response types and the estimation of non-linear or other types of effects. A recent work has incorporated boosting into PRS modelling *via* a three-step approach (Maj et al., 2022): First, a marginal screening approach was applied on all variants to identify potentially informative ones. Then, multivariable algorithms including probing with boosting (Thomas et al., 2017) were applied on blocks of variants in LD to select ("fine-map") the most informative variants. Finally, a statistical boosting model was fitted on the variant set created by joining the selected variants of all chunks. This approach yielded particularly sparse and interpretable models, whose predictive performance was superior to PRS derived by univariate methods like clumping and thresholding (Euesden et al., 2014) and was outperformed by the predictive performance achieved by the lasso *via* the BASIL algorithm. However this approach includes pre-filtering of the variants and is computationally demanding.

In this article we introduce a new framework to boost PRS, starting with a new computational approach to build L_2-boosting models on large-scale genotype data for quantitative traits. Similar to the snpnet approach for the lasso, our algorithm iteratively works on smaller batches of variants. Yet, in contrast to recent boosting methods (Staerk and Mayr, 2021; Maj et al., 2022), the variants do not need to be pre-filtered in our snpboost approach and the batches are not pre-defined or randomly sampled, but chosen iteratively and deterministically in a data-driven way based on the correlations of the variants to the remaining residuals. By restricting the set of available base-learners in each step to those variants which were most correlated with residuals from a previous iteration, we are able to reduce the search space and decrease the computational time compared to a classical component-wise boosting algorithm.

We conducted a simulation study to examine the performance of our adapted boosting algorithm snpboost compared to the original L_2-boosting on a reduced but still high-dimensional data set, on which the application of standard L_2-boosting was still computationally feasible. Furthermore, we simulated data of higher dimensionality and larger sample size to investigate the influence of various hyperparameters (including the batch size) on the prediction accuracy and computational burden of the snpboost approach in a typical large-scale setting. We discuss reasonable default values for the hyperparameters which are incorporated in the provided R implementation (https://github.com/hklinkhammer/snpboost). Finally, we constructed multivariable PRS for various traits on data from the UK Biobank *via* application of snpboost and compared the performance of our approach to the lasso estimates from the BASIL algorithm proposed by Qian et al. as well as to further commonly used methods. On the examined phenotypes we found highly comparable predictive performance while our adapted boosting approach had a tendency to select sparser models compared to the lasso and the other methods. Finally, we illustrate how the framework can be conveniently extended to the classification of binary phenotypes by the incorporation of different loss functions.

2 Methods

For $n \in \mathbb{N}$ individuals, let $\mathbf{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ denote a particular continuous phenotype of interest. Furthermore, let X_j correspond to the genetic variant j, for $j = 1, \ldots, p$. The observed dosage data of n individuals is given in the genotype matrix $\mathbf{X} = (x_{i,j}) \in [0, 2]^{n \times p}$, where $\mathbf{x}_j \in [0, 2]^n$ corresponds to the jth column of $\mathbf{X}$. We consider a linear regression model

$$\mathbb{E}(y_i|\mathbf{X}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}, \quad i = 1, \ldots, n, \tag{1}$$

With coefficients $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$. The aim is to determine coefficients $\hat{\beta}_0 \in \mathbb{R}$ and $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ such that the estimator $\hat{\mathbf{y}} = \hat{\beta}_0 + \mathbf{X}\hat{\boldsymbol{\beta}}$ minimizes the mean squared error of prediction on an independent test set MSEP $= \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\hat{y}_{\text{test},i} - y_{\text{test},i})^2$. Additionally, one is often interested in relatively sparse models in the sense that only a fraction of the coefficient vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is non-zero.

In high-dimensional settings with $p > n$ it is not feasible to apply classical estimation techniques like the ordinary least squares estimator. A commonly-used solution is to consider further constraints on the coefficient vector resulting in penalized regression methods including the lasso (Tibshirani, 1996). The lasso incorporates an L_1-penalty on the coefficient vector such that the lasso estimate $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ is given by

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \text{argmin}_{(\beta_0,\boldsymbol{\beta})}\left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{i,j}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right] \quad (2)$$

for some $\lambda \geq 0$. The explicit L_1-penalization of the coefficient vector leads to shrinkage of the coefficient estimates. In contrast to ridge regression (Hoerl and Kennard, 2000), the use of the L_1-penalty enables to set some parameters exactly to zero corresponding to sparse models. There has been extensive research on the theoretical properties of the lasso including oracle inequalities in high-dimensional settings (e.g., Fu and Knight (2000); Greenshtein and Ritov (2004); Bunea et al. (2007); van de Geer (2008)). Nevertheless, there are situations leading to variable selection problems of the lasso, particularly in the presence of high correlations between signal and noise variables (Hepp et al., 2016). When working with genotype data, high correlations between signal and noise variables might often be present as a result of LD, i.e., genetic variants that have close positions on the DNA strand tend to be highly correlated.

An alternative to explicitly penalized regression methods such as the lasso is statistical gradient boosting (Bühlmann and Hothorn, 2007; Mayr and Hofner, 2018). Gradient boosting requires the specification of a loss function $f(\boldsymbol{y}, \hat{\boldsymbol{y}})$ and so-called base-learners h_j that are iteratively fitted to the response. In detail, the aim is again to fit the linear regression model (1) which is performed in an iterative fashion. Starting at iteration $m = 0$ with a starting value $\hat{\boldsymbol{y}}^{(0)} = \mathbf{0}$, the following steps are repeated until a maximum number m_{stop} of boosting iterations is reached (Bühlmann and Hothorn, 2007):

1. Set $m := m+1$ and compute the negative gradient vector of the loss function:

$$\boldsymbol{u}^{(m)} = -\frac{\partial f(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \hat{\boldsymbol{y}}}\bigg|_{\hat{\boldsymbol{y}}=\hat{\boldsymbol{y}}^{(m-1)}}$$

2. Fit every base-learner h_j separately to the negative gradient vector $\boldsymbol{u}^{(m)}$ and select the best fitting base-learner $\hat{h}_{j^*}^{(m)}(X_j)$.
3. Update the predictor with the learning rate $0 \leq \nu \leq 1$: $\hat{\boldsymbol{y}}^{(m)} = \hat{\boldsymbol{y}}^{(m-1)} + \nu\hat{h}_{j^*}^{(m)}(X_j)$
4. Stop if $m = m_{\text{stop}}$.

Stopping the algorithm before it converges (early stopping) leads to implicit regularization and shrinkage of effect estimates. The component-wise L_2-boosting algorithm (Bühlmann and Yu, 2003; Bühlmann and Hothorn, 2007) employs the squared error $f(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2$ as a loss function (Bühlmann and Yu, 2003) and separate univariate linear regression models of the residuals on the jth genetic variant as base-learners (i.e., $h_j(X_j) = \beta_0 + \beta_j X_j$, for $j = 1, \ldots, p$). In low-dimensional ($p < n$) settings this set-up mimics a classical Gaussian linear model and converges to the least squares solution for large values of m_{stop}. The general boosting procedure can be interpreted as gradient descent in function space, where the residual vector represents the gradient of the L_2 loss and the function space is provided by the different base-learner solutions (Friedman, 2001; Bühlmann and Yu, 2003; Mayr and Hofner, 2018). The previously described steps transform therefore into the following procedure (shown in grey in Figure 1): The best fitting base-learner in boosting step m+1 corresponds to the variant j^* with the highest Pearson correlation $\rho(\boldsymbol{x}_{j^*}, \boldsymbol{r}^{(m)})$ to the residuals $\boldsymbol{r}^{(m)} = \boldsymbol{y} - \hat{\boldsymbol{y}}^{(m)}$ resulting from the previous boosting step m. We then fit a linear regression model of the current residuals $\boldsymbol{r}^{(m)}$ on the variant j^* and update the corresponding coefficient $\hat{\beta}_{j^*}^{(m+1)}$ as well as the intercept $\hat{\beta}_0^{(m+1)}$. This is repeated until a maximum number of boosting iterations is reached or any other early stopping criterion is fulfilled. If additional covariates apart from the genetic variants are included in the model, they are treated as mandatory covariates—similar to the intercept. The additional covariates are included in each single base-learner and are hence updated in each boosting step without competing with the genetic variants.

Hepp et al. (2016) investigated the commonalities and differences between the lasso and statistical boosting: while there are (low-dimensional) settings in which the gradient boosting approximates the lasso coefficient paths arbitrarily close when the learning rate ν is approaching 0, their results generally differ if the coefficient paths are not monotone. The authors note that, in contrast to the lasso which limits the sum of the absolute values of the coefficients for each penalty parameter λ separately, boosting limits the total L_1-arc-length of all coefficient curves (Hepp et al., 2016). Interpreting this as the total absolute distance "travelled" by all coefficients among the coefficient paths through the iterations $m = 1, \ldots, m_{\text{stop}}$, it becomes clear that the solution in a certain iteration depends on all previous solutions of the iterative algorithm. This might lead to more stable pathways particularly in settings with high correlations between independent variables, which is typical for genetic data. Hepp et al. conducted several numerical experiments including high-dimensional settings in which they found similar predictive performance of lasso and boosting. In detail, boosting tended to yield slightly better prediction results while the lasso tended to result in sparser models with faster computations. On the other hand, the boosting algorithm can be easily extended to different response types as well as to different effects, including non-linear and interaction effects. In terms of genetic data, interaction effects can be used to model and identify epistatic effects and gene-environment interactions.

When working on genetic data from large cohort studies we do not only face a high-dimensional setting with $p > n$ but also a large-scale setting with large sample sizes n *and* large numbers of variants p. Large-scale settings often lead to extended computational times as well as memory issues. To overcome these and apply statistical boosting on genotype data, we implemented an adapted component-wise L_2-boosting algorithm that is built on the snpnet framework (Qian et al., 2020) and works on batches of variants. To do so, we additionally incorporate a batch-building step before starting the boosting iterations (shown in blue in Figure 1). In this step we extract the p_{batch} variants ($p_{\text{batch}} \ll p$) with the highest correlation $\rho(\boldsymbol{x}_j, \boldsymbol{r}^{(m)})$ to the current residual vector and include them in the batch B_k. A maximum number of m_{batch} boosting iterations is performed on batch B_k before the next batch is built based on the correlations of all p variants to the updated residuals. In total, we fit a maximum of $b_{\max}$ batches or stop early if an early stopping criterion is fulfilled. The algorithm is summarised in Table 1 and Figure 1.

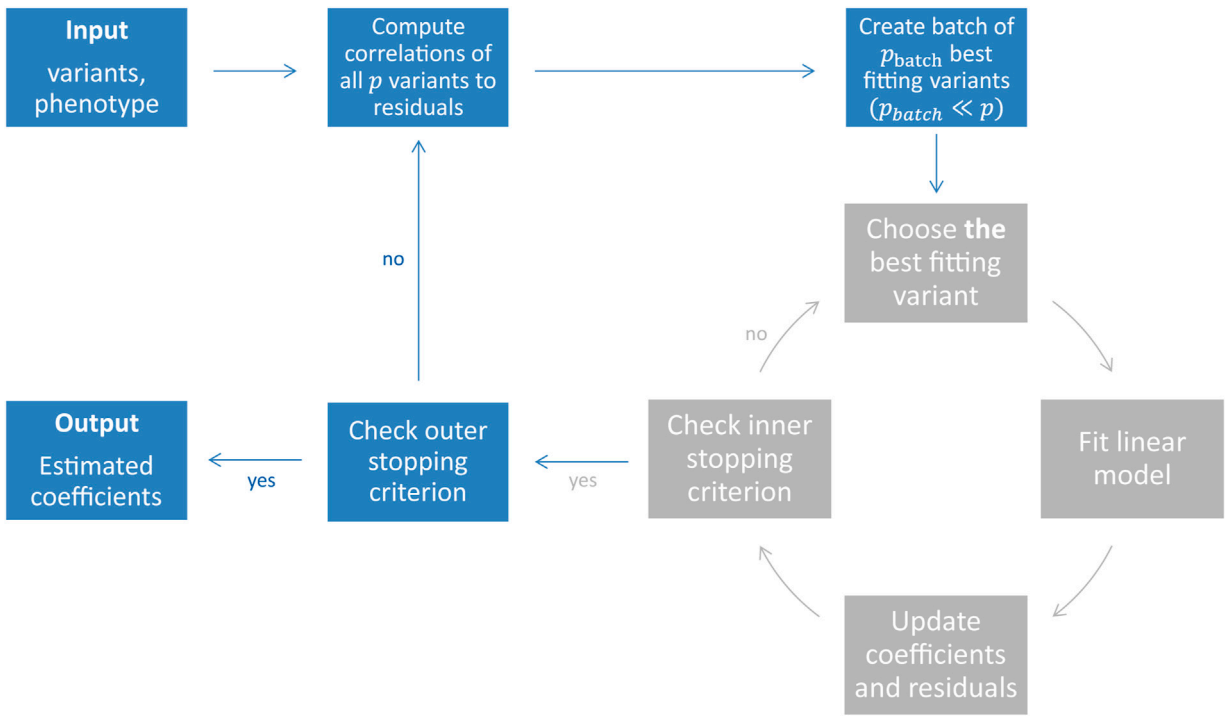


FIGURE 1
Illustration of the snpboost algorithm. The snpboost algorithm consists of an outer loop to create batches (shown in blue) and an inner loop representing the boosting on one batch (shown in grey).

TABLE 1 Definition of the snpboost algorithm without additional covariates. If additional covariates apart from the genetic variants should be included in the model, they are treated as mandatory covariates—similar to the intercept. The additional covariates are included in each single base-learner and are hence updated in each boosting step without competing with the genetic variants.

Algorithm: snpboost

Input: phenotype data $\mathbf{y} \in \mathbb{R}^n$, genotype data $\mathbf{X} \in [0,2]^{n \times p}$,
batch size $p_{\text{batch}} \in \{1, \ldots, p\}$,
learning rate $\nu > 0$,
maximum number of boosting iterations per batch $m_{\text{batch}} \in \mathbb{N}$,
maximum number of batches $b_{\max} \in \mathbb{N}$,
stopping lag for outer stopping criterion $b_{\text{stop}} \in \mathbb{N}$.

Algorithm:

1. **Initialization**: Set boosting index $m = 0$,
residuals $\mathbf{r}^{(0)} = \mathbf{y} - \bar{\mathbf{y}}$,
coefficients $\hat{\beta}_0^{(0)} = \bar{\mathbf{y}}$, $\hat{\beta}_j^{(0)} = 0$, $j = 1, \ldots, p$.
2. **Outer loop**: Set outer counter $k = 1$.
   a. Screening: Compute correlations $c_j^{(m)} = \rho(\mathbf{r}^{(m)}, \mathbf{x}_j)$, $j = 1, \ldots, p$.
   Create batch B_k of p_{batch} variants with highest absolute correlations $|c_j^{(m)}|$.
   Save the highest absolute correlation outside the batch $c_{\text{stop}} = \max_{j \notin B_k} |c_j^{(m)}|$.
   b. **Inner loop**: Set inner counter $l = 1$.
   (1) If $l > 1$, compute correlations inside batch: $c_j^{(m)} = \rho(\mathbf{r}^{(m)}, \mathbf{x}_j), \quad j \in B_k$
   (2) Choose variant j^* with the highest absolute correlation $|c_{j^*}^{(m)}| = \max_{j \in B_k} |c_j^{(m)}|$.
   If the current maximum absolute correlation inside the batch is smaller than the highest correlation outside the batch, i.e. if $|c_{j^*}^{(m)}| < c_{\text{stop}}$, stop the inner loop; else set $m := m + 1$.
   (3) Fit linear model: $\mathbb{E}(r^{(m-1)} \mid \mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_{j^*} X_{j^*}$
   (4) Update coefficients with learning rate ν and $\hat{\boldsymbol{\beta}}$ from iii.:
   $\hat{\beta}_0^{(m)} = \hat{\beta}_0^{(m-1)} + \nu \hat{\beta}_0, \quad \hat{\beta}_{j^*}^{(m)} = \hat{\beta}_{j^*}^{(m-1)} + \nu \hat{\beta}_{j^*},$
   $\hat{\beta}_j^{(m)} = \hat{\beta}_j^{(m-1)}, \quad j \in \{1, \ldots, p\} \backslash \{j^*\}$
   as well as residuals $\mathbf{r}^{(m)} = \mathbf{y} - \hat{\beta}_0^{(m)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(m)}$
   (5) If $l = m_{\text{batch}}$, end inner loop, else increase inner counter $l := l + 1$.
   c. If $k = b_{\max}$ or if the MSEP on the validation set has not decreased for b_{stop} batches, stop the outer loop; else increase the outer counter $k := k + 1$.
3. **Final model choice:** Find $m_{\text{stop}} \in \{1, \ldots, m\}$ corresponding to the lowest MSEP on the validation set. The final coefficient estimates are given by $\hat{\beta}_0^{(m_{\text{stop}})} \in \mathbb{R}$ and $\hat{\boldsymbol{\beta}}^{(m_{\text{stop}})} \in \mathbb{R}^p$.

By iteratively working on batches of variants we save computational time and memory because only parts of the variants have to be loaded into memory at once. Additionally, not every step requires the calculation of all potential base-learner solutions and the updated correlations for all variants. By this, we encourage additional sparsity by restricting the search space in terms of the set of available base-learners (as variants not included in the current batch cannot be selected). To examine when a new set of base-learners should be considered, which corresponds to the question when to stop the inner loop (inside the batches) and create a new batch of variants, we incorporated another step: we monitor the correlations of the variants inside a batch to the residuals and compare them to the correlations of variants outside of the batch. When creating a batch B_k we therefore compute and store the highest outer correlation $c_{\text{stop}} := \max_{j \notin B_k} |\rho(\mathbf{r}^{(m)}, \mathbf{x}_j)|$. After each boosting step m we check if the greatest absolute correlation of the variants inside the batch B_k to the current residual vector $\mathbf{r}^{(m)}$ is smaller than c_{stop}:

$$c_{\text{stop}} > \max_{j \in B_k} \left| \rho\left(\mathbf{r}^{(m)}, \mathbf{x}_j\right) \right|. \tag{3}$$

If inequality Eq. 3 holds true, we stop the inner loop and create a new batch since a variant outside the batch may provide a better fit to the current residual vector. In the original L_2-boosting without batches, the variant with the highest correlation to the residuals would be chosen in each boosting step. The incorporation of batches in general limits this choice to the variants inside the batch. However, the proposed stopping criterion provides an indication to consider variants outside the batch which may be higher correlated with the current residuals. Actually, if all variants were independent, the proposed stopping criterion would lead to the same choice of variants in each boosting step in snpboost as in the original L_2-boosting. Despite LD, our simulation results show that the

proposed stopping criterion yields reasonable variant choices and results in a competitive predictive performance (Section 3.1.1). Additionally, the inner loop is also stopped if the number of updates inside the batch reaches m_{batch}.

Furthermore, we need to determine after how many batches the algorithm should terminate. In classical statistical boosting the number of boosting iterations is often selected by cross-validation or resampling techniques—mimicking an additional data set to validate the predictive performance of the resulting models. However, if the data set is large enough, one can also directly divide the data into training and validation set. As in Qian et al. (2020), we hence simultaneously monitor the predictive performance of our model on an independent validation set while fitting on the training set. As a validation criterion for the predictive performance we use the MSEP on the validation set. The outer loop consisting of the batch-building step is stopped if the MSEP on the validation set has not decreased for b_{stop} batches or after a maximum number of b_{max} batches have been processed.

The proposed method is implemented as an add-on to the snpnet package by Qian et al. (2020) in the statistical computing environment R (R Core Team (2021), https://github.com/hklinkhammer/snpboost). While we are also incorporating PLINK 2.0 (Chang et al., 2015) to compute the correlations and build the batches in the outer loop, we replaced the fitting of the lasso by the adapted component-wise L_2-boosting algorithm on the resulting batches (Table 1; Figure 1).

3 Empirical results

3.1 Simulation study

We conducted a simulation study to investigate the behaviour of the proposed snpboost algorithm in various controlled data scenarios. The simulation study aims at two main goals: first, to examine potential differences in performance compared to the original component-wise L_2-boosting (Bühlmann and Yu, 2003) in smaller settings and, second, to gain insights on how to choose the included hyperparameters in practical situations.

Simulations are based on the UK Biobank genotype data (Bycroft et al., 2018) obtained under application number 81202 combined with simulated phenotypes. We restricted the individuals to white British ancestry and used the PLINK 2.0 function–thin-indiv-count to randomly sample n individuals, of which 50%, 20% and 30% were assigned to the training, validation and test set, respectively (Chang et al., 2015; Purcell and Chang, 2015). Then, p variants with minor allele frequency not less than 1% were randomly sampled using PLINK 2.0's–thin-count. Missing genotypes were replaced by the reference allele using the R package bigsnpr (Privé et al., 2018).

Continuous phenotypes were simulated from a linear model with Gaussian distributed noise and effect sizes using bigsnpr. To account for different genetic architectures, we considered varying heritability h^2 and sparsity s, defined as the amount of variance explained by the genetic liability and the proportion of causal variants, respectively. For each setting of h^2 and s, we simulated 100 different datasets. PRS models were derived by snpboost and evaluated by using various metrics regarding the predictive performance and the accuracy of the estimated coefficients. In detail, the predictive performance was measured by the MSEP and the R^2 value defined as the squared correlation between the predicted and the true phenotype on the independent test set. To assess the computational efficiency we measured the computation time of the algorithm. The accuracy of the resulting estimates was evaluated by the number of included variants in the final model and the mean squared error (MSE) of the estimates as well as the true positive (TP) rates and precision regarding variant selection. Additional results for all considered settings as well as comparisons to snpnet can be found in the Supplementary Material (Supplementary Figures S1–S6).

3.1.1 Comparison to original L_2-boosting in smaller settings

To analyse the performance of snpboost compared to the original component-wise L_2-boosting algorithm (Bühlmann and Yu, 2003), we used a single large batch with batch size $p_{batch} = p$ in the snpboost algorithm on simulated data with reduced dimensionality. We then compared the results to the ones derived by using smaller batches in terms of predictive performance, computation time, mean squared errors of the estimated coefficients as well as true positive rates and precision regarding variant selection. The simulations were conducted for n = 20,000 observations (10,000 training set, 4,000 validation set, 6,000 test set) and p = 20,000 variants as well as for varying degrees of heritability and sparsity. To obtain comparable results we chose a fixed number of boosting iterations independent of the batch size p_{batch} and a fixed learning rate $\nu = 0.1$. For each simulation, 10 CPUs with 1 GB memory each were used.

Figure 2 displays the boxplots of each metric obtained after 1,500 boosting iterations for heritability h^2 = 50% and sparsity s = 0.1%, i.e., 20 influential variants. Incorporating batches did not largely affect the predictive performance in terms of R^2 and MSEP nor the MSE of the coefficient estimates (MSE results not shown). However, different batch sizes do not always yield the same models as L_2-boosting as can be observed from the number of variants included in the final models. The models resulting from a batch size of p_{batch} = 1,000 tend to contain less variants than the ones from the original L_2-boosting (batch size p_{batch} = 20,000). This could be explained by the reduced search space in each boosting step and a trade-off between exploration (genome-wide search) and exploitation (search inside the batch). As a consequence, variants within the batch that are already in the model are more often updated instead of including new variants outside of the batch. The same holds true when comparing the number of chosen variants for batch size p_{batch} = 1,000 to smaller batch sizes (i.e., p_{batch} = 10 and p_{batch} = 100). As all models tend to overestimate the number of influential variants, the lower number of selected variants for batch size p_{batch} = 1,000 corresponds to a higher precision since less false positives are included. The fact that the other metrics remain almost constant suggests that either only variants with very small effects are not included when using a larger batch size or the variants that are updated are highly correlated with the ones not included. Furthermore, incorporating batches in the algorithm has a major effect on the computation time. To interpret the results shown in Figure 2 it is important to understand the two drivers of the computation time. On the one hand, it increases with the number of correlations that have to be calculated in each boosting step which explains the increased computation time of the original L_2-boosting (i.e., a batch size of p_{batch} = 20,000 and 20,000 computed correlations in each boosting step) compared to smaller batch sizes such as p_{batch} = 100 and p_{batch} = 1,000. On the other hand, reading the genotype data from disk when building the batches also increases the

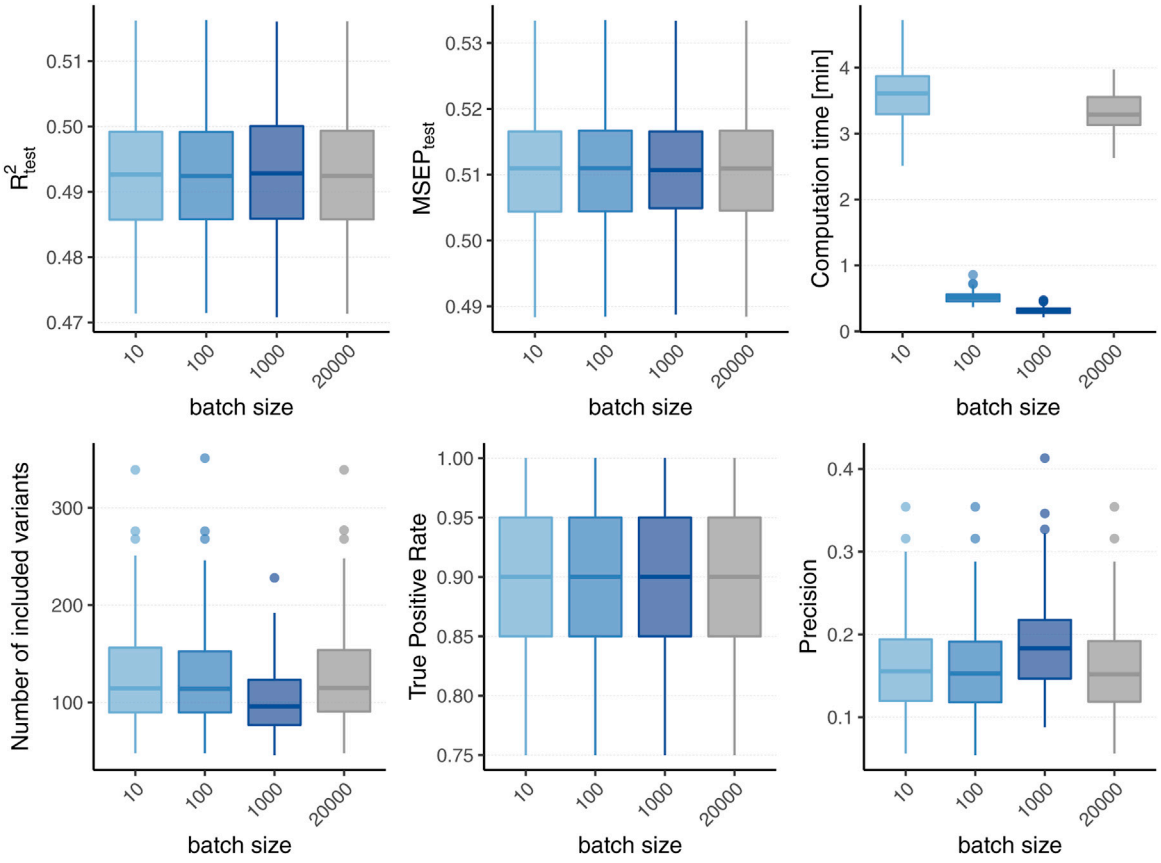


FIGURE 2
Comparison to original L_2-boosting. Results of 100 simulated phenotypes with heritability h^2 = 50% and sparsity s = 0.1% for p = 20,000 variants and n = 20,000 individuals (divided into 50% training, 20% validation and 30% test set). Boxplots of the evaluation metrics obtained after 1,500 boosting iterations are shown depending on the batch size. Batch size p_{batch} = 20,000 corresponds to the original L_2-boosting (shown in grey).

computation time leading to a higher computation time for smaller batches with p_{batch} = 10 for which more reads-from-disk have to be carried out. The varying computation times therefore reflect a trade-off between the number of correlations computed in each boosting step and the number of created batches.

In summary, the incorporation of batches in the boosting algorithm did not affect the predictive performance of the model in our scenarios, while computation time was substantially reduced. However, snpboost does not always lead to the same models as the original L_2-boosting algorithm, in particular in terms of the included variants and sparsity. The results for further settings with different heritability and sparsity were comparable and can be found in the Supplementary Material.

3.1.2 Choice of hyperparameters for large-scale applications

The proposed snpboost algorithm includes various hyperparameters, namely the batch size p_{batch}, the learning rate ν, the maximum number of boosting iterations per batch m_{batch}, the maximum number of processed batches b_{max} and the stopping lag for the outer early stopping criterion b_{stop}. In this section we discuss default values for the hyperparameters to facilitate the applicability of the algorithm in practice. The majority of these parameters do not need to be tuned but can be specified with reasonable default values, e.g., based on results from the literature and experience with the original boosting algorithm. For the remaining ones (p_{batch} and b_{stop}) we examine how they influence the computational and predictive performance of snpboost in a simulation study.

The choice of the learning rate ν can be leaned on widely-used boosting algorithms. A rather small learning rate prevents boosting algorithms from overfitting on single base-learners and is therefore favorable regarding predictive performance. Nevertheless, a smaller learning rate will increase the number of needed boosting iterations to fit the full effect of the base-learners and simultaneously increase the algorithm's computation time. Widely used R packages such as mboost (Bühlmann and Hothorn, 2007; Hothorn et al., 2010) and xgboost (Chen and Guestrin, 2016) use default learning rates of 0.1 and 0.3, respectively. As the effect of the learning rate will be comparable in the proposed adapted boosting algorithm, we decided to specify a fixed default value of $\nu = 0.1$ in all our simulations. For the batch-related hyperparameters we varied the batch size p_{batch} over a range of possible values namely $p_{batch} \in \{10, 100, 1{,}000, 5{,}000\}$ to analyse its effect. For each batch we allow a maximum number of boosting iterations m_{batch} equivalent to the batch size p_{batch}. Since we specified the learning rate with a rather small fixed value and due to the correlation-based early stopping criterion, this choice should prevent the algorithm from overfitting on one batch. If one or more variants inside the batch are still among the most influential ones out of all

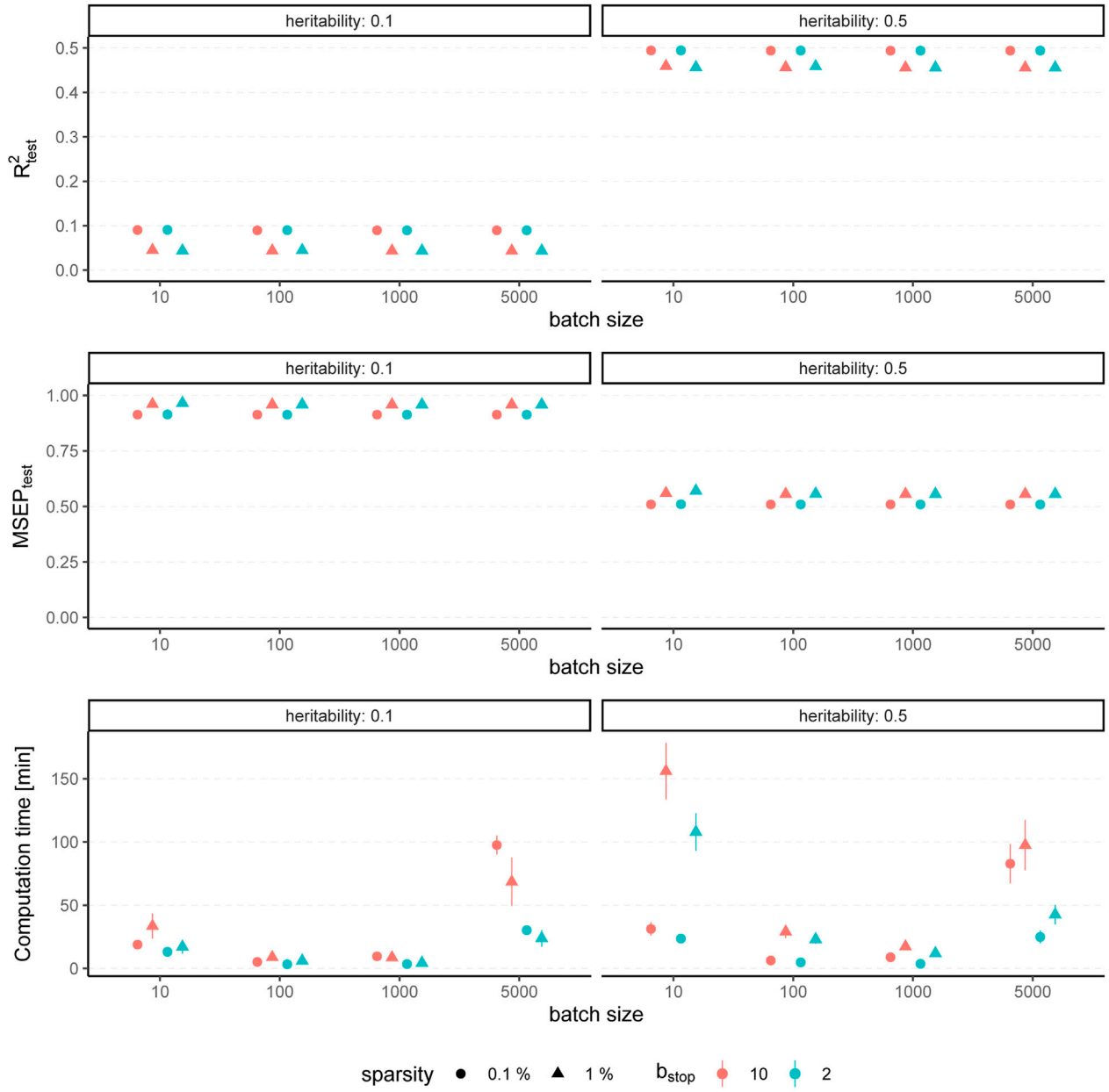


FIGURE 3
Predictive performance for varying batch size and stopping criteria. Results of 100 simulated phenotypes with heritability $h^2 \in \{10\%, 50\%\}$, sparsity $s \in \{0.1\%, 1\%\}$ and $b_{stop} \in \{2, 10\}$ for p = 100, 000 variants and n = 100, 000 individuals (divided into 50% training, 20% validation and 30% test set). Mean and standard deviation of the evaluation metrics are shown depending on the batch size.

variants they will also be included in the next batch. For the outer stopping criterion we specified a large maximum number of batches b_{max} = 20,000 to ensure that the algorithm terminates even in case the MSEP on the validation set has not decreased for b_{stop} batches. Since we do not want the algorithm to stop too early and simultaneously minimize the computation time, in our simulations we consider the choices b_{stop} = 2 and b_{stop} = 10. We then fitted PRS models using snpboost with the previously described hyperparameters. For the computations we used 10 CPUs with 2 GB RAM each.

The results for simulated phenotypes with 10% and 50% heritability are shown in Figure 3 and Figure 4. Results for further degrees of heritability can be found in the supplement. Independently of the heritability and the sparsity of the simulated data, the predictive performance was not affected in our settings by varying batch sizes in terms of R^2 and MSEP. However, the computation time differed crucially, resulting in considerably higher values for rather small (p_{batch} = 10) or rather large (p_{batch} = 5,000) batches. Furthermore, larger batches led to a higher number of included variants in the final model. This effect was stronger for phenotypes which have a less sparse genetic architecture and associated with a later stopping of the algorithm, i.e., more boosting steps were required to derive the final model. A higher number of variants in the final model was associated with a slightly higher MSE of the coefficients as well as higher true positive rates on the one hand but also smaller precision on the other

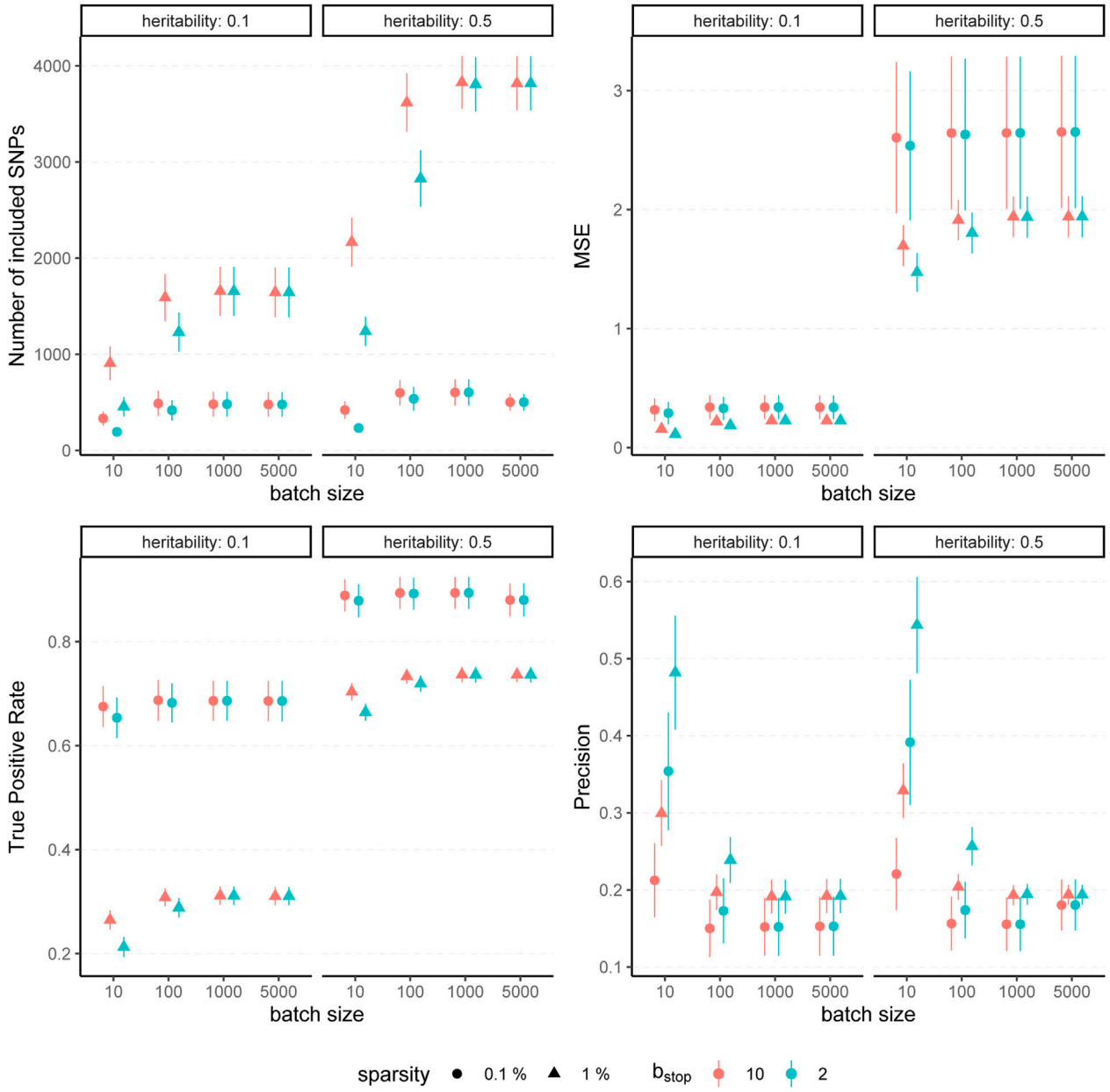


FIGURE 4
Evaluation metrics of the estimated coefficients for varying batch size and stopping criteria. Results of 100 simulated phenotypes with heritability $h^2 \in \{10\%, 50\%\}$, sparsity $s \in \{0.1\%, 1\%\}$ and $b_{stop} \in \{2, 10\}$ for $p = 100{,}000$ variants and $n = 100{,}000$ individuals (divided into 50% training, 20% validation and 30% test set). Mean and standard deviation of the evaluation metrics are shown depending on the batch size.

hand. As expected, a higher m_{stop} increased the computation time of the fitting process for all batch sizes. In contrast, there was no considerable effect on the predictive performance. However, $b_{stop} = 2$ and $b_{stop} = 10$ had an impact on the coefficient estimates as can be seen in Figure 4, e.g., by a tendency to include more variants in the model when choosing $b_{stop} = 10$. This tendency was only apparent for batch sizes $p_{batch} < 1{,}000$, suggesting that for larger batches the choice of b_{stop} is only of minor importance for both, prediction performance and coefficient estimates. The results clearly indicate that a more favorable signal-to-noise ratio (i.e., a higher heritability) and less influential variants (i.e., a higher sparsity) are in general beneficial for the performance of our approach. For phenotypes with a sparser genetic architecture, the considered evaluation metrics tended to show less variability.

In summary, the choice of the hyperparameters had no major influence on the predictive performance measures R^2 and MSEP but on the computation time, which was lowest for medium size batches ($100 \leq p_{batch} \leq 1{,}000$). The accuracy of the coefficient estimates measured *via* MSE, TP and TN rate varied with the batch size, as larger batches tended to lead to more (true positive) variants included in the final model, but also to a slightly higher MSE and a smaller TN rate. While the differences in MSE, TP, and TN rate were only small, smaller batches yielded sparser models in particular for phenotypes with a high heritability.

To conclude, batch sizes of $100 \leq p_{\text{batch}} \leq 1{,}000$ seem to be the most favorable regarding the computation time and the other evaluation metrics. We propose a batch size of $p_{\text{batch}} = 1{,}000$ as the default value because the results suggest less dependency on the b_{stop} parameter than for a batch size of 100 variants. Accordingly, we recommend a default value of $b_{\text{stop}} = 2$ to keep the computation time as low as possible. In practice, genotype data most often contain more than 100,000 variants, which further supports the choice of $p_{\text{batch}} = 1{,}000$ with regard to the computation time. Although our simulation study suggests that those default values should provide reasonable results in most cases, it is recommendable to take the genetic architecture of the examined phenotype as well as the main aim of the analysis into account. Phenotypes with a high expected heritability might be better fitted by using smaller batches, while for phenotypes with many causal variants larger batches might be favorable to increase the TP rate. If one is interested in extremely sparse models identifying only the most-informative variants one could also try to use smaller batches to avoid an overestimation of the number of causal variants.

3.2 Application to the UK Biobank

We applied our proposed method on data from the UK Biobank resource under Application Number 81202. Besides the validation of the results from the previous section, we compared our boosting models fitted *via* the proposed snpboost approach to the ones derived by fitting the lasso *via* the BASIL algorithm implemented in the snpnet package, which have already been shown to outperform commonly-used PRS models for various traits (Qian et al., 2020). Furthermore, we compared our results to PRScs (Ge et al., 2019), LDpred2 (Privé et al., 2021) and SBayesR (Lloyd-Jones et al., 2019), which are based on summary statistics, as well as to multivariable methods *via* LDAK (Zhang et al., 2021) based on Bolt-LMM (Loh et al., 2015), Ridge Regression (Henderson, 1950) and BayesR (Moser et al., 2015).The UK Biobank (UKBB, Bycroft et al., 2018) is a large-scale prospective cohort study including more than half a million participants from the United Kingdom aged between 40 and 69 years when recruited. The database comprises genome-wide genotype data of each individual as well as various in-depth phenotypic information such as biological measurements as well as blood and urine biomarkers. The data have been collected since 2006 and are continually updated with follow-up data.

Our aim is to estimate PRS for various phenotypes, covering several heritability and sparsity levels. The heritability of a trait is an upper bound for the predictive performance based on genotype information. Thus, we used the analyses of Tanigawa et al. (2022) as a proxy and specifically considered five appropriate continuous phenotypes: standing height in cm (UKBB field 50), LDL-cholesterol in mmol/l (UKBB field 30780), blood glucose level in mmol/L (UKBB field 30740), lipoprotein A in nmol/L (UKBB field 30790) and BMI in kg/m^2 (UKBB field 21001).

Height and BMI are quantitative traits with a relatively high heritability and a rather polygenic structure. Twin-studies estimated a heritability of approximately 69% for height and 42% for BMI after adjusting for covariates (Hemani et al., 2013). For a long time, genetic models could not explain this estimated heritability, a phenomenon known as "missing heritability" (Maher, 2008; Gibson, 2010). More recent studies have indicated that this may be primarily due to many influential common variants with small effect sizes (Yang et al., 2010; Wood et al., 2014; Yang et al., 2015) underlining the high polygenicity of those traits. In contrast to this, the distribution of the biomarker lipoprotein A, which is a strong risk factor for coronary heart disease, is mainly explained by variants in the LPA gene on chromosome 6 (Kronenberg and Utermann, 2013). Thus, we expect a sparse PRS with a relatively high prediction accuracy for this trait. For LDL-cholesterol it is known that it is associated with several genes such as LDLR and PCSK9 (Sanna et al., 2011; Sabatine, 2019). Therefore, we expect signal in several genomic regions. Recent studies compared different approaches including the lasso to derive PRS, and found that multivariable methods can reach a predictive performance of up to 20% (Maj et al., 2022; Tanigawa et al., 2022). As in previous works (Sinnott-Armstrong et al., 2021), we adjusted the measured LDL-cholesterol value by a factor of 0.684 for individuals taking statins lowering the blood lipid. For blood glucose we are not aware of a genetic impact and also Tanigawa et al. (2022) found the genetic background only explaining a small fraction (less than 2%) of the biomarker's variance.

Out of the over 500,000 individuals from UK Biobank we filtered for unrelated (based on UKBB resource 668) individuals with self-reported white British ancestry (UKBB field 21000) and available data for all chosen phenotypes, resulting in $n = 262{,}171$ observations. Additionally, the covariates age and sex as well as the first ten principal components of the genotype matrix are available. We randomly divided the data set into training ($n_{\text{train}} = 157{,}204$), validation ($n_{\text{val}} = 52{,}416$) and test set ($n_{\text{test}} = 52{,}551$). We used genome-wide genotype data and filtered for variants with a genotyped rate of at least 90% and a minor allele frequency of at least 0.1%, resulting in $p = 562{,}723$ genetic variants. Missing genotypes are imputed by the corresponding mean of the complete observations.

For both the boosting and lasso approaches, we first estimated a PRS using only the genotyped variants as predictors. We used the training set to fit the model and the validation set to simultaneously monitor the predictive performance for choosing the main tuning parameters of the algorithms (i.e., the number of iterations for boosting and the penalty parameter for the lasso). To fit the lasso we used the R package snpnet (Qian et al., 2020) with the provided default hyperparameters. Following the results of our simulation study, for the snpboost algorithm we chose a batch size of $p_{\text{batch}} = 1{,}000$ variants, a learning rate of $\nu = 0.1$ and an outer stopping lag of $b_{\text{stop}} = 2$ batches. Using the resulting estimated $\widehat{\text{PRS}}$ we fitted two linear models on the combined training and validation set, namely the first one (M_{PRS}) incorporating only the PRS as a single predictor variable:

$$M_{\text{PRS}}\colon Y = \gamma_0 + \gamma_{\text{PRS}}\widehat{\text{PRS}} \tag{4}$$

and the second one (M_f) including the first ten principal components, sex and age as additional covariates:

$$M_f\colon Y = \gamma_0 + \gamma_{\text{PRS}}\widehat{\text{PRS}} + \gamma_1\text{PC}_1 + \cdots + \gamma_{10}\text{PC}_{10} + \gamma_{\text{sex}}\text{sex} + \gamma_{\text{age}}\text{age}. \tag{5}$$

To measure the actual benefit in accuracy of including a PRS in the prediction model, we also fitted a model including only covariates (M_c):

$$M_c\colon Y = \gamma_0 + \gamma_1\text{PC}_1 + \cdots + \gamma_{10}\text{PC}_{10} + \gamma_{\text{sex}}\text{sex} + \gamma_{\text{age}}\text{age}. \tag{6}$$

Finally, we also included the covariates in the fitting process to derive the PRS, corresponding to the model $M_{PRS,c}$:

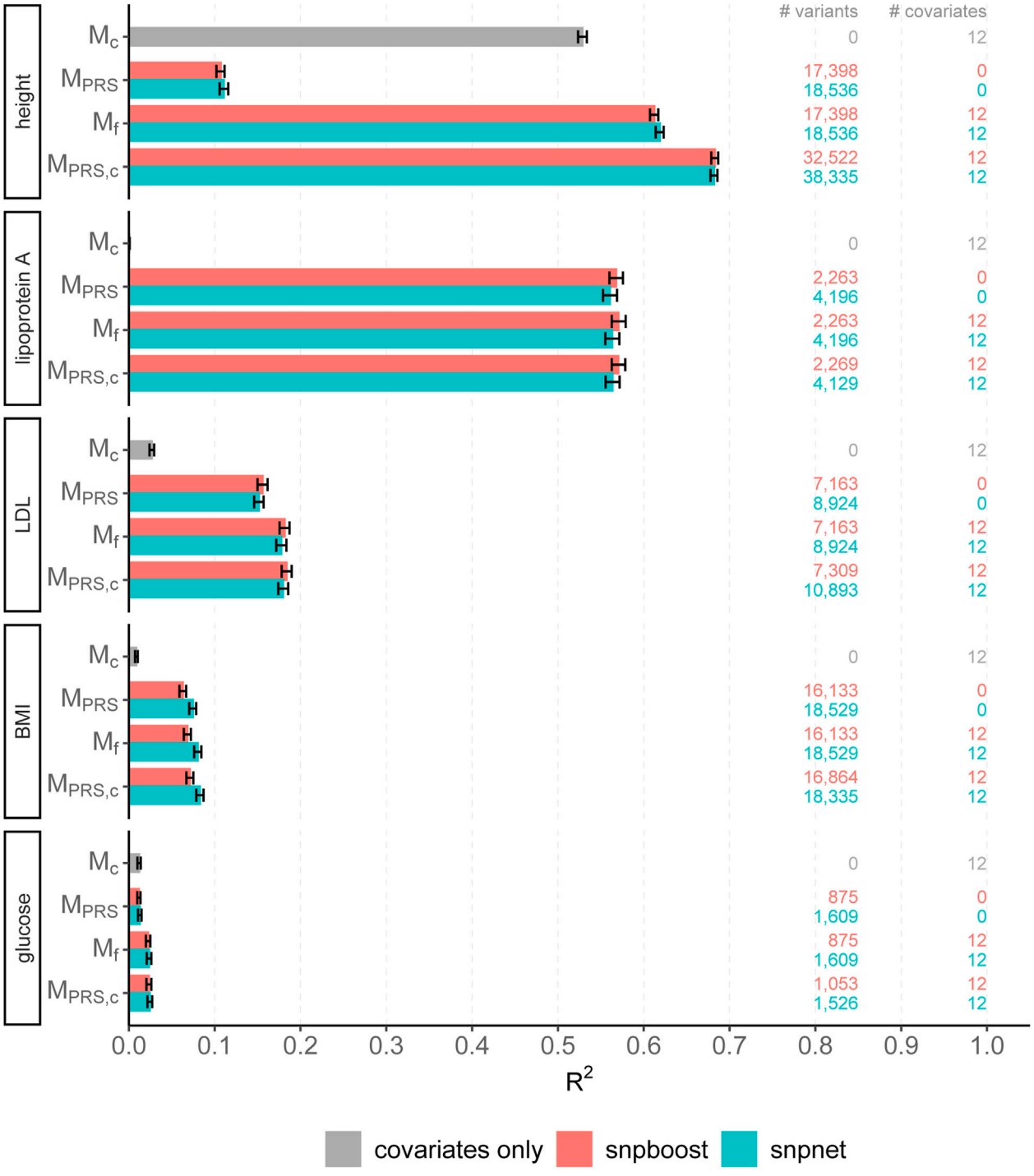


FIGURE 5
Comparison of predictive performance of snpnet and snpboost for five continuous phenotypes from the UKBB. Results of the covariate-only model (M_c; grey bars) and multivariable polygenic models with and without inclusion of the covariates derived by lasso (snpnet; petrol-colored bars) and statistical boosting (snpboost; red bars) for the prediction of five phenotypes from the UKBB. The barplots show the predictive performance (R^2) on the test set of 52,551 unrelated white British individuals. M_{PRS} corresponds to a linear model incorporating the PRS as a single predictor variable and M_f to a linear model incorporation sex, age and the first ten principal components as additional covariates. $M_{PRS,c}$ includes the covariates already in the fitting process of the PRS. Bootstrapped 95% confidence intervals are indicated by error bars. Furthermore, information on the number of selected genetic variants (# variants) and the number of additionally included covariates (# covariates) is given.

TABLE 2 Comparison of computational efficiency of snpnet and snpboost on eight phenotypes from the UKBB. Computational times of the algorithms snpnet and snpboost for multivariable polygenic models with and without inclusion of the covariates for the prediction of eight phenotypes from the UKBB. M_{PRS} corresponds to the application of the algorithms without including covariates and $M_{PRS,c}$ to the inclusion of the covariates sex, age and the first ten principal components. The experiments were run on 16 CPUs with 2 GB RAM each.

		Computation time in minutes	
Phenotype	Model	snpnet	snpboost
Height	M_{PRS}	132.44	116.65
Height	$M_{PRS,c}$	97.49	299.98
BMI	M_{PRS}	54.36	94.34
BMI	$M_{PRS,c}$	54.81	156.49
LDL	M_{PRS}	37.92	50.64
LDL	$M_{PRS,c}$	45.61	64.27
glucose	M_{PRS}	14.86	11.38
glucose	$M_{PRS,c}$	14.71	16.33
lipoprotein A	M_{PRS}	28.99	25.08
lipoprotein A	$M_{PRS,c}$	33.67	30.14
asthma	M_{PRS}	3.97	5.31
asthma	$M_{PRS,c}$	4.21	6.63
coeliac	M_{PRS}	3.11	5.46
coeliac	$M_{PRS,c}$	5.00	6.69
HBP	M_{PRS}	46.63	90.27
HBP	$M_{PRS,c}$	30.07	181.23

$$M_{PRS,c}: Y = \beta_0 + \beta_{PC1}\mathrm{PC}_1 + \cdots + \beta_{PC10}\mathrm{PC}_{10} + \beta_{sex}\mathrm{sex} + \beta_{age}\mathrm{age} + \sum_{j=1}^{P} \beta_j X_j. \quad (7)$$

All models were evaluated on the independent test set and compared with respect to their predictive performance, computational efficiency and sparsity. To measure the predictive performance we used the R^2 value on the test set given by the squared correlation between the observed and predicted phenotypes as well as the root mean squared error of prediction (RMSEP). The computational efficiency was measured as the computation time in minutes of the respective algorithm and the sparsity is given by the number of included variants in the final PRS. All computations were conducted on a computer cluster with 16 CPUs and 2 GB RAM each. The derivation of the PRS by the use of further methods (namely PRScs, LDpred2, SBayesR, Bolt-LMM, Ridge Regression and BayesR) was based on the same training and validation data and is described in the Supplementary Material. All models were tested on the same independent test set.

The results of snpboost as well as of snpnet for all phenotypes are given in Figure 5 and Table 2. The resulting RMSEP is shown in Supplementary Figure S7. Overall, snpnet and snpboost yield comparable results regarding the predictive performance, without one approach being consistently superior to the other. Both the resulting R^2 and RMSEP are very close. Furthermore, the shown R^2 values are in line with previously reported R^2 resulting from snpnet, which has been shown to be highly competitive to various other (univariate) PRS methods (Qian et al., 2020; Li et al., 2022; Tanigawa et al., 2022). The PRS estimated *via* snpnet and snpboost both clearly increase the predictive performance compared to the covariates-only model M_c for all shown phenotypes. With respect to sparsity, our boosting approach tends to select less variants (on average 26% less variants compared to the lasso). The computation time of both approaches is highly dependent on the genetic architecture, i.e., the heritability and sparsity of the phenotype. In particular, a higher and more polygenic signal tends to lead to longer computation times. In case of fitting the PRS based solely on the genotype data and including the covariates in a subsequent linear model, snpboost tends to be faster than snpnet; however, the computation times for snpboost increase substantially when including covariates in the fitting process for LDL-cholesterol and height. This is partly due to more coefficients being fitted and updated in each boosting step and partly due to larger PRS models resulting from more boosting steps. Nevertheless, the models are still fitted in reasonable time using our batch-based approach. As described in Hepp et al. (2016), boosting is generally expected to be slower than the lasso, which can only be observed for less sparse models in the examined phenotypes. In general, the model $M_{PRS,c}$ outperforms the model M_f regarding the predictive performance, implying that including the covariates already in the fitting of the PRS is favorable regarding the detection of the genetic signal. However, the effect is only substantial for phenotypes with a high association to covariates (i.e., height). Furthermore, the model $M_{PRS,c}$ tends to select more variables than estimating the PRS based only on the genotypes (M_{PRS}) and the computation time is considerably increased when using the snpboost approach. Therefore, it might be advisable to only consider the covariates in the fitting process if there is a large association already in the covariates-only model.

Figure 6 displays the absolute values of the resulting estimated non-zero coefficients for LDL-cholesterol for the boosting and lasso approaches. Both tend to detect variants with higher effect sizes in the same genetic regions, e.g., at chromosome 2 and chromosome 19. In total, there are 3,030 genetic variants that are present in both PRS, out of 7,163 variants selected by snpboost and 8,924 variants selected by snpnet. While snpboost results in less variants, the included variants have larger effect sizes and less variants with very small effect sizes close to zero are included in the model. Supplementary Figure S8 displays the coefficients again with shared variants marked in black. All SNPs with comparably high effect sizes in the snpnet PRS are included in both models but the snpboost PRS incorporates further SNPs with stronger effects. The results are similar for the other phenotypes and included in the Supplementary Material. In conclusion, the snpboost PRS tends to include less variants in total, but more variants with comparably high effect sizes corresponding to less shrinkage for the variants included in the model compared to the lasso.

The Supplementary Material comprises results for comparisons to further commonly used methods to derive PRS (Supplementary Tables S1, S2). Our proposed algorithm yielded consistently higher prediction performance compared to the summary statistics based PRScs and LDpred2 methods; furthermore, it yielded competitive results compared to summary statistics based SBayesR and four different multivariable approaches, while tending to select the sparsest models.

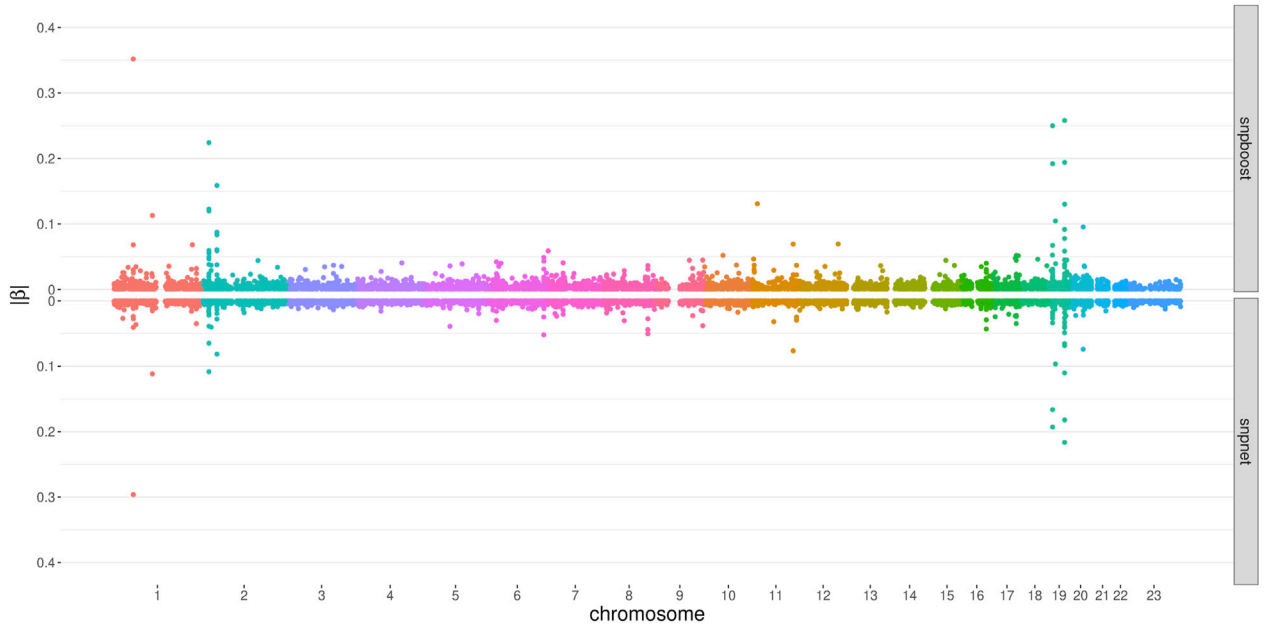


FIGURE 6
Absolute values of coefficient estimates for PRS models for LDL-cholesterol derived by boosting (snpboost) and lasso (snpnet) in dependence of the genomic position of the variants.

4 Extension to binary data

While traits like blood biomarkers or physical measurements are often quantitative, it is, for example, also of interest to predict the probability of the occurrence of a disease for a particular patient. In this case we deal with binary data $y_i \in \{0, 1\}$ and proceed as in a logistic regression by modelling the logit of the expected value as a linear model

$$\text{logit}\left(\mathrm{P}\left(y_i = 1|\mathbf{X}\right)\right) = \ln\left(\frac{\mathrm{P}\left(y_i = 1|\mathbf{X}\right)}{1 - \mathrm{P}\left(y_i = 1|\mathbf{X}\right)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}, \quad i = 1, \ldots, n. \tag{8}$$

The estimated probability $\widehat{p}_i(\mathbf{X}) = \mathrm{P}\left(y_i = 1|\mathbf{X}\right)$ is then given by

$$\widehat{p}_i(X) = \mathrm{P}\left(y_i = 1|\mathbf{X}\right) = \frac{\exp\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j}\right)}{1 + \exp\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j}\right)}. \tag{9}$$

To fit binary outcomes *via* boosting we replace the L_2 loss by the log loss

$$f_{\ln}\left(\mathbf{y}, \hat{\mathbf{p}}\right) = -\frac{1}{n} \sum_{i=1}^{n} y_i \ln\left(\hat{p}_i\right) + \left(1 - y_i\right) \ln\left(1 - \hat{p}_i\right). \tag{10}$$

Note that following this new loss function, the gradient is no longer represented by the residuals. The base-learners are hence fitted now to the first derivative of the loss function in Eq. 10. Consequently, batches are built out of the $p_b atch$ variants with the highest absolute correlation to the first derivative of the loss in Eq. 10 instead of the residual. However, the other components of the algorithm including the base learners remain unchanged. We also keep the hyperparameters derived in Section 3.1.2 fixed. We applied the extended algorithm on data of the UKBB for three binary phenotypes: the occurrence of asthma (UKBB field 22127), coeliac disease (UKBB field 21068) and high blood pressure (UKBB field 6150). All three traits are associated to many environmental factors but also have a genetic component (Arora and Newton-Cheh, 2010; Trynka et al., 2010; Yang et al., 2017; El-Husseini et al., 2020). Tanigawa et al. (2022) estimated high blood pressure to be a rather polygenic trait while the genetic component of asthma and coeliac disease is determined by fewer common variants.

We incorporated unrelated individuals of white British ancestry in our analysis and divided the samples randomly into training, validation and test sets. In total we used 8,397 cases (n_{train} = 4,266, n_{val} = 1,709 and n_{test} = 2,522) and 58,428 controls (n_{train} = 29,079, n_{val} = 11,707, n_{test} = 17,642) for asthma, 1,793 cases (n_{train} = 882, n_{val} = 361 and n_{test} = 550) and 92,646 controls (n_{train} = 46,234, n_{val} = 18,449, n_{test} = 27,963) for coeliac disease and 71,235 cases (n_{train} = 35,720, n_{val} = 14,210 and n_{test} = 21,305) and 190,422 controls (n_{train} = 94,740, n_{val} = 38,246, n_{test} = 57,436) for high blood pressure.

The applicability to binary traits was also one of the first extension of snpnet and Qian et al. (2020) showed impressive results for a number of binary traits. Due to that, we again also apply snpnet to the same data to evaluate the quality of our results.

We evaluated the accuracy of the resulting predictions on the test set using both the log loss as well as the AUC. Results are shown in Figure 7 and in Supplementary Figure S17. The overall predictive performance is comparable for all three phenotypes. For high blood pressure with a polygenic genetic component snpboost yields a sparse model with a high predictive performance. For sparse binary phenotypes as asthma and coeliac disease, snpboost and snpnet yield similar sparse models. The result for coeliac disease, which appears to be rather oliogenic than polygenic, for snpnet is outstanding, but in line with the results of Tanigawa et al. (2022). Nevertheless, also snpboost also estimates a very sparse PRS with a

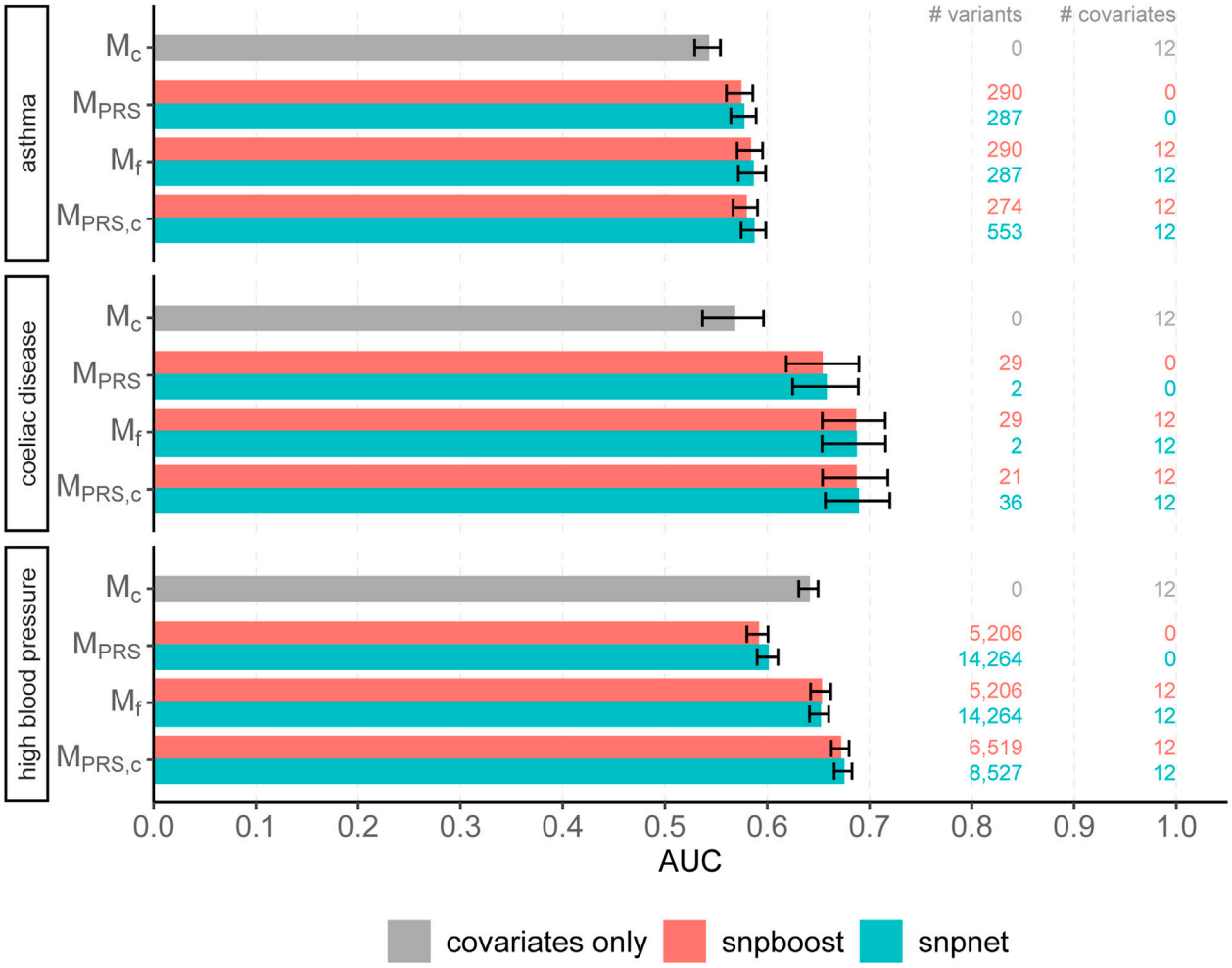


FIGURE 7
Comparison of predictive performance of snpnet and snpboost for three binary phenotypes from the UKBB. Results of the covariate-only model (M_c; grey bars) and multivariable polygenic models with and without inclusion of the covariates derived by lasso (snpnet; petrol-coloured bars) and statistical boosting (snpboost; red bars) for the prediction of three binary phenotypes from the UKBB. The barplots show the AUC on the test set of 20,164 (asthma), 28,513 (coeliac disease) and 78,741 (high blood pressure) unrelated white British individuals. M_{PRS} corresponds to a logistic regression model incorporating the PRS as a single predictor variable and M_f to a logistic regression model incorporating sex, age and the first ten principal components as additional covariates. $M_{PRS,c}$ includes the covariates already in the fitting process of the PRS. Bootstrapped 95% confidence intervals are indicated by error bars. Furthermore, information on the number of selected genetic variants (# variants) and the number of additionally included covariates (# covariates) is given.

high predictive performance. Table 2 illustrates the computation time for binary data of snpnet and snpboost on a computer cluster with 16 CPUs and 2 GB RAM each. Both, snpboost and snpnet yield very limited computation times, with snpnet being faster.

In summary, this illustrates how easily and conveniently the snpboost framework can be extended to different data types by incorporating different loss functions. Even though we simply plugged in the log loss and did not optimize the hyperparameters such as the batch size or the learning rate of our algorithm for binary data, snpboost yields a competitive predictive performance compared to the BASIL algorithm.

5 Discussion

In this work we have proposed a new methodological framework to derive multivariable PRS models *via* applying a statistical boosting approach directly on genotype data. Currently, PRS are most often built based on summary statistics from GWAS that were estimated by simple and univariate linear regression models (Choi et al., 2020). This methodologically simple approach is mainly justified by the computational hurdle resulting from the ultra-high-dimensionality of the genotype data. For example, in the past it had been unfeasible to fit a lasso model on the complete genotype data due to the high computational complexity. To overcome this, Mak et al. (2017) developed lassosum, an approach to approximate the lasso path by only using summary statistics and LD reference panels. However, recently published works provided methods to enable statistical modelling by penalized multivariable regression approaches on genotype data (Privé et al., 2018; Qian et al., 2020). Qian et al. demonstrated that lasso-based PRS were able to outperform several PRS derived by methods based on univariate summary statistics (Qian et al., 2020). First approaches to apply statistical boosting on genotype data employed a three-step-approach to fit multivariable PRS (Maj

et al., 2022): first, variants are pre-filtered based on their univariate associations with the examined phenotype. Second, statistical modelling and variable selection approaches such as AdaSub (Staerk et al., 2021) and boosting with probing (Thomas et al., 2017) are used to identify the informative variants on blocks of variants in LD. Finally, a multivariable PRS based on the selected variants is constructed *via* component-wise L_2-boosting (Bühlmann and Hothorn, 2007). While this approach yielded particularly sparse models and could compete with common methods like clumping and thresholding (Euesden et al., 2014), lasso *via* snpnet yielded more accurate results regarding the predictive performance which is usually the main objective of PRS modelling.

In the present article we introduced the boosting algorithm snpboost that works on smaller batches of variants similar to the BASIL algorithm. Our framework enables the application of statistical boosting directly on the complete original genotype data. In a smaller but still high-dimensional simulation setting, we were able to show that the adapted boosting algorithm yields similar performance to the original component-wise L_2-boosting, indicating that we do not lose predictive performance due to the incorporation of batches. In a further setting with more realistic dimensionality we have derived appropriate default values for the application of snpboost on large-scale data. We were able to show that the specified default values resulted in models with good performance in most cases but also gave advice on how to adapt them based on the genetic architecture of the examined phenotype and the specific research questions.

We applied the newly proposed snpboost algorithm on large-scale genotype data of the UKBB. In particular, we have compared the performance of snpboost to the one achieved by the lasso *via* snpnet, which has been shown to outperform many classical PRS (Qian et al., 2020). Our results indicate that the snpboost algorithm leads to PRS models that are highly competitive to lasso-based PRS models in both predictive performance and computation time. Although it might have been expected that the computation time would be higher for statistical boosting than for the lasso (Hepp et al., 2016), our approach had a tendency to result in sparser models. These sparser models correspond to an earlier stopping of the algorithm which reduces the computation time of boosting. The incorporation of further covariates such as age, sex and principal components in the fitting process of the PRS resulted in increased computation times for some phenotypes, particularly for height. However, in such cases, the boosting algorithm yielded an improved predictive performance with larger numbers of included variants. This illustrates that sparsity is not always favorable in regards of predictive performance. Additionally, we compared the performance of snpboost to further predictive PRS tools, which are either summary statistics based as PRScs, LDpred2 and SBayesR or multivariable approaches *via* the LDAK implementation of BayesR, Ridge Regression and Bolt-LMM (Zhang et al., 2021). While these methods do not apply variable selection, the predictive performance of snpboost was still highly competitive.

Our analyses show that there is a large overlap of the chosen variants by lasso and boosting, in particular regarding the variants with high estimated effect sizes. However, boosting has been found to include less variants in the final model and to induce less shrinkage on the effect estimates compared to the lasso. In clinical practice, a sparser PRS model might be of particular interest if the aim is not only prediction but also the identification of risk loci in the genome. In fact, functional annotations of the selected variants can better elucidate the underlying biological mechanisms influencing the analyzed trait. Thus, statistical boosting might be one way to yield more biologically interpretable PRS models.

Despite the presented promising results, the proposed method also inherited some limitations from statistical boosting. In contrast to classical regression methods, boosting does not provide closed formulas for standard errors of effect estimates or confidence intervals that could be used for inference. Furthermore, as mentioned before, statistical boosting is in general associated with a slightly higher computational complexity compared to methods such as the lasso (Hepp et al., 2016) and has a known tendency to include too many variables in low-dimensional settings (Staerk and Mayr, 2021; Strömer et al., 2022). Our results suggest that the incorporation of batches substantially reduced the computational time. Additionally, the reduction of the search space in each boosting step might partially prevent the algorithm from selecting too many variables. However, the implementation of the batch-based statistical boosting in snpboost is currently limited to linear base-learners, each corresponding to one genetic variant.

Apart from those technical limitations, using individual-level data raises ethical and logistical questions: While summary statistics are easily shared and do not allow for identification of unique individuals, individual-level data involve the risk of identification. It is therefore crucial, that providers as well as researchers using individual-level data follow ethical standards. Furthermore, the storage and transfer of individual-level data require more capacities which might not be at everyone's disposal in the complete research community. However, the resulting PRS can be published by sharing only the included variants, alleles and coefficients—exactly like summary statistics (Lambert et al., 2021). To make use of available summary statistics and to avoid the limitations associated with individual-level data, it might be of interest to develop an approximation of a component-wise boosting algorithm based on summary statistics and LD panels, analogously as lassosum for the lasso. From a computational perspective, this is not necessary as snpboost only requires limited resources (e.g., our analysis of the UKBB data was run with only 32 GB RAM in total).

By incorporating the log loss we made our framework applicable also to binary traits and demonstrated the convenience of further extensions of the snpboost framework beyond the case of continuous phenotypes. Without re-specifying our hyperparameters we were able to yield similar results as the snpnet framework.

In future research we want to further exploit the modular structure of boosting to model more complex biological phenomena. We will incorporate different loss functions to extend the snpboost framework to be applicable also to count and time-to-event data. To account for the uncertainty in the prediction, one could also construct subject-specific prediction intervals based on quantile regression (Mayr et al., 2012). Besides extending the approach *via* new loss functions, one could also change the base-learners in various ways. For example, base-learners could be adapted to take into account different models of inheritance beside the classical additive component typically used in the polygenic models, such as dominant and recessive hereditary schemes. Further possibilities for future research include the extension of the set of possible base-learners, e.g., to model gene-environment interactions as well as epistatic effects across variants which can play a relevant role in biological phenotypes (Li and Lehner, 2020). To do so, base-learners including interactions between variants and variant-covariate interactions could be incorporated. Apart from that, biological knowledge can also be used a priori. Márquez-Luna et al. (2021) have shown that the incorporation of functional annotations of the genetic variants contribute to a rise in

prediction accuracy. Previous works in the field of penalized regression and boosting have proposed to make use of biologically meaningful groups of genomic variants such as genes or pathways as described by Luan and Li (2008), Wei and Li (2007) as well as Liu et al. (2013). While those previous methods were computationally limited to smaller datasets our framework opens the possibility to include those ideas in the multivariable modelling of PRS. Besides those methodological extensions of our proposed snpboost framework, future research will also focus on the practical application of PRS derived by our framework. An important aspect of PRS research is the transferability of PRS models to different ethnicities, as PRS are often derived on cohorts of European ancestry and a substantial loss of predictive performance is observed when applied on further cohorts with different ethnicities (Landry et al., 2018; Evans et al., 2022). Previous studies have indicated that sparser models may contribute to overcome this issue (Maj et al., 2022) and it is of particular interest to examine the transferability of PRS derived by statistical boosting.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: This research has been conducted using the UK Biobank resource under application number 81202 (http://www.ukbiobank.ac.uk). Requests to access these datasets should be directed to UK Biobank, http://www.ukbiobank.ac.uk.

Ethics statement

The studies involving human participants were reviewed and approved by the North West Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB) approval. (UK Biobank, https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics). The patients/participants provided their written informed consent to participate in this study.

Author contributions

HK, AM, CS, CM, and PK contributed to conception and design of the method. HK wrote the code, performed the experiments and wrote the first draft of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

Funding

The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, grant number 428239776, MA7304/1-1).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1076440/full#supplementary-material

References

Arora, P., and Newton-Cheh, C. (2010). Blood pressure and human genetic variation in the general population. *Curr. Opin. Cardiol.* 25, 229–237. doi:10.1097/hco.0b013e3283383e2c

Beesley, L. J., Salvatore, M., Fritsche, L. G., Pandit, A., Rao, A., Brummett, C., et al. (2020). The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Statistics Med.* 39, 773–800. doi:10.1002/sim.8445

Bühlmann, P., and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Stat. Sci.* 22, 477–505. doi:10.1214/07-STS242

Bühlmann, P., and Yu, B. (2003). Boosting with the l2 loss. *J. Am. Stat. Assoc.* 98, 324–339. doi:10.1198/016214503000125

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Statistics* 1, 169–194. doi:10.1214/07-EJS008

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi:10.1038/s41586-018-0579-z

Chang, C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* 4, 7. doi:10.1186/s13742-015-0047-8

Chen, T., and Guestrin, C. (2016). "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (New York, NY, USA: Association for Computing Machinery), KDD '16), 785. doi:10.1145/2939672.2939785

Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. *Nat. Protoc.* 15, 2759–2772. doi:10.1038/s41596-020-0353-1

El-Husseini, Z. W., Gosens, R., Dekker, F., and Koppelman, G. H. (2020). The genetics of asthma and the promise of genomics-guided drug target discovery. *Lancet Respir. Med.* 8, 1045–1056. doi:10.1016/s2213-2600(20)30363-5

Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2014). PRSice: Polygenic risk score software. *Bioinformatics* 31, 1466–1468. doi:10.1093/bioinformatics/btu848

Evans, D. G., van Veen, E. M., Byers, H., Roberts, E., Howell, A., Howell, S. J., et al. (2022). The importance of ethnicity: Are breast cancer polygenic risk scores ready for women who are not of white European origin? *Int. J. Cancer* 150, 73–79. doi:10.1002/ijc.33782

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statistics* 29, 1189–1232. doi:10.1214/aos/1013203451

Fu, W., and Knight, K. (2000). Asymptotics for lasso-type estimators. *Ann. Statistics* 28, 1356–1378. doi:10.1214/aos/1015957397

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nat. Commun.* 10, 1776. doi:10.1038/s41467-019-09718-5

Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nat. Genet.* 42, 558–560. doi:10.1038/ng0710-558

Greenshtein, E., and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* 10, 971–988. doi:10.3150/bj/1106314846

Hassanin, E., May, P., Aldisi, R., Spier, I., Forstner, A. J., Nöthen, M. M., et al. (2021). Breast and prostate cancer risk: The interplay of polygenic risk, rare pathogenic germline variants, and family history. *Genet. Med.* 24, 576–585. doi:10.1016/j.gim.2021.11.009

Hemani, G., Yang, J., Vinkhuyzen, A., Powell, J., Willemsen, G., Hottenga, J.-J., et al. (2013). Inference of the genetic architecture underlying bmi and height with the use of 20, 240 sibling pairs. *Am. J. Hum. Genet.* 93, 865–875. doi:10.1016/j.ajhg.2013.10.005

Henderson, C. R. (1950). Estimation of genetic parameters. *Ann. Math. Stud.* 21, 309.

Hepp, T., Schmid, M., Gefeller, O., Waldmann, E., and Mayr, A. (2016). Approaches to regularized regression – a comparison between gradient boosting and the lasso. *Methods Inf. Med.* 55, 422–430. doi:10.3414/ME16-01-0033

Hoerl, A. E., and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 42, 80–86. doi:10.1080/00401706.2000.10485983

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-based Boosting 2.0. *J. Mach. Learn. Res.* 11, 2109–2113.

National Human Genome Research Institute (2021). The cost of sequencing a human genome. Available at: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (Accessed 11 04, 2021).

Kronenberg, F., and Utermann, G. (2013). Lipoprotein(a): Resurrected by genetics. *J. Intern. Med.* 273, 6–30. doi:10.1111/j.1365-2796.2012.02592.x

Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K.-A., Mooij, T. M., Roos-Blom, M.-J., et al. (2017). Risks of breast, ovarian, and contralateral breast cancer for brca1 and brca2 mutation carriers. *JAMA* 317, 2402–2416. doi:10.1001/jama.2017.7112

Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., Buniello, A., et al. (2021). The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* 53, 420–425. doi:10.1038/s41588-021-00783-5

Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., and Bonham, V. L. (2018). Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Aff.* 37, 780–785. doi:10.1377/hlthaff.2017.1595

Li, R., Chang, C., Justesen, J. M., Tanigawa, Y., Qian, J., Hastie, T., et al. (2022). Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. *Biostatistics* 23, 522–540. doi:10.1093/biostatistics/kxaa038

Li, X., and Lehner, B. (2020). Biophysical ambiguities prevent accurate genetic prediction. *Nat. Commun.* 11, 4923. doi:10.1038/s41467-020-18694-0

Liu, J., Huang, J., Ma, S., and Wang, K. (2013). Incorporating group correlations in genome-wide association studies using smoothed group lasso. *Biostatistics* 14, 205–219. doi:10.1093/biostatistics/kxs034

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., et al. (2019). Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nat. Commun.* 10, 5086. doi:10.1038/s41467-019-12653-0

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., et al. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* 47, 284–290. doi:10.1038/ng.3190

Luan, Y., and Li, H. (2008). Group additive regression models for genomic data analysis. *Biostatistics* 9, 100–113. doi:10.1093/biostatistics/kxm015

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature* 456, 18–21. doi:10.1038/456018a

Maj, C., Staerk, C., Borisov, O., Klinkhammer, H., Yeung, M. W., Krawitz, P., et al. (2022). Statistical learning for sparser fine-mapped polygenic models: The prediction of LDL-cholesterol. *Genet. Epidemiol.* 46, 589–603. doi:10.1002/gepi.22495

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* 41, 469–480. doi:10.1002/gepi.22050

Márquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S. S., Furlotte, N., Auton, A., et al. (2021). Incorporating functional priors improves polygenic prediction accuracy in UK biobank and 23andme data sets. *Nat. Commun.* 12, 6052. doi:10.1038/s41467-021-25171-9

Mayr, A., and Hofner, B. (2018). Boosting for statistical modelling-a non-technical introduction. *Stat. Model.* 18, 365–384. doi:10.1177/1471082X17748086

Mayr, A., Hothorn, T., and Fenske, N. (2012). Prediction intervals for future BMI values of individual children - a non-parametric approach by quantile boosting. *BMC Med. Res. Methodol.* 12, 6. doi:10.1186/1471-2288-12-6

Meinshausen, N. (2007). Relaxed lasso. *Comput. Statistics Data Analysis* 52, 374–393. doi:10.1016/j.csda.2006.12.019

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLOS Genet.* 11, e1004969. doi:10.1371/journal.pgen.1004969

Privé, F., Arbel, J., and Vilhjálmsson, B. J. (2021). Ldpred2: Better, faster, stronger. *Bioinformatics* 36, 5424–5431. doi:10.1093/bioinformatics/btaa1029

Privé, F., Aschard, H., Ziyatdinov, A., and Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: Bigstatsr and bigsnpr. *Bioinformatics* 34, 2781–2787. doi:10.1093/bioinformatics/bty185

Purcell, S., and Chang, C. (2015). Plink 2.0. Available at: www.cog-genomics.org/plink/2.0/.

Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., et al. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLOS Genet.* 16, e1009141. doi:10.1371/journal.pgen.1009141

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sabatine, M. S. (2019). PCSK9 inhibitors: Clinical evidence and implementation. *Nat. Rev. Cardiol.* 16, 155–165. doi:10.1038/s41569-018-0107-8

Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H. M., Jackson, A. U., et al. (2011). Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet.* 7, e1002198. doi:10.1371/journal.pgen.1002198

Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., et al. (2021). Genetics of 35 blood and urine biomarkers in the UK biobank. *Nat. Genet.* 53, 185–194. doi:10.1038/s41588-020-00757-z

Staerk, C., Kateri, M., and Ntzoufras, I. (2021). High-dimensional variable selection via low-dimensional adaptive learning. *Electron. J. Statistics* 15, 1797. doi:10.1214/21-ejs1797

Staerk, C., and Mayr, A. (2021). Randomized boosting with multivariable base-learners for high-dimensional variable selection and prediction. *BMC Bioinforma.* 22, 441. doi:10.1186/s12859-021-04340-z

Strömer, A., Staerk, C., Klein, N., Weinhold, L., Titze, S., and Mayr, A. (2022). Deselection of base-learners for statistical boosting—With an application to distributional regression. *Stat. Methods Med. Res.* 31, 207–224. doi:10.1177/09622802211051088

Tanigawa, Y., Qian, J., Venkataraman, G., Justesen, J. M., Li, R., Tibshirani, R., et al. (2022). Significant sparse polygenic risk scores across 813 traits in UK biobank. *PLOS Genet.* 18, e1010105–e1010121. doi:10.1371/journal.pgen.1010105

Thomas, J., Hepp, T., Mayr, A., and Bischl, B. (2017). Probing for sparse and fast variable selection with model-based boosting. *Comput. Math. Methods Med.* 2017, 1421409–1421418. doi:10.1155/2017/1421409

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Trynka, G., Wijmenga, C., and van Heel, D. A. (2010). A genetic perspective on coeliac disease. *Trends Mol. Med.* 16, 537–550. doi:10.1016/j.molmed.2010.09.003

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statistics* 36, 614–645. doi:10.1214/009053607000000929

Vilhjálmsson, B., Yang, J., Finucane, H., Gusev, A., Lindström, S., Ripke, S., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97, 576–592. doi:10.1016/j.ajhg.2015.09.001

Wei, Z., and Li, H. (2007). Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics* 8, 265–284. doi:10.1093/biostatistics/kxl007

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46, 1173–1186. doi:10.1038/ng.3097

Yang, I. V., Lozupone, C. A., and Schwartz, D. A. (2017). The environment, epigenome, and asthma. *J. Allergy Clin. Immunol.* 140, 14–23. doi:10.1016/j.jaci.2017.05.011

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47, 1114–1120. doi:10.1038/ng.3390

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi:10.1038/ng.608

Zhang, Q., Privé, F., Vilhjálmsson, B., and Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* 12, 4192. doi:10.1038/s41467-021-24485-y

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67, 301–320. doi:10.1111/j.1467-9868.2005.00503.x

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 15 February 2023
DOI 10.3389/fgene.2023.1109269

OPEN ACCESS

EDITED BY
Angelo Facchiano,
Institute of Food Sciences (CNR), Italy

REVIEWED BY
Jiaogen Zhou,
Huaiyin Normal University, China
Duolin Wang,
University of Missouri, United States

*CORRESPONDENCE
Yi Zhao,
biozy@ict.ac.cn

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 27 November 2022
ACCEPTED 25 January 2023
PUBLISHED 15 February 2023

CITATION
Luo C, Wu Y and Zhao Y (2023), SupCAM: Chromosome cluster types identification using supervised contrastive learning with category-variant augmentation and self-margin loss.
Front. Genet. 14:1109269.
doi: 10.3389/fgene.2023.1109269

SupCAM: Chromosome cluster types identification using supervised contrastive learning with category-variant augmentation and self-margin loss

Chunlong Luo[1,2], Yang Wu[1] and Yi Zhao[1]*

[1]Research Center for Ubiquitous Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, [2]University of Chinese Academy of Sciences, Beijing, China

Chromosome segmentation is a crucial analyzing task in karyotyping, a technique used in experiments to discover chromosomal abnormalities. Chromosomes often touch and occlude with each other in images, forming various chromosome clusters. The majority of chromosome segmentation methods only work on a single type of chromosome cluster. Therefore, the pre-task of chromosome segmentation, the identification of chromosome cluster types, requires more focus. Unfortunately, the previous method used for this task is limited by the small-scale chromosome cluster dataset, ChrCluster, and needs the help of large-scale natural image datasets, such as ImageNet. We realized that semantic differences between chromosomes and natural objects should not be ignored, and thus developed a novel two-step method called SupCAM, which could avoid overfitting only using ChrCluster and achieve a better performance. In the first step, we pre-trained the backbone network on ChrCluster following the supervised contrastive learning framework. We introduced two improvements to the model. One is called the category-variant image composition method, which augments samples by synthesizing valid images and proper labels. The other introduces angular margin into large-scale instance contrastive loss, namely *self*-margin loss, to increase the intraclass consistency and decrease interclass similarity. In the second step, we fine-tuned the network and obtained the final classification model. We validated the effectiveness of modules through massive ablation studies. Finally, SupCAM achieved an accuracy of 94.99% with the ChrCluster dataset, which outperformed the method used previously for this task. In summary, SupCAM significantly supports the chromosome cluster type identification task to achieve better automatic chromosome segmentation.

KEYWORDS

supervised contrastive learning, category-variant data augmentation, angular margin loss, chromosome cluster types identification, karyotyping

1 Introduction

Karyotyping is an essential cytogenetic experiment technique that aims to find numerical and structural abnormalities of chromosomes. Normally, human tissue cells have 23 pairs of chromosomes, including autosomes and sex chromosomes. These chromosomes are stained using Giemsa staining techniques and then photographed using advanced microscope cameras to generate metaphase images. The karyotyping analysis usually requires the segmentation of

chromosome instances from metaphase images. Owing to the inefficiency and high cost of manual analysis, researchers have presented many automatic algorithms to ease the burden.

Most existing studies focus on the chromosome segmentation task but ignore its pre-task, chromosome cluster types identification. As non-rigid chromosomes float in an oil droplet when photographed, it is usual that touching and severely overlapping chromosomes occur in metaphase images, namely chromosome clusters. However, using classical geometric connectivity techniques, it is easy to obtain individual instances or clusters from a metaphase image. Most existing chromosome segmentation studies only dive into a specific type of chromosome cluster. To segment touching clusters, Arora (2019) and Yilmaz et al. (2018) present algorithms that make full use of the geometric characteristics between touching areas. To segment overlapping chromosome clusters, Hu et al. (2017) tries to design a new customized neural network for better performance. To segment touching-overlapping clusters, Minaee et al. (2014) dives into the geometric features of this type of cluster and proposes a geometric-based method. Alternatively, Lin et al. (2020) chooses to improve the state-of-the-art deep-learning model to tackle this issue. Nevertheless, if we can automatically identify the type of chromosome cluster first and then input it to the above segmentation methods, we can automatically segment chromosomes directly from metaphase images.

In 2021, Lin et al. (2021) proposed the chromosome cluster type identification task. In this work, 6,592 chromosome clusters were obtained from the hospital, and they created and made available the first chromosome cluster dataset (ChrCluster for simplicity). All samples are manually annotated into four categories: instance, overlapping, touching, and touching-overlapping, as shown in Figure 1. Finally, they propose a classification model as the benchmark of the ChrCluster dataset. To avoid overfitting on the small-scale ChrCluster dataset, Lin et al. (2021) takes Instagram weakly supervised learning pretrained weights [Mahajan et al. (2018)] and the customized ResNeXt [Xie et al. (2017)] classification model to achieve an accuracy of 94.09%.

However, chromosome cluster images are gray images and only contain specific domain objects, which results in different distributions between the ChrCluster dataset and the ImageNet/Instagram dataset. Therefore, pre-training the model with the ImageNet or Instagram dataset is not the ideal option. Given this point, we attempt to pre-train domain-friendly weights only using the ChrCluster dataset for better downstream task performance.

Self-supervised contrastive learning [Wu et al. (2018); van den Oord et al. (2018); Hénaff (2020); Chen T. et al. (2020a); He et al. (2020); Chen X. et al. (2020b)] is an unsupervised learning mechanism that aims to pre-train representative features (output of specific weights) that can be transferred to downstream tasks by fine-tuning. They achieve contrastive learning through a Siamese network structure. Large-scale instance contrastive loss, such as InfoNCE, is used to attract the positive pairs and repulse the negative pairs. Specifically, they regard the different augmentation views of the same instance as positive pairs and views from different instances as negative pairs. Finally, pre-trained weights are transferred to downstream tasks, such as classification, detection, and segmentation. Supervised contrastive learning methods [Khosla et al. (2020); Cui et al. (2021); Kang et al. (2021)] are further proposed to achieve better performance on the downstream classification task. They add label information into self-supervised contrastive learning. With the help of label information, not only embeddings from the different views of the same instance should be gathered together but also embeddings of instances from the same class should be pulled close, which will result in many positives for each embedding as opposed to a single positive in self-supervised contrastive learning. Given this way, we can utilize the supervised contrastive learning framework to pre-train domain-friendly features that can capture more similarity among intraclass. However, both contrastive learning methods train the model using instance contrastive loss like the SupCon loss [Khosla et al. (2020)], which means that they are non-parametric and do not have a final FC layer as a classifier. As a result, fine-tuning at the downstream chromosome cluster identification task is essential.

For both self-supervised and supervised contrastive learning, category-invariant data augmentation approaches are essential. SimCLR [Chen T. et al. (2020a)] has systematically proved the importance of category-invariant data augmentation (*RandomResizedCrop*, *RandomColorJittering*, and *GaussianBlur*) for self-supervised contrastive learning. However, stronger category-variant augmentation techniques [Zhang et al. (2018); Yun et al. (2019)] are ignored due to the lack of label information. Supervised contrastive learning methods have added label information, but the instance contrastive loss they use is not yet able to adapt to continuous labels generated by previous category-variant augmentation methods. Therefore, we introduce a category-variant image composition method with discrete targets for our proposed supervised contrastive learning method, which can further enrich the visual schemas of the ChrCluster dataset and achieve better performance.

In addition, large-scale instance contrastive loss is important for supervised contrastive learning. It is obvious that the inner product of normalized embeddings in both InfoNCE [Chen T. et al. (2020a)] and SupCon [Khosla et al. (2020)] is equal to the cosine similarity operation.

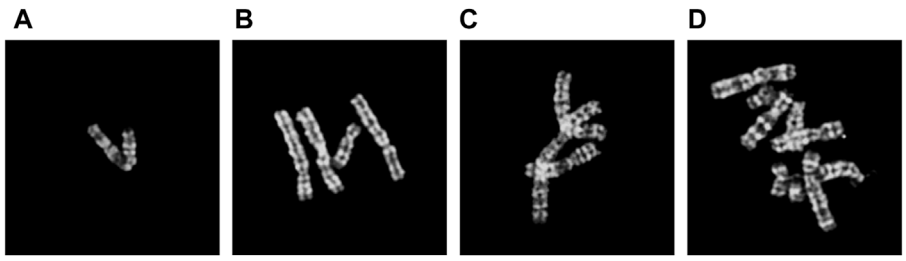


FIGURE 1
Examples of different types of chromosome clusters include **(A)** instance, **(B)** touching, **(C)** overlapping, and **(D)** touching-overlapping.

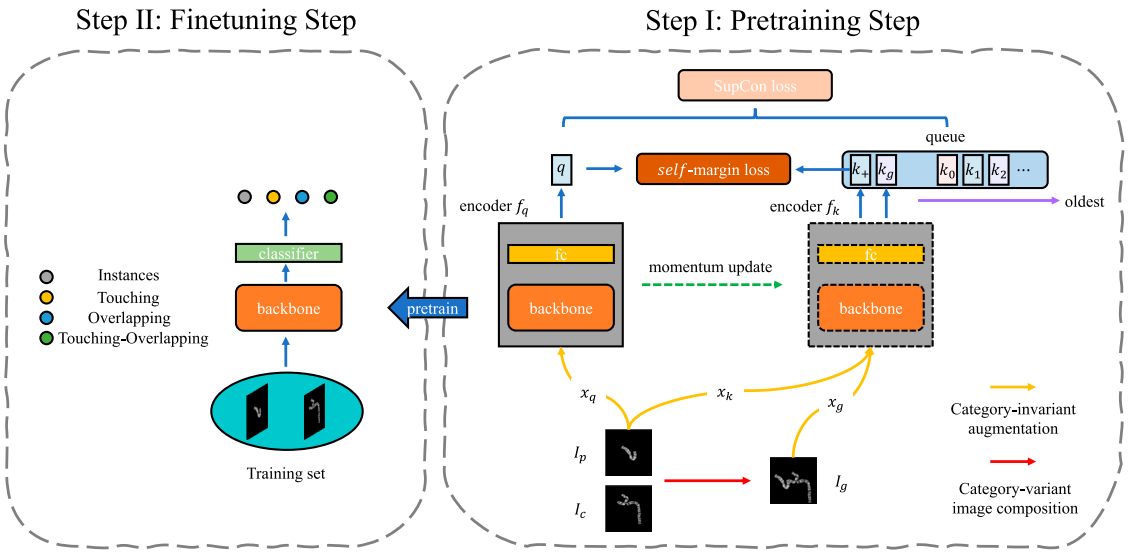


FIGURE 2
The framework of SupCAM. The pretraining step shown is based on MoCo [He et al. (2020)] structure but imports two modifications, namely category-variant image composition and *self*-margin loss. The fine-tuning step will initialize the backbone using weights pre-trained in the first step and finally reports identification results.

The angular between two embeddings is the only variable in the loss. Thus, adding an angular margin can achieve better intraclass compactness and interclass discrimination. For example, previous angular margin-based losses [Liu et al. (2016); Liu et al. (2017); Wang et al. (2018a); Wang et al. (2018b); Deng et al. (2019)] encourage sharper feature distribution and better discriminating performance by adding various angular margins between instance features and class weights. Among them, the Additive Angular Margin loss [Deng et al. (2019)] performs best. Given this way, we can design a new large-scale instance contrastive loss using additive angular margin to enhance the semantic discrimination capability of pre-trained features.

To sum up, inspired by the supervised contrastive learning method SupCon [Khosla et al. (2020)], we propose the two-step SupCAM approach to identify the various chromosome cluster types. In the first pretraining step, considering that the MoCo [He et al. (2020)] style network can save more storage space by the memory queue, we take MoCo as feature extractor to encode images. To learn category-related features, we take SupCon loss to maximize the consistency across all views of all samples in the same class rather than only that of the various views of the same sample. Additionally, we provide a category-variant image composition method to augment chromosome cluster images, which combines two randomly chosen images and assigns a new discrete label in accordance with the rule to create a new valid sample. We also import an angular margin into different embeddings of the instance contrastive loss to bring embeddings from the same class closer together. Owing to the poor synchronization between the query and the old keys, a straightforward extension that simply adds angular margins to all positive pairs may fall short of achieving model convergency. Therefore, we only import an angular margin between the different views of the same sample, known as *self*-margin loss, which is the first attempt to enforce more compact embeddings using large-scale instance contrastive loss with angular margin. In the second step, we fine-tune the final classification model based on the pre-trained backbone from the first step. We prove the effectiveness of our methods by fine-tuning multiple classical classification networks, such as ResNet and its variants. Overall, our main contributions in this paper can be summarized as follows:

- We solve chromosome cluster identification through a two-step method, named SupCAM, that pre-trains the backbone in a supervised contrastive learning manner and fine-tunes classification models. In this way, SupCAM obtains more representative features to avoid overfitting and domain-friendly pre-trained weights as a better alternative to ImageNet pre-trained weights.
- We propose a category-variant image composition method that will reassign the category according to the overlapping area of the chromosome clusters.
- We import angular margin into instance contrastive-based loss, named *self*-margin loss. The *self*-margin loss will enforce higher intraclass compactness and interclass discrepancy of the model.
- We prove the efficiency of our contributions through the public chromosome cluster types dataset, ChrCluster. We also achieve 94.99% accuracy, which is higher than the 94.09% accuracy proposed by Lin et al. (2021).

2 Methods

We will go into more depth about the suggested method in this section. In the section entitled 'Two-Step Framework' 2.1, we fully detail the SupCAM pre-training and fine-tuning steps and emphasize the significance of the new loss function and novel data augmentation method. The details of the new category-variant image composition approach, including the composing algorithm and principles of label assigning, will thereafter be covered in the section entitled 'Category-Variant Image Composition' 2.2. In the section entitled '*Self*-margin loss' 2.3, we will deduce new *self*-margin loss through merging label information and angular margin step by step.

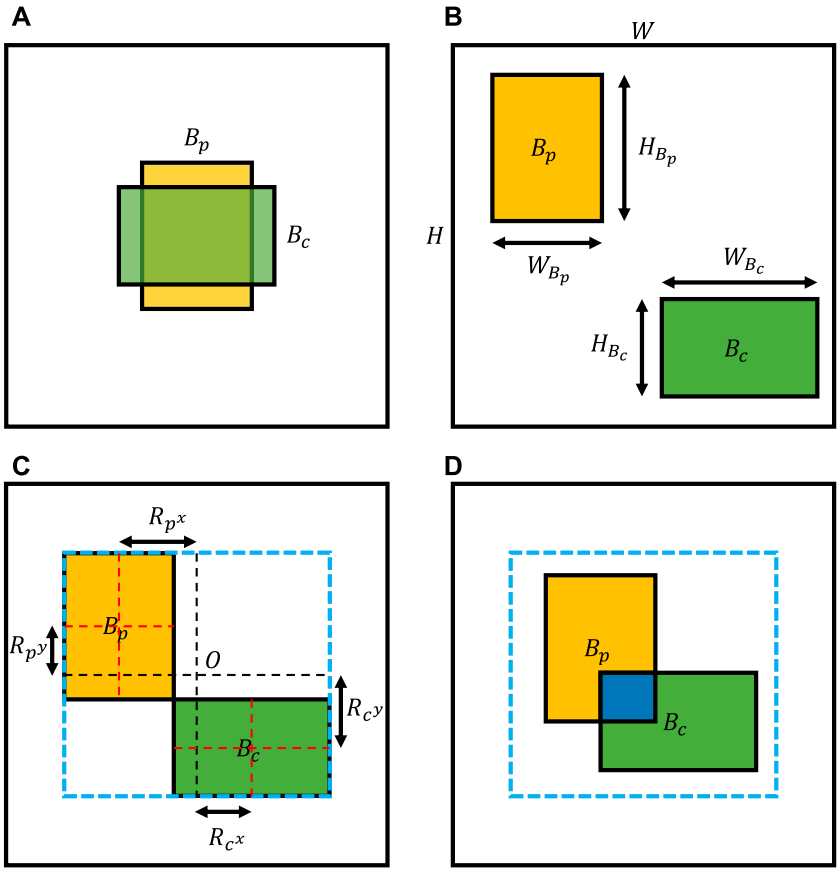


FIGURE 3
Illustration of image shift. **(A)** Common composing image without image shift. **(B)** Invalid composing image because we do not limit image shift ranges. **(C)** The image shift range determined by the maximum outer enclosing box of two bounding boxes. **(D)** Valid composing result, as we sample image shifts under a reasonable range.

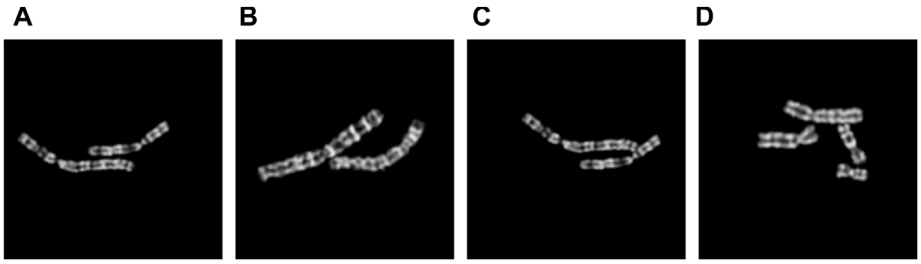


FIGURE 4
Examples of composed images with different number *N* of overlapping pixels. **(A)** *N* = 3; **(B)** *N* = 30; **(C)** *N* = 84; **(D)** *N* = 150.

2.1 Two-step framework

In this study, we present a two-step method called SupCAM that consists of the pre-training and fine-tuning steps, as shown in Figure 2, to tackle the chromosome cluster types classification problem. We pre-trained our backbone using the supervised contrastive learning framework in the first stage. In the second step, we extracted representative features through a pre-trained backbone and fine-tuned a few traditional classification models for final identification.

2.1.1 Pre-training step

In the pre-training step, we took the MoCo as the basic architecture in this work, but it is free to be replaced with other self-supervised contrastive learning models. As shown in Figure 2, SupCAM owns query encoder f_q, and key encoder f_k. f_q was trained in

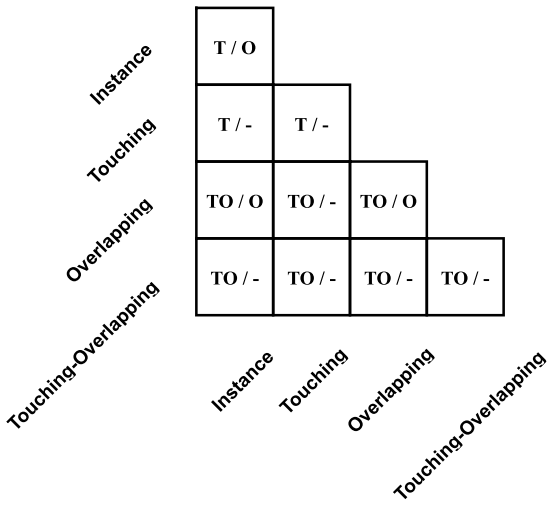


FIGURE 5
Detail of y_g in the *middle*-Table. The left signs in each cell represent assigned labels when the number of pixel intersections of warped I_p and I_c is below or equal to the pixel intersection threshold P_n, and the right side signs represent assigned labels when pixel intersections are larger than the threshold. Furthermore, mark 'T', 'O', 'TO', and '-' represent touching, overlapping, touching-overlapping, and uncertainty tags, respectively.

an end-to-end manner but f_k was implemented as a momentum-based moving average of f_q. We also inherited the dynamically updated queue but ignored the projection head used in the MoCo.

To gain multiple views of the sampled images during training, we first used category-invariant and category-variant augmentation approaches. Specifically, we randomly sampled primary image I_p and candidate image I_c. The primary images were augmented by the category-invariant augmentation methods as usual, resulting in two views with the same class, denoted as x_q and x_k. The I_c was first augmented using a category-variant image composition method, which combines with the I_p to create a new image, called generated image I_g. A new class label was assigned according to the look-up table. Then, the same category-invariant augmentation modules were applied on the I_g, leading to the x_g. We will further describe the details of the category-variant image composition method in the category-variant image composition Section 2.2. Afterward, as shown in Figure 2, through the query encoder f_q and key encoder f_k, augmented samples were mapped to a tuple of representation vectors:

$$\{q, k_+, k_g\} = \{f_q(x_q), f_k(x_k), f_k(x_g)\} \quad (1)$$

where key encoder f_k encodes both x_k and x_g to embeddings k_+ and k_g. (q, k_+) is the intrinsic positive pair as it comes from the same image, but k_g is positive or negative depending on whether I_g has the same class label with I_p. Besides, k_+ and k_g are used to update the memory queue in a first input first output (FIFO) manner. Benefiting from the slowly progressing key encoder and progressively replaced queue, representations in the queue can remain as consistent as possible with the latest q, which helps the contrastive model converge.

Inspired by the excellent performance of angular margin loss [Liu et al. (2017); Wang H. et al. (2018b); Deng et al. (2019)], we present *self*-margin loss in this study for better discriminative power of the pre-trained backbone. Specifically, our final loss consisted of the SupCon loss and *self*-margin loss. For each query q, a set of encoded keys $\{k_0, k_1, k_2, \dots\}$ and k_+ and k_g were used to compute SupCon loss. Meanwhile, as k_+ was not only the newest key compared with other keys in the memory queue but also had the same class as q, we only applied an additional angular margin between q and k_+. In this way, we achieved better performance while keeping the training

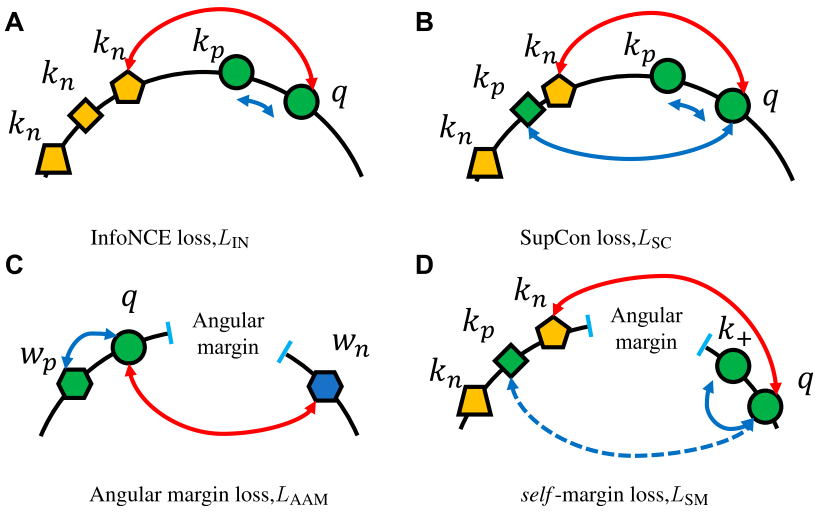


FIGURE 6
Examples of various loss functions. **(A)** InfoNCE loss, which pulls query q and current key k_p together and regards each old embedding in the memory queue as negative key k_n, which should be pushed away. **(B)** SupCon loss, in which not only the current key but old keys that have the same class as the query should be pulled close. The example of angular margin loss in **(C)** shows how hard margin constrains parametric weights and makes the same class embeddings more compact. **(D)** Depiction of our *self*-margin loss, which only enforces the angular margin between the query and the current key and ignores other positive keys in the memory queue.

TABLE 1 Comparison with previous methods. All experiments were conducted following the division principle in Lin et al. (2021). ResNeXt101: ResNeXt101 32 × 8d; †: ResNeXt101-32 × 8d attached with a customized header network invented by Lin et al. (2021); ImageNet: 1.28 million images with 1,000-class ImageNet dataset; Instagram: 940 million public images with a ~ 1500 hashtags dataset proposed by Mahajan et al. (2018).

Methods	Backbone	Pre-train dataset	Accuracy	Precision	Sensitivity	Specificity	F1
Lin et al. (2021)	ResNet101	ImageNet(1.28 M)	91.89	90.65	87.92	97.30	88.32
	DenseNet121	ImageNet(1.28 M)	87.65	85.59	81.68	95.88	82.23
	ResNeXt101	ImageNet(1.28 M)	92.27	90.79	89.10	97.42	89.36
	ResNeXt101†	Instagram(940 M)	94.09	93.08	92.79	98.03	**92.84**
SupCAM	ResNet101	ChrCluster(6.5K)	94.24	92.54	92.00	97.74	91.37
	DenseNet121	ChrCluster(6.5K)	94.69	92.92	92.89	97.94	92.11
	ResNeXt101	ChrCluster(6.5K)	**94.99**	**93.25**	**92.81**	**98.12**	92.26

The bold values represent that they are the best performance in this metric.

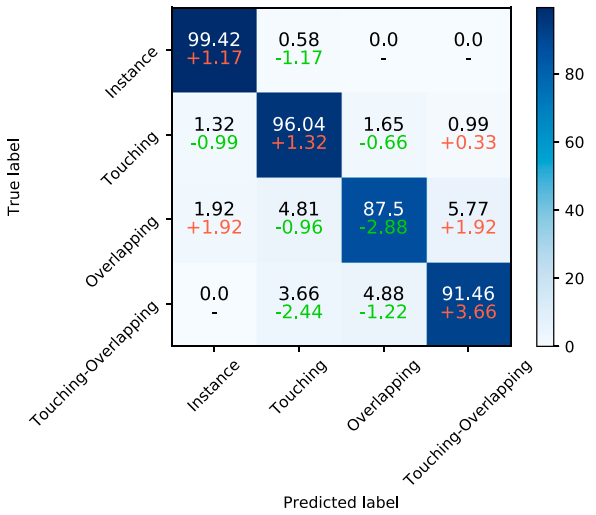


FIGURE 7
The first line of each cell is the confusion matrix of SupCAM with ResNeXt101 backbone on the test set. Besides, on the second line of each cell, we show differences between Lin et al. (2021) and our method, in which red and green represent increment or decrement of percentage, respectively.

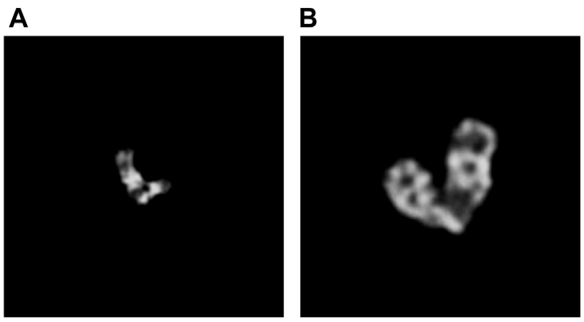


FIGURE 8
Overlapping clusters are misclassified as the instance class by SupCAM. **(A)** shows a misclassified example where the bottom of the left chromosome occludes the other one. **(B)** is another misclassified example where the bottom of the right chromosome occludes the bottom of the left one.

process stable. In Section 2.3, the analysis of the *self*-margin loss will be shown in detail.

2.1.2 Fine-tuning step

All results shown in the section entitled 'Experimental results and discussion' 3 are from the fine-tuned classification model. As shown in Figure 2, in the fine-tuning step, we reused the pre-trained backbone network and attached a fully connected layer, a four-classes linear classifier, on top of it as our chromosome cluster types identification model. After loading the pre-trained weights of the backbone network and randomly initializing the fully connected layer, we trained the model on the training set for several epochs. In the end, we evaluated the SupCAM model on the test set for the module's effectiveness and final performance. The details of the classification model and training process will be described in the section entitled 'Implementation Details' 3.3.

2.2 Category-variant image composition

In this section, we will introduce a category-variant image composition algorithm as a strong data augmentation policy in SupCAM. Traditional category-invariant data augmentation methods dominate self-supervised and supervised contrastive learning methods. However, stronger category-variant data augmentation methods, such as *Mixup* [Zhang et al. (2018)] and *CutMix* [Yun et al. (2019)], are ignored because they do not satisfy the discrete targets requirements of large-scale instance contrastive loss. Thus, we propose a category-variant image composition algorithm to synthesize new chromosome cluster samples with discrete labels for enriching visual schemas.

2.2.1 Algorithm

Let (I_p, y_p) and (I_c, y_c) denote primary and candidate samples, respectively, where $\{I_p, I_c\} \in \mathbb{R}^{W \times H \times C}$. The goal of category-variant image composition is to generate a new training sample (I_g, y_g) by combining primary and candidate samples. We defined the composing process as:

$$I_g = \lambda \mathcal{W}(T_p, I_p) \oplus (1-\lambda)\mathcal{W}(T_c, I_c)$$
$$y_g = \mathcal{L}(y_p, y_c) \tag{2}$$

where $\mathcal{W}$ represents affine function, T_p, T_c are the transformation matrix of primary and candidate images, and λ is the combination ratio. Besides, $\oplus$ is complex combination operation and $\mathcal{L}$ means look-up table operation, which will be described in the look-up table Section 2.2.2

Input: primary sample (I_p, y_p), candidate sample (I_c, y_c), upper limit of sampling number N, pixel intersection $P_\cap$
Output: generated sample (I_g, y_g)

Initialize y_g is uncertainty and sampling count $n = 0$
$W, H = Size(I_p)$
Binary mask of I_p and I_c:

$$M_p = \mathcal{I}_{[I_p \neq 0]} I_p; M_c = \mathcal{I}_{[I_c \neq 0]} I_c$$

Bounding box of chromosome cluster in M_p and M_c:

$$B_{i\in\{p,c\}} = \left(\min_x M_i, \max_x M_i, \min_y M_i, \max_y M_i\right)$$

Shift range of image I_p and I_c:

$$R_{i^x|i\in\{p,c\}} = \left[\min\left(W, W_{B_p} + W_{B_c}\right) - W_{B_i}\right]/2$$
$$R_{i^y|i\in\{p,c\}} = \left[\min\left(H, H_{B_p} + H_{B_c}\right) - H_{B_i}\right]/2$$

while y_g is uncertainty and $n < N$ **do**
Shift bias are uniformly sampled according to:

$$S_{i^x|i\in\{p,c\}} = U(-R_{i^x}, R_{i^x})$$
$$S_{i^y|i\in\{p,c\}} = U(-R_{i^y}, R_{i^y})$$

Warp the images using transformation matrix T_p and T_c: $T_{i\in\{p,c\}} = \begin{bmatrix} 1 & 0 & S_{i^x} \\ 0 & 1 & S_{i^y} \end{bmatrix}$, $\hat{I}_{i\in\{p,c\}} = \mathcal{W}(T_i, I_i)$.
Generate I_g according to the warped images through combination operation $\oplus$:

$$I_g^{i,j} = \begin{cases} \hat{I}_p^{i,j} & , \text{ if } \hat{I}_p^{i,j} > 0, \hat{I}_c^{i,j} = 0 \\ \hat{I}_c^{i,j} & , \text{ if } \hat{I}_p^{i,j} = 0, \hat{I}_c^{i,j} > 0 \\ 0.5\hat{I}_p^{i,j} + 0.5\hat{I}_c^{i,j} & , \text{ if } \hat{I}_p^{i,j} > 0, \hat{I}_c^{i,j} > 0 \\ 0 & , \text{ Others} \end{cases}$$

Assign label by look-up table: $\mathcal{L}(y_p, y_c, N_{\hat{I}_p^{i,j} \cap \hat{I}_c^{i,j}}, P_\cap)$
$n = n + 1$
if y_g is not uncertainty **then**
return Generated sample (I_g, y_g)
end if
end while
return Candidate sample (I_c, y_c)

Algorithm 1. Category-Variant Image Composition.

As shown in Algorithm 1, we first extracted the foreground-background mask of I_p and I_c through indicator function $\mathcal{I}$ and then obtained the bounding box of chromosome cluster area by min-max operation. The shift range along the *x*-axis and *y*-axis of two images is restricted by the size relation between the images and bounding boxes. Given the range, we uniformly sampled the shift bias and utilized them to construct transformation matrix $T \in \mathbb{R}^{2\times3}$ of image I_p and I_c. The affine function $\mathcal{W}$ will generate transformed images according to the transformation matrix and origin images. To generate I_g and avoid unnatural artifacts, we designed a complex combination operation $\oplus$, which assigned linear interpolations of pixels only in the overlapping area. The foreground and background areas were assigned original pixels. Meanwhile, the label of I_g was achieved through the look-up table $\mathcal{L}$. However, because of the uncertainty of y_g, we sampled the shift bias multiple times for meaningful results but also imported an upper limit of sampling number N (normally 10 in our experiments) to balance the efficiency and effectiveness. Therefore, if we have sampled more than N times, candidate sample (I_c, y_c) will be directly output. The uncertainty of y_g will be detailed in the look-up table Section 2.2.2.

Here, the importance of image shift should be clarified. Unlike *Mixup*, which conducts linear interpolations of all pixels, and *CutMix*, which replaces a random image region with a patch from another image, we need to shift the image to simulate specific properties of different types of chromosome clusters. As shown in Figure 3A, chromosome clusters are commonly distributed in the central region of the image, which means that we combine images directly without random shift, leading to overlapping instances dominating the generated samples. Additionally, we should set a limited range for the shift bias. On the one hand, unlimited shifting may lead to the loss of characteristic areas, such as overlapping or touching regions. On the other hand, as shown in the invalid image illustrated in Figure 3B, most composing results may show as two individual chromosome clusters, which do not satisfy any definition of chromosome cluster types proposed by Lin et al. (2021). To determine the range of shift bias, we simplified the irregular concave polygons of chromosome clusters to rectangles of bounding boxes. Then, two bounding boxes could uniquely confirm a maximum outer enclosing box as the border of shift bias, like Figure 3C. In this way, we are much more likely to be able to generate chromosome clusters that satisfy the definition, as shown in Figure 3D.

2.2.2 Look-up table

In this section, we will clarify the process of assigning the correct class label to each generated sample, namely the look-up table. Considering the image composition processing and the chromosome cluster definition, the generated image will not belong to the instance category in the first place. Besides, according to Lin et al. (2021), the crucial difference between overlapping and touching chromosome clusters is whether any connectivity between two chromosome instances entails pixels intersection. However, as shown in Figure 4, it is counterintuitive if we consider these results as overlapping cases but only a few pixel intersections distribute in the pixel connectivity region. Given this point, before assigning four chromosome cluster types and an uncertainty tag, we first need to set a pixel intersection threshold $P_\cap$ greater than zero to decide whether generated image I_g is touching case $(N_{\hat{I}_p^{i,j} \cap \hat{I}_c^{i,j}} \leq P_\cap)$ or overlapping case $(N_{\hat{I}_p^{i,j} \cap \hat{I}_c^{i,j}} > P_\cap)$.

The table in Figure 5 shows the guidance for assigning a cluster type to generated images I_g, and for simplicity, we call it *middle*-Table. Original categories can pair into 20 possible touching and overlapping cases. As listed in *middle*-Table, the left of each cell is the candidate cluster types of touching cases, and the right is the candidate cluster types of overlapping cases. Specifically, for touching cases, their class type depends on whether touching or overlapping clusters exist in original sample pairs. In other words, only if overlapping clusters exist

TABLE 2 Ablation study of the SupCAM model with ResNet50 on the 30% test set of the ChrCluster dataset. We repeated all experiments 10 times and report the mean and standard deviation. IN indicates that the backbone network has been pre-trained by the ImageNet dataset. CatVar, category-variant image composition method; L_{SM}, *self*-margin loss.

IN	MoCo	SupCon	CatVar	L_{SM}	Accuracy	Precision	Sensitivity	Specificity	F1
					88.38 ± .60	84.07 ± .82	83.94 ± .79	95.79 ± .19	82.11 ± .88
✓					92.65 ± .30	90.24 ± .65	89.87 ± .73	97.31 ± .10	88.79 ± .72
	✓				89.15 ± .34	85.31 ± .60	85.23 ± .49	95.97 ± .15	83.61 ± .53
	✓	✓			91.65 ± .32	88.08 ± .51	88.60 ± .38	96.97 ± .13	87.00 ± .47
	✓	✓	✓		93.24 ± .20	91.18 ± .46	90.60 ± .40	97.49 ± .09	89.75 ± .46
	✓	✓	✓	✓	**93.56 ± .18**	**91.65 ± .42**	**91.31 ±.36**	**97.63 ± .06**	**90.34 ± .41**

The bold values represent that they are the best performance in this metric.

in original sample pairs can composed touching cases be tagged as a touching-overlapping type, such as an overlapping-instance pair. Otherwise, y_g should assign the touching type, such as the instance-instance pair and the touching-instance pair.

As for overlapping cases, most of the uncertainty of label y_g happens in this case that the number of pixel intersections beyond pixel intersection threshold $P_{\cap}$. Strictly speaking, except for overlapping cases of instance-instance pair, all overlapping cases should be assigned the uncertainty tag as we cannot be sure about the number of touching and overlapping regions, such as in the *light*-Table described in the section entitled Category-Variant Image Composition 3.5.3. For example, given a touching-instance pair, we can assign the touching-overlapping type or the overlapping type according to the size and position of overlapping areas between two chromosome clusters. However, we should emphasize the overlapping-instance pair and the overlapping-overlapping pair. Although two pairs can be assigned the touching-overlapping type or the overlapping type, we hypothesize that when these pairs are overlapping cases, they are unlikely to have touching areas and should directly mark the overlapping type. Finally, experiment results in Table 4 support the above hypothesis.

2.3 *Self*-margin loss

As in the framework shown in Figure 2, we extended the InfoNCE loss to *self*-margin loss by gradually merging label information and additive angular margin.

Given an encoded query $q \in \mathbb{R}^d$ and a set of encoded samples $\{k_0, k_1, k_2, \dots\}$ stored in the memory queue, the InfoNCE loss L_{IN}, as shown in Figure 6A considered as the following:

$$L_{\mathrm{IN}} = -\log \frac{e^{q\cdot k_p/\tau}}{e^{q\cdot k_p/\tau} + \sum_{k_i \in K_N} e^{q\cdot k_i/\tau}} \tag{3}$$

where k_p is the only positive key in the memory queue that q matches and K_N represent the remaining negative key set. $\tau \in \mathbb{R}^+$ is a scalar temperature parameter. In this way, L_{IN} is low if q is more in agreement with its positive key k_p than other negative keys, which is intuitively like a $(K_N + 1)$ classes cross-entropy loss in the form.

Different from only augmented views of the same image should be considered as positives in InfoNCE loss, SupCon loss L_{SC} as shown in Figure 6B, imports label information and generalizes to an arbitrary number of positives as long as they belong to the same class:

$$L_{\mathrm{SC}} = -\frac{1}{\|K_P\|} \sum_{k_p \in K_P} \log \frac{e^{q\cdot k_p/\tau}}{e^{q\cdot k_p/\tau} + \sum_{k_i \in K_N} e^{q\cdot k_i/\tau}} \tag{4}$$

where K_P is a set of positive keys that have the same class label as query q. The SupCon loss function can be regarded as the average of multiple times of InfoNCE loss value, as each k_p can be considered as the only positive key at some point. The loss encourages the encoder to pull embeddings of the same class closer, resulting in a more reasonable distribution of representations for the subsequent supervised learning task.

Now we move on to the additive angular margin loss L_{AAM} proposed in ArcFace [Wang F. et al. (2018a)]. As illustrated in Figure 6C, a larger angular margin, which exists between q and negative class weight w_n, will enforce the same class queries q closer and make them easily identifiable. We suppose we have normalized weights $W \in \mathbb{R}^{d\times(\|K_N\|+1)}$ of the last fully connected layer where it can be redefined as one positive class center $w_p \in \mathbb{R}^d$ that the input matches to and remaining negative class centers $W_N \in \mathbb{R}^{d\times\|K_N\|}$. Additionally, we normalize its inputs q and ignore the bias term for simplicity. Then, the L_{AAM} can be reformulated as follows using our notation:

$$L_{\mathrm{AAM}} = -\log \frac{e^{\cos(\theta_{q,w_p}+m)/\tau}}{e^{\cos(\theta_{q,w_p}+m)/\tau} + \sum_{w_i \in W_N} e^{\cos\theta_{q,w_i}/\tau}}, \tag{5}$$

where $\theta_{q,w_i} = \arccos\left(\frac{q\cdot w_i}{\|q\|\|w_i\|}\right)$ represents the angle between w_i and query q. An additional margin penalty m is added on $\theta_{q,w_p} = \arccos\left(\frac{q\cdot w_p}{\|q\|\|w_p\|}\right)$ to enforce higher intraclass compactness and interclass discrimination.

If we set $w_p = k_p$, $W_N = K_N$, and $w_i = k_i$ in L_{AAM}, then from Eqs 4, 5 we have *self*-margin loss L_{SM}:

$$L_{\mathrm{SM}} = -\frac{1}{\|K_P\|} \sum_{k_p \in K_P} \log \frac{e^{\cos\left(\theta_{q,k_p}+m\right)/\tau}}{e^{\cos\left(\theta_{q,k_p}+m\right)/\tau} + \sum_{k_i \in K_N} e^{\cos\theta_{q,k_i}/\tau}} \tag{6}$$

However, L_{AAM} relies on parametric weights from the last fully connected layer. These weight vectors are the latest and are smoothly updated by end-to-end backpropagation, which results in enough synchronization between embeddings and weights. On the contrary, although a slowly evolving key encoder exists, all keys used in contrastive losses (such as L_{IN} and L_{SC}) are non-parametric and rapidly changing in a FIFO manner. Given this point, positive keys are consistent enough for the contrastive-based loss but not synchronized enough for the angular margin-based loss. We cannot even make the model converge using Eq. 6.

TABLE 3 Ablation study of composition methods. Equal weights mean that the overlapping area of I_p and I_c are combined half and half. λ shows that we sampled a λ from beta distribution and then applied linear interpolations in the overlapping areas of two images. The final maximum experiment represents the operation of taking the maximum pixel value in overlapping areas.

Composition method	Accuracy	Precision	Sensitivity	Specificity	F1
Equal weights	**93.56 ± .18**	91.65 ± .42	91.31 ± .36	**97.63 ± .06**	90.34 ± .41
λ-interpolation [Yun et al. (2019)]	93.41 ± .21	**91.66 ± .37**	**91.66 ± .35**	**97.62 ± .09**	**90.47 ± .35**
Maximum	93.04 ± .17	89.93 ± .37	89.93 ± .36	97.44 ± .08	88.77 ± .36

The bold values represent that they are the best performance in this metric.

TABLE 4 Ablation study of the look-up table. Besides the *middle*-Table, which was our final choice, we tried to extend the label assigning to the extreme, namely through the *heavy*-Table and *light*-Table schemes. The goal of the *no*-Table is to examine the effects of candidate image I_c.

Scheme	Accuracy	Precision	Sensitivity	Specificity	F1
middle-Table	**93.56** ± .18	**91.65** ± .42	**91.31** ± .36	**97.63** ± .06	**90.34** ± .41
heavy-Table	93.30 ± .28	90.76 ± .56	90.56 ± .56	97.55 ± .11	89.52 ± .57
light-Table	93.09 ± .11	90.77 ± .34	90.66 ± **.29**	97.47 ± **.05**	89.50 ± **.31**
no-Table	93.19 ± **.10**	90.84 ± **.31**	90.71 ± .40	97.53 ± **.05**	89.56 ± .36

The bold values represent that they are the best performance in this metric.

TABLE 5 The table below shows SupCAM performance with different angular margin values (*m* in Eq. 7) used in the *self*-margin loss during the first pre-training step.

Angular margin	Accuracy	Precision	Sensitivity	Specificity	F1
$m = 0.1$	93.39 ± **.15**	91.10 ± **.29**	90.92 ± **.35**	97.59 ± .07	89.90 ± **.37**
$m = 0.2$	**93.56** ± .18	**91.65** ± .42	**91.31** ± .36	**97.63** ± **.06**	**90.34** ± .41
$m = 0.3$	93.21 ± .26	90.62 ± .57	90.73 ± .52	97.51 ± .10	89.46 ± .58
$m = 0.4$	91.87 ± .35	88.50 ± .42	88.42 ± .42	97.10 ± .12	87.04 ± .53
$m = 0.5$	\	\	\	\	\

The bold values represent that they are the best performance in this metric.

As shown in Figure 2, the synchronization between query q and positive key $k_+ \in K_P$ has been guaranteed by the similar weights (moving-average key encoder f_k and the same batch). Therefore, in Eq. 6, for query q, we only hold on to the latest inherent positive key k_+ and ignore the remaining positive keys, including possible k_g. The final formulation of *self*-margin loss L_{SM} is:

$$L_{\mathrm{SM}} = -\log \frac{e^{\cos\left(\theta_{q,k_+}+m\right)/\tau}}{e^{\cos\left(\theta_{q,k_+}+m\right)/\tau} + \sum_{k_i \in K_N} e^{\cos\theta_{q,k_i}/\tau}} \tag{7}$$

and illustrated in Figure 6D.

We will prove the performance of L_{SM} in Experiments 3.5 and compare it with some intuitive candidate methods.

3 Experimental results and discussion

3.1 Dataset

In this study, we used the dataset reported by Lin et al. (2021) to evaluate our model performance and demonstrate the effectiveness of modules. The dataset is the first clinical chromosome cluster dataset that has 6,592 samples, called ChrCluster. All samples are padded to the 224 × 224 size and manually labeled into four categories: 1,712 chromosome instance, 3,029 touching chromosomes cluster, 1,038 overlapping chromosomes cluster, and 813 touching-overlapping chromosomes cluster. In the ablation study Section 3.5, we described how we split the dataset into 3,955 training samples, 659 validation samples, and 1,978 test samples in a class-based random stratified fashion. For the final comparison in the Section entitled 'Comparison Result' 3.4, we followed the division principle described by Lin et al. (2021), which has 80% training data, 10% validation data, and 10% test data. To avoid leaking test set information from the pre-training step to the fine-tuning step, we pre-trained the backbone network only using the training set no matter whether the goal is an ablation study or final comparisons.

3.2 Evaluation metrics

To fairly evaluate the performance of SupCAM, we followed the main evaluation metrics described by Lin et al. (2021) including

TABLE 6 Other candidate angular margin based scenarios and the main differences are detailed in Section 3.5.4.

Other angular margin method	Accuracy	Precision	Sensitivity	Specificity	F1
self-margin loss	**93.56** ± .18	**91.65** ± .42	**91.31** ± .36	**97.63** ± .06	**90.34** ± .41
Parametric margin(m = 0.2)	93.29 ± .20	90.88 ± .38	91.10 ± .34	97.56 ± .08	89.94 ± .35
Parametric margin(m = 0.3)	93.13 ± .15	91.03 ± .34	90.88 ± .32	97.49 ± .07	89.83 ± **.29**
Parametric margin(m = 0.4)	93.08 ± .19	90.66 ± .36	90.44 ± **.24**	97.47 ± .06	89.40 ± .32
Parametric margin(m = 0.5)	93.26 ± **.13**	91.42 ± **.31**	91.03 ± .29	97.56 ± **.05**	90.07 ± .30
Cluster margin	93.09 ± .18	90.57 ± .42	90.74 ± .49	97.51 ± .08	89.49 ± .47

The bold values represent that they are the best performance in this metric.

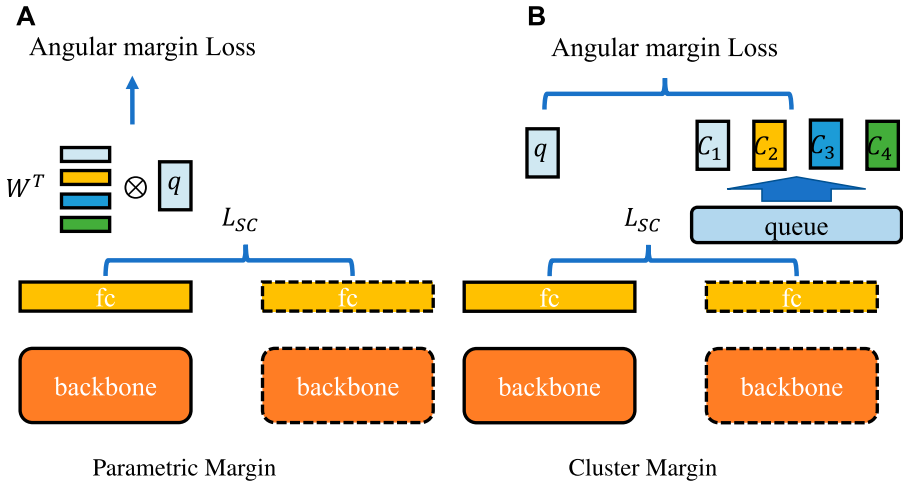


FIGURE 9
Comparison of margin-based methods. The backbone network is shared between the first and second steps, and the fully connected layer outputs queries and keys. The dashed lines indicate that all layers are updated in a moving-average manner. L_{SC} represents SupCon loss. **(A)** represents the "parametric margin" schema which applies angular margin loss between query embedding and additional parametric weights W. **(B)** is the "cluster margin" method that clusters all key embeddings into four class centers $\{C_1, C_2, C_3, C_4\}$ according to class label, and applies angular margin between query embeddings and cluster centers.

accuracy, *precision*, *sensitivity*, *specificity*, and *F1*. It is worth noticing that except for the *accuracy*, all the above-mentioned metrics were averaged in a 'macro' fashion. The 'macro' fashion will first calculate metrics for each category individually and then average the metrics across classes with equal weights.

Now, we should clarify the definition of the following four basic criteria in a multi-classification task:

- True positive(TP_i): given a test sample that belongs to i-th class, if the model correctly predicts it as i-th class, we regard it as true positive.
- False positive(FP_i): given a test sample that does not belong to i-th class, if the model incorrectly predicts it as i-th class, we regard it as false positive.
- False negative(FN_i): given a test sample that belongs to i-th class, if the model incorrectly predicts it as other classes, we regard it as false negative.
- True negative(TN_i): given a test sample that does not belong to i-th class, if the model correctly predicts it as other classes, we regard it as true negative.

Assume that N_c represents the number of chromosome cluster categories and N is the number of test set instances, then we have:

$$accuracy = \frac{1}{N}\sum_{i=0}^{N_c} TP_i \tag{8}$$

$$\begin{aligned} precision &= \frac{1}{N_c}\sum_{i=0}^{N_c} precision_i \\ &= \frac{1}{N_c}\sum_{i=0}^{N_c} \frac{TP_i}{TP_i + FP_i} \end{aligned} \tag{9}$$

$$\begin{aligned} sensitivity &= \frac{1}{N_c}\sum_{i=0}^{N_c} sensitivity_i \\ &= \frac{1}{N_c}\sum_{i=0}^{N_c} \frac{TP_i}{TP_i + FN_i} \end{aligned} \tag{10}$$

$$\begin{aligned} specificity &= \frac{1}{N_c}\sum_{i=0}^{N_c} specificity_i \\ &= \frac{1}{N_c}\sum_{i=0}^{N_c} \frac{TN_i}{TN_i + FP_i} \end{aligned} \tag{11}$$

$$F1 = \frac{1}{N_c}\sum_{i=0}^{N_c} 2 \cdot \frac{precision_i \cdot sensitivity_i}{precision_i + sensitivity_i} \tag{12}$$

All above-mentioned metrics are as higher as better. We use percentages for them and keep two decimal places.

3.3 Implementation details

We implemented our work on the Pytorch Lightning[1] toolbox based on the Pytorch [Paszke et al. (2019)] deep-learning library. We finished all experiments on an Ubuntu OS Server with one NVIDIA GTX Titan Xp GPU.

In the first pre-training phase, following the MoCo pipeline, we optimized the structure and some hyperparameters for the chromosome cluster type identification task. As described in the two-step framework Section 2.1, besides the conventional query q and key k_+ used in MoCo, we additionally generated x_g using the category-variant image composition method and encoded it as the third embedding k_g through the key encoder. Limited by the size of the dataset, we reduced the embedding dimension to 128-d and the queue capacity to 1,024 accordingly. The scalar temperature τ used in SupCon loss and *self*-margin loss was set as 0.07. We chose 0.2 for angular margin m and 200 for pixel intersection hyperparameter $P_\cap$. We used SGD as our optimizer, where momentum is 0.9, and weight decay is 0.0001. We set the mini-batch size as 32 for the single GPU and trained the model for 200 epochs. Furthermore, we applied linear warm-up during the first 20 epochs until achieving the initial learning rate 0.03 and then decayed it through a cosine annealing schedule. Category-invariant image augmentation methods used in the first step included *RandomResizedCrop* and *HorizontalFlip*. The total loss is the sum of SupCon loss L_{SC} and *self*-margin loss L_{SM}:

$$L = L_{SC} + L_{SM} \tag{13}$$

In the second fine-tuning step, for the classification model, we first loaded the corresponding pre-trained backbone module and randomly initialized the weights and bias of the final classifier. Only *RandomRotate* was employed during the training phase to reduce overfitting. We used SGD as our optimizer and had the same setting as the pre-training step. We trained the classification model for 15 epochs with a mini-batch of 16 images. Unlike the first step, the initial learning rate was set as 0.01 and decreased by 0.1 after 8 and 12 epochs individually. The loss function adopted in the fine-tuning step was cross-entropy loss enhanced by label smoothing (hyperparameter $\sigma = 0.1$) [Szegedy et al. (2016)].

3.4 Comparison result

3.4.1 Overview

In this section, we report the final results following the division principle of Lin et al. (2021). Table 1 shows the comparison results between SupCAM and previous methods. On the top of Table 1, we list some representative experiment results of previous methods with different backbones, including ResNet101 [He et al. (2016)], DenseNet121 [Huang et al. (2017)], and ResNeXt101 [Xie et al. (2017)]. Specifically, ResNeXt† optimizes the header of the classification model using a mixed pooling layer and multiple linear-dropout groups. Meanwhile, not only 1.28 million images from the ImageNet dataset but also approximately 940 million images from the Instagram dataset are used to pre-train backbone weights, which are loaded as initial weights of the ResNeXt†. Owing to above-mentioned improvements, ResNeXt† proposed by Lin et al. (2021) achieves the previous state-of-the-art performance, which is 94.09 accurate and has the best results with other evaluation metrics.

In this study, benefiting from the supervised contrastive learning framework enhanced by the category-variant image composition methods and *self*-margin loss, SupCAM achieved the best performance. Specifically, SupCAM improved the accuracy by a large margin of 2.35 under ResNet101 and 2.72 under the original ResNeXt101. Finally, although Lin et al. (2021) used an extremely large Instagram dataset, which was almost 140,000 times larger than ChrCluster, we still increased the accuracy by approximately 0.9 compared with ResNeXt†. Except for F1, other metrics also performed better. It is worth noting that previous methods may suffer from heavy overfitting, as shown in the result that used the DenseNet121 as the backbone network in Lin et al. (2021). As a more powerful backbone than ResNet101, DenseNet121 performed less well in all metrics. By contrast, under SupCAM, DenseNet121 successfully outperformed ResNet101, which means that SupCAM can alleviate the risk of overfitting without relying on a large dataset but using only the ChrCluster dataset. To sum up, Table 1 shows the high data utilization efficiency and robustness of the SupCAM for solving the task of chromosome cluster type identification. In addition, we evaluated the performance using pre-trained weights from ImageNet instead of random initialization in the first step of SupCAM, and as shown in Supplementary Table S1, it also outperformed the previous method, but was worse than the final SupCAM.

3.4.2 Confusion matrix

Besides the above metrics across classes, we used a confusion matrix to further reveal the performance of the SupCAM method in each class. As shown in Figure 7, SupCAM outperformed a previous study [Lin et al. (2021)] on instance, overlapping, and touching-overlapping classes but was weak in the overlapping category. Specifically, the number of touching-overlapping clusters incorrectly predicted as touching and overlapping types were reduced simultaneously, which resulted in an increment of 3.66 in the accuracy of the touching-overlapping class. Additionally, the accuracy of the instance class and the touching class was increased to 99.42 and 96.04, respectively. It is obvious that the combination of the category-variant image composition method and *self*-margin loss can improve the performance of the identification model in most chromosome cluster categories.

At the same time, to try to explain the degeneracy of SupCAM in the overlapping category, we list some false negative samples, especially those misclassified as the instance type. As illustrated by Figure 8, they are puzzling samples, and it is hard to decide whether they belong to the overlapping type at first glance. On the other side, a hard threshold of pixel intersection in the category-variant image composition method may import artificial disturbance to the label system, which adds confusion to the final prediction. Therefore, these weaknesses inspire us to propose more reasonable and natural image composition methods in the future.

1 https://pytorch-lightning.readthedocs.io/en/latest/

3.5 Ablation study

3.5.1 Overview

To evaluate the effectiveness of each model, we applied the ablation study at the 30% test set of the ChrCluster dataset. To avoid performance fluctuations due to the small size of the dataset, all experiments during the ablation study were repeated 10 times and we obtained the mean and standard deviation of each evaluation metric. In this way, as well as comparing the performance through the mean value, we can further justify the stableness of each method.

As shown in Table 2, we first trained the chromosome cluster types classification model from scratch as the baseline, which was 88.38 ± 0.60 accurate. Pre-training on the large ImageNet dataset further improved the accuracy to 92.65 ± 0.30. However, the above experiments suffer from larger performance fluctuation than our methods, which reminds us that a huge domain gap exists between the ImageNet and ChrCluster. Therefore, pre-training the chromosome cluster types identification model on the large ImageNet dataset is not the best choice. Finally, we proved that the key factor driving the model performance improvement is the model structure itself as SupCAM achieved the best performance among all experiments under the same fine-tuning strategies.

3.5.2 Supervised contrastive learning

To verify the contribution of supervised contrastive learning to the performance, before completing the basic classification task, we imported the pre-training step, which pre-trained the backbone in a supervised contrastive manner with SupCon loss through MoCo architecture. We took the MoCo augmentation setting [Chen X. et al. (2020b)] as the initial augmentation method in this experiment. Table 2 shows that the MoCo-style supervised contrastive pre-training step increased accuracy by 3.27 points and had a F1 score 4.89 points higher than the model trained from scratch. It is notable here that the direct employment of the MoCo-style supervised contrastive pre-training step was worse than the identification model pre-trained by the ImageNet dataset, but it was more stable in some cases. In conclusion, pre-training the backbone in a supervised contrastive manner is effective but we need more specific optimizations to adapt the chromosome cluster types identification task.

3.5.3 Category-variant image composition

The experiment results in Table 2 show that the category-variant image composition method improves accuracy from 91.65 ± 0.32 to 93.25 ± 0.20 and specificity from 96.97 ± 0.13 to 97.49 ± 0.09. Both the performance and stableness of this model were increased and even outperformed the model trained by the MoCo setting, which validates that the category-variant image composition method can more reasonably and effectively augment chromosome cluster data than the original MoCo augmentation setting.

To be more specific, as shown in Supplementary Figure S1, we experimented with multiple candidate pixel intersection threshold $P_{\cap}$, and box plots show that when the $P_{\cap}$ is set as 200 pixels, the model achieves the best performance in all metrics. Meanwhile, we also examined the choices of composition methods in overlapping areas, as shown in Table 3. Besides the equal weights method used in this study, we list two representative composition methods. Linear interpolation through a sampled λ ~B(1,1) is widely used in Yun et al. (2019) and Zhang et al. (2018). Another straightforward idea is taking the maximum pixel value from the primary image I_p and the candidate image I_c as the final pixel in overlapping areas. Experiments show that the 'maximum' method is not suitable for the chromosome cluster types identification task and the "λ-interpolation" method performs badly on the most important accuracy criterion, although slightly outperforms the 'equal weights' method on other metrics.

Furthermore, we confirmed the design of the look-up table in Table 4. As shown in the results, the *middle*-Table scheme achieved the best performance. In addition, we evaluated some extreme scenarios, such as the *heavy*-Table scheme and the *light*-Table scheme. Specifically, the *heavy*-Table scheme assigns an explicit label to each (I_p, I_c) pair directly no matter whether disagreements exist in overlapping cases. Suppose there is a touching-instance pair in an overlapping case, the *middle*-Table will tag them with an uncertainty label, but with a *heavy*-Table, we roughly assign the touching-overlapping category. The *light*-Table solution takes the opposite approach by not providing any valid label for overlapping cases unless they all belong to the instance type. The results in Table 4 show that the *heavy*-Table achieved an accuracy of 93.30 ± .28, which outperformed the 93.09 ± .11 accuracy of the *light*-Table scheme. Through the comparison between *light*-*middle*-*heavy* solutions, we can conclude that 1) category-variant image composition method indeed improves the performance of the cluster type identification task ($\mu^{Acc}_{heavy} > \mu^{Acc}_{light}$); 2) we should avoid roughly assigning a label for complicated cases ($\mu^{Acc}_{middle} > \mu^{Acc}_{heavy}$); and 3) manually composing an image and assigning a label inevitably imports unnatural counterfeits, resulting in performance fluctuation ($\sigma^{Acc}_{heavy} > \sigma^{Acc}_{middle} > \sigma^{Acc}_{light}$).

Moreover, to clarify the effects of taking I_c as I_g as in line 16 of Algorithm 1; Table 4 shows the results from a contrast experiment we conducted, called a *no*-Table scheme, that only used existing candidate image I_c rather than composed images. As expected, *no*-Table achieved an accuracy of 93.19 ± .10, which was lower but more stable than that of *middle*-Table, proving the effectiveness and relative unstableness of the category-variant image composition method once more.

3.5.4 SupCAM with *Self*-margin loss

As shown in Table 2, *self*-margin loss improved the accuracy from 93.24 ± .20 to 93.56 ± .18 and the precision, sensitivity, specificity, and F1 scores were also improved. Besides, it is worth noting that weights pre-trained with *self*-margin loss could further stabilize the final classification performance. Thus, we validated the effectiveness of *self*-margin loss of the first pre-training step.

It is important to find the optimal margin m for the chromosome cluster types identification task, and the best margin m observed in Table 5 was 0.2. Specifically, smaller additional angular margin penalties, such as $m = 0.1$ and $m = 0.2$, improved the performance. However, when margin penalties was large, e.g., $m = 0.3$ and $m = 0.4$, *self*-margin loss not only decreased the performance but also made the model more unstable. When the margin penalty increased to 0.5, the model could not be converged. Therefore, we conclude that although we ensure synchronization by (q, k_+) pair, the moving-average update manner makes the model more sensitive to the large margin penalty than the model updated in an end-to-end manner, which is further described in the next paragraph.

Furthermore, margin-based architectures are diverse, and we justified the advantages of *self*-margin loss through the results shown in Table 6. As illustrated in Figure 9A, with 'parametric margin' as one of the candidate schemes, we additionally added an end-to-end updating weight $W \in \mathcal{R}^{d\times4}$ as classes centers after the original fully connected layer and the angular margin-based loss is applied between the parametric weights and query embedding q. Results proved that the 'Parametric Margin' scheme is not good at the chromosome cluster types identification task; however, its better stability also confirms the

conclusion in the above paragraph. Another candidate scheme is 'cluster margin', as shown in Figure 9B. To form meaningful class centers for each query q, we clustered all key embeddings stored in the memory queue according to their label and renormalized the center of each cluster. Cluster centers were updated in a moving-average manner. However, the results in Table 6 confirmed what we inferred in the *self*-margin loss' Section 2.3, i.e., that terrible synchronization leads to worse performance under the angular margin framework.

4 Conclusion

In this study, we proposed a two-step SupCAM method to solve the chromosome cluster types identification task. In the first step, we improved the supervised contrastive learning method through a strong category-variant image composition algorithm and *self*-margin loss. After pre-training, we further fine-tuned the classification models in the second step. The effectiveness of each module was proved by massive ablation studies. The top prediction performance suggested that SupCAM has state-of-the-art performance in the chromosome cluster identification task. All these experimental findings demonstrate that the proposed SupCAM, as a supervised contrastive learning method, can effectively extract more representative and domain-friendly weights from the small-scale ChrCluster and is a better alternative to previous ImageNet pre-trained weights as it alleviates overfitting risks, resulting in better performance. Specifically, SupCAM introduces a strong category-variant image composition method with discrete labels to generate more abundant visual schemas. Meanwhile, we designed and implemented a new stable *self*-margin loss by adding an angular margin between the different embeddings of the instance contrastive loss, resulting in higher intraclass compactness and interclass discrepancy. Although our study focuses on chromosome cluster identification, our proposed method can inspire more researchers to analyze medical images using only small-scale medical image datasets rather than large natural image datasets. In the future, we will refine image composition processing and the look-up table to achieve more stable performance. In addition, other schemes that add angular margin into instance contrastive-based loss should be further studied.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: [https://github.com/ChengchuangLin/ChromosomeClusterIdentification].

Author contributions

CL, YW, and YZ contributed to conception and design of the study. CL and YW analyzed the dataset. CL performed the model experiments. CL wrote the first draft of the manuscript. All authors contributed to the manuscript revision and read and approved the submitted version.

Funding

This work was supported by the Institute of Computing Technology, Chinese Academy of Sciences (55E161080).

Acknowledgments

We thank Dr. Lianhe Zhao for correcting the article grammar and some spelling errors, and PhD student Yufan Luo for his comments on the Introduction section and Category-variant image composition section.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1109269/full#supplementary-material

References

Arora, T. (2019). A novel approach for segmentation of human metaphase chromosome images using region based active contours. *Int. Arab. J. Inf. Technol.* 16, 132–137.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. (2020a). "A simple framework for contrastive learning of visual representations," in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual Event (PMLR), 13-18 July 2020 (Vienna, Austria: PMLR), 1597–1607. vol. 119 of Proceedings of Machine Learning Research.

Chen, X., Fan, H., Girshick, R. B., and He, K. (2020b). *Improved baselines with momentum contrastive learning*. CoRR abs/2003.04297.

Cui, J., Zhong, Z., Liu, S., Yu, B., and Jia, J. (2021). "Parametric contrastive learning," in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, October 10-17, 2021 (Montreal, QC, Canada: IEEE), 695–704. doi:10.1109/ICCV48922.2021.00075

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). "Arcface: Additive angular margin loss for deep face recognition," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, June 16-20, 2019 (Long Beach, CA, USA: Computer Vision Foundation/IEEE), 4690–4699. doi:10.1109/CVPR.2019.00482

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. (2020). "Momentum contrast for unsupervised visual representation learning," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, June 13-19, 2020 (Seattle, WA, USA: Computer Vision Foundation/IEEE), 9726–9735. doi:10.1109/CVPR42600.2020.00975

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, June 27-30, 2016 (Las Vegas, NV, USA: IEEE Computer Society), 770–778. doi:10.1109/CVPR.2016.90

Hénaff, O. J. (2020). "Data-efficient image recognition with contrastive predictive coding," in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual Event (PMLR), vol. 119 of Proceedings of Machine Learning Research, 13-18 July 2020 (Vienna, Austria: PMLR), 4182–4192.

Hu, R. L., Karnowski, J., Fadely, R., and Pommier, J. (2017). *Image segmentation to distinguish between overlapping human chromosomes*. CoRR abs/1712.07639.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, July 21-26, 2017 (Honolulu, HI, USA: IEEE Computer Society), 2261–2269. doi:10.1109/CVPR.2017.243

Kang, B., Li, Y., Xie, S., Yuan, Z., and Feng, J. (2021). "Exploring balanced feature spaces for representation learning," in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, May 3-7, 2021 (Austria: OpenReview.net).

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., et al. (2020). "Supervised contrastive learning," in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, virtual, December 6-12, 2020.

Lin, C., Zhao, G., Yin, A., Ding, B., Guo, L., and Chen, H. (2020). As-panet: A chromosome instance segmentation method based on improved path aggregation network architecture. *J. Image Graph.* 25, 2271–2280.

Lin, C., Zhao, G., Yin, A., Yang, Z., Guo, L., Chen, H., et al. (2021). A novel chromosome cluster types identification method using resnext WSL model. *Med. Image Anal.* 69, 101943. doi:10.1016/j.media.2020.101943

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. (2017). "Sphereface: Deep hypersphere embedding for face recognition," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, July 21-26, 2017 (Honolulu, HI, USA: IEEE Computer Society), 6738–6746. doi:10.1109/CVPR.2017.713

Liu, W., Wen, Y., Yu, Z., and Yang, M. (2016). "Large-margin softmax loss for convolutional neural networks," in Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, June 19-24, 2016 (New York City, NY, USA: JMLR.org), 507–516. vol. 48 of JMLR Workshop and Conference Proceedings.

Mahajan, D., Girshick, R. B., Ramanathan, V., He, K., Paluri, M., Li, Y., et al. (2018). "Exploring the limits of weakly supervised pretraining," in Computer Vision - ECCV 2018 - 15th European Conference, September 8-14, 2018 (Munich, Germany: Springer), 185–201. Proceedings, Part II. vol. 11206 of Lecture Notes in Computer Science. doi:10.1007/978-3-030-01216-8_12

Minaee, S., Fotouhi, M., and Khalaj, B. H. (2014). "A geometric approach to fully automatic chromosome segmentation," in 2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), 20 July 2015 (Philadelphia, PA, USA: IEEE), 1–6. doi:10.1109/SPMB.2014.7163174

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, December 8-14, 2019, 8024–8035.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, June 27-30, 2016 (Las Vegas, NV, USA: IEEE Computer Society), 2818–2826. doi:10.1109/CVPR.2016.308

van den Oord, A., Li, Y., and Vinyals, O. (2018). *Representation learning with contrastive predictive coding*. CoRR abs/1807.03748.

Wang, F., Cheng, J., Liu, W., and Liu, H. (2018a). Additive margin softmax for face verification. *IEEE Signal Process. Lett.* 25, 926–930. doi:10.1109/LSP.2018.2822810

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., et al. (2018b). "Cosface: Large margin cosine loss for deep face recognition," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, June 18-22, 2018 (Salt Lake City, UT, USA: Computer Vision Foundation/IEEE Computer Society), 5265–5274. doi:10.1109/CVPR.2018.00552

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. (2018). "Unsupervised feature learning via non-parametric instance discrimination," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, June 18-22, 2018 (Salt Lake City, UT, USA: Computer Vision Foundation/IEEE Computer Society), 3733–3742. doi:10.1109/CVPR.2018.00393

Xie, S., Girshick, R. B., Dollár, P., Tu, Z., and He, K. (2017). "Aggregated residual transformations for deep neural networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, July 21-26, 2017 (Honolulu, HI, USA: IEEE Computer Society), 5987–5995. doi:10.1109/CVPR.2017.634

Yilmaz, I. C., Yang, J., Altinsoy, E., and Zhou, L. (2018). "An improved segmentation for raw g-band chromosome images," in 5th International Conference on Systems and Informatics, ICSAI 2018, November 10-12, 2018 (Nanjing, China: IEEE), 944–950. doi:10.1109/ICSAI.2018.8599328

Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., and Choe, J. (2019). "Cutmix: Regularization strategy to train strong classifiers with localizable features," in 2019 IEEE/CVF International Conference on Computer Vision, ICCV, October 27 - November 2, 2019 (Seoul, Korea (South): IEEE), 6022–6031. doi:10.1109/ICCV.2019.00612

Zhang, H., Cissé, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). "mixup: Beyond empirical risk minimization," in 6th International Conference on Learning Representations, ICLR 2018, April 30 - May 3, 2018 (Vancouver, BC, Canada: OpenReview.net). Conference Track Proceedings.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 18 April 2023
DOI 10.3389/fgene.2023.1133775

OPEN ACCESS

EDITED BY
Chunyu Wang,
Harbin Institute of Technology, China

REVIEWED BY
Antonio Neme,
National Autonomous University of Mexico, Mexico
Ru Xiaoqing,
University of Tsukuba, Japan

*CORRESPONDENCE
Wen Zhu,
syzhuwen@163.com

RECEIVED 29 December 2022
ACCEPTED 04 April 2023
PUBLISHED 18 April 2023

CITATION
Zheng L, Liu L, Zhu W, Ding Y and Wu F (2023), Predicting enhancer-promoter interaction based on epigenomic signals. *Front. Genet.* 14:1133775. doi: 10.3389/fgene.2023.1133775

Predicting enhancer-promoter interaction based on epigenomic signals

Leqiong Zheng[1,2,3], Li Liu[2], Wen Zhu[1,3]*, Yijie Ding[3] and Fangxiang Wu[1]

[1]School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [2]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, [3]Key Laboratory of Computational Science and Application of Hainan Province, Haikou, China

Introduction: The physical interactions between enhancers and promoters are often involved in gene transcriptional regulation. High tissue-specific enhancer-promoter interactions (EPIs) are responsible for the differential expression of genes. Experimental methods are time-consuming and labor-intensive in measuring EPIs. An alternative approach, machine learning, has been widely used to predict EPIs. However, most existing machine learning methods require a large number of functional genomic and epigenomic features as input, which limits the application to different cell lines.

Methods: In this paper, we developed a random forest model, HARD (H3K27ac, ATAC-seq, RAD21, and Distance), to predict EPI using only four types of features.

Results: Independent tests on a benchmark dataset showed that HARD outperforms other models with the fewest features.

Discussion: Our results revealed that chromatin accessibility and the binding of cohesin are important for cell-line-specific EPIs. Furthermore, we trained the HARD model in the GM12878 cell line and performed testing in the HeLa cell line. The cross-cell-lines prediction also performs well, suggesting it has the potential to be applied to other cell lines.

KEYWORDS

enhancer-promoter interaction, machine learning, ChIA-PET, random forest, epigenomic signals

1 Introduction

Enhancers and promoters are two of the most critical regulatory elements of gene transcription in the eukaryotic genome (Maston et al., 2006). The physical interactions between them precisely regulate spatiotemporal gene expression, which contributes to complex cellular functions. Aberrant connections between enhancers and promoters may lead to abnormal expression of disease-related genes (Krijger and De Laat, 2016). Therefore, the study of how enhancers and promoters interact can improve our understanding of health and disease. The primary mechanism of enhancer-promoter interaction is chromatin looping (Rubtsov et al., 2006; Miele and Dekker, 2008), which allows distal enhancers to contact the target gene promoters in three-dimensional space (Lv et al., 2021). Such long-range regulatory interactions play a significant role in tissue-specific gene expression (Maston et al., 2006; De Laat and Duboule, 2013) and can link the regulatory element to the target gene (Corradin et al., 2014). In recent decades, the identification of EPIs has relied

on high-throughput experimental techniques, such as chromosome conformation capture (3C) (Dekker et al., 2002), 4C (Splinter et al., 2012), 5C (Sanyal et al., 2012), Hi-C (Lieberman-Aiden et al., 2009), Hi-C capture (Schoenfelder et al., 2015), DNase-Hi-C (Ma et al., 2015), and ChIA-PET (Li et al., 2012; Heidari et al., 2014). These experimental approaches are effective in identifying EPIs but are time-consuming and expensive (Ecker et al., 2012). Thus, a more cost-effective method is required for predicting enhancer-promoter interactions. To address this problem, machine learning methods are used to predict EPIs by using available genomic or epigenomic data.

Many deep learning methods have been proposed for predicting EPIs based on DNA sequence, including SPEID, SIMCNN, and EPIVAN. SPEID (Singh et al., 2019) and SIMCNN (Zhuang et al., 2019) employ CNN-based approaches, while EPIVAN (Hong et al., 2020) incorporates an attention mechanism for improved prediction accuracy. Although they achieved good results using only DNA sequences, the cell-line-specific nature of EPIs (Heidari et al., 2014; Ma et al., 2015) presents a challenge (Lv et al., 2021; Ao et al., 2022a). For instance, the same pair of enhancer and promoter contacts in some cell lines, but not in others, despite the DNA sequences have not changed (Schöler and Gruss, 1984). To address this issue, several models have been developed to identify cell-line-specific EPIs using epigenomic signals, including chromatin accessibility, the binding of special transcription factors, and histone modification levels. For example, RIPPLE (Roy et al., 2015) provides a systematic approach for predicting and interpreting

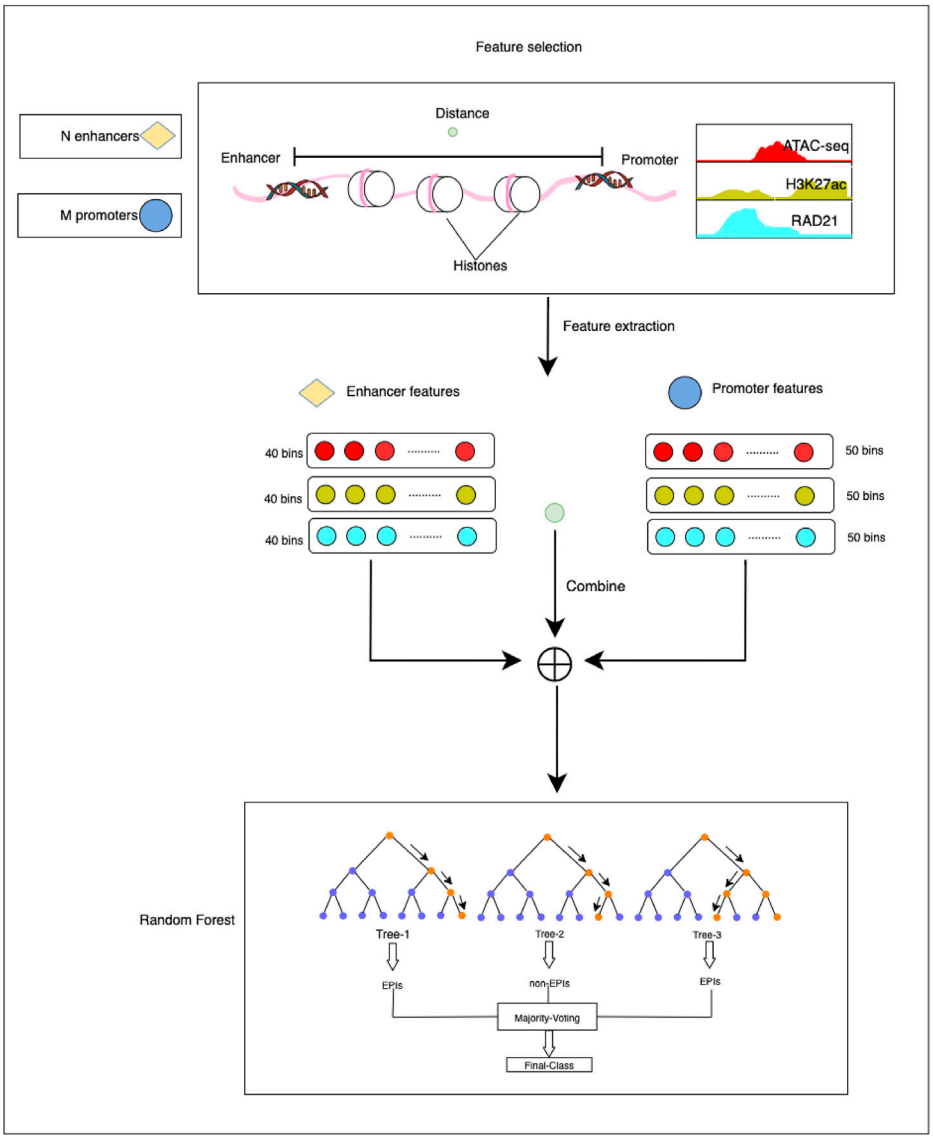


FIGURE 1
The overall framework of the HARD model. First, ATAC-seq, H3K27ac, and RAD21 epigenomic signals were selected as essential features to predict EPIs. Then, the enhancer and promoter were divided into 40 and 50 bins, respectively, with 50 bp per bin. Deeptools was used to extract the epigenomic signals. The epigenomic signal matrix was combined with the distance between the enhancer and promoter. Finally, we input the final feature matrix to the random forest learning machine for training and testing.

TABLE 1 Distribution of samples.

Data set	Positive samples	Negative samples
GM12878 training	6251	25,005
GM12878 test	1563	6251
Hela	347	1388

EPIs in a cell-line-specific manner using a variety of epigenomic features. However, many epigenomic signals are not available for all cell lines.

Based on the aforementioned analyses, we considered using as few epigenomic features as possible to build machine learning models to predict cell-line-specific EPIs. Loose chromatin is a prerequisite for loop formation. The H3K27ac ChIP-seq and ATAC-seq data are often used to represent chromatin accessibility. Chromatin interaction decays with distance. RAD21 is a subunit of cohesin that play important role in a loop formation. Therefore, the four types of features were extracted to train the models. By comparing several machine learning classifiers, the random forest was selected due to the high accuracy. Finally, we compared our HARD model with the sequence-based and other epigenomic features-based models. The results showed that our model outperformed them both in the same cell line and cross-cell-lines.

2 Materials and methods

The HARD model consists of three primary steps: 1) constructing positive and negative sets based on the benchmark database. 2) Extracting epigenomic features that can influence the formation of EPI. 3) predicting EPIs within the same cell line and across cell lines (Figure 1).

2.1 Data collection and processing

The enhancer-promoter interaction data were obtained from the BENGI (Moore et al., 2020) database. To construct a benchmark of enhancer-promoter interactions, BENGI integrated various experimental datasets, such as Hi-C, ChIA-PET, genetic interactions (cis-eQTLs), and CRISPR/Cas9 perturbations. After removing ambiguous pairs, we selected the RNAPII ChIAPET data of GM12878 and HeLa cell lines with a fixed positive and negative sample ratio. Both data have a positive-to-negative sample ratio of 1:4. To ensure the data is more accurate, the ambiguous interaction pairs were removed. The RNAPII ChIAPET data only provides the IDs of cCRE-ELS (cCREs with enhancer-like signatures) and TSS (transcription start site) without the position of cCRE-ELS and TSS. We located the cCRE-ELS and TSS in the genome according to the IDs of hg19-cCREs and GENCODEv19-TSS, respectively. Then, duplicate data was removed to retain unique data.

Next, 2,000 bp upstream and 500 bp downstream of the TSS were defined as the promoter region. For enhancers, upstream 1000 bp and downstream 1000 bp were extracted from the midpoint of the cCRE-ELS region. Ultimately, 39,070 pairs of enhancer-promoter interaction were obtained in the GM12878 dataset, and 1,735 pairs of enhancer-promoter interaction were obtained in the HeLa dataset. Then, the dataset was divided into a training set and a test set for the GM12878 sample. Specifically, 80% of the data was used for training, and the remaining 20% was used as an independent test set. To ensure consistency in data distribution across both datasets, the positive and negative sample ratios of both divided datasets were maintained at a 1:4 ratio. The above data processing part and the subsequent classification experiments were implemented in the python language environment, and the sklearn library is used. The detailed data distribution is shown in Table 1.

We selected three epigenomic signal features as our experimental features, including ATAC-seq, H3K27ac, and RAD21. The epigenomic signal data, which included ATAC-seq, H3K27ac, and RAD21, were obtained from the ENCODE (Ecker et al., 2012) database. The data with IDs ENCFF000XKM, ENCFF051PGV, and ENCFF706HLO corresponded to sequencing data in bigWig format of RAD21, ATAC-seq, and H3K27ac in the HeLa cell line, respectively. Similarly, the data with IDs ENCFF000WCT, ENCFF180ZAY, and ENCFF440GZA

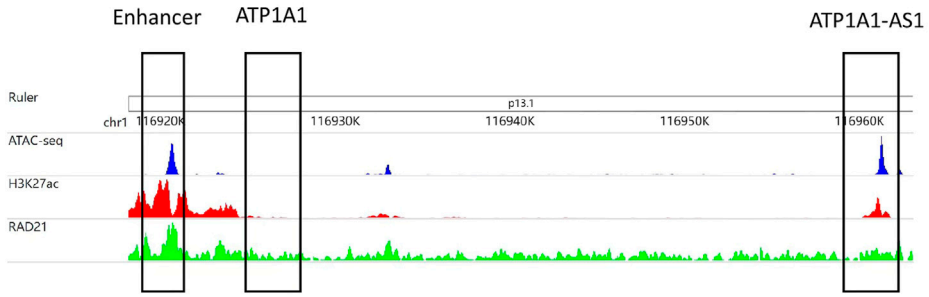


FIGURE 2
The area chr1:116,919,153–116,921,153 selected in the first matrix box is an enhancer subarea. The second matrix box selected region cr1: 116,924,718–116,927,218 is the promoter region of the ATP1A1 gene. The third matrix box selected region chr1:116,959,158–116,961,658 is the promoter region of the ATP1A1-AS1 gene. The three tracks in the figure were generated from the bigWig data of ATAC-seq, H3K27ac and RAD21 of GM12878 cell lines.

TABLE 2 Comparison of the predictive EPI performance of each classifier in the GM12878 cell line.

Classifier	Sn	Sp	Precision	Acc	AUC	AUPRC
RF	**0.578**	**0.964**	**0.799**	**0.887**	**0.919**	**0.773**
Adaboost	0.555	0.947	0.725	0.869	0.881	0.688
GBDT	0.568	0.955	0.759	0.878	0.896	0.739

The meaning of bold values is the highest value of a specific performance indicator under different classifiers.

TABLE 3 Comparison of HARD, EPIVAN and RF (10) models in the GM12878 cell line.

Classifier	Sn	Sp	Precision	Acc	AUC	AUPRC
HARD	0.578	0.964	**0.799**	**0.887**	**0.919**	**0.773**
EPIVAN	0.365	**0.971**	0.720	0.850	0.809	0.603
RF (10)	**0.709**	0.730	0.396	0.726	0.799	0.540

The meaning of bold values is the highest value of a specific performance indicator under different classifiers.

corresponded to sequencing data in bigWig format of RAD21, ATAC-seq, and H3K27ac in the GM12878 cell line, respectively.

2.2 Feature extraction

The above-mentioned features were extracted through the following steps. First, the genomic site data of EPIs and epigenomic signal data were imported into deeptools (Ramírez et al., 2014), a bioinformatics tool used for feature extraction. Then the enhancer and promoter regions were divided into bins of 50 bp. Each enhancer region was further divided into 40 bins, whereas each promoter region was divided into 50 bins. For ATAC-seq, H3K27ac, and RAD21, it generated a signal value for each bin. Following feature extraction, the enhancers and promoters were represented by 120-dimensional and 150-dimensional feature vectors, respectively. The distance is defined as the number of base pairs from the midpoint of the enhancer to the midpoint of the promoter. The epigenomic feature vector and distance feature vector were concatenated to obtain the final feature matrix. This step involved combining the feature vectors obtained from the enhancer and promoter regions into a single matrix, with each row of the matrix representing a pair of enhancer-promoter interactions. The final feature matrix was then used as input for the classification experiments.

2.3 Classification algorithms

We compared three classifiers, random forest (RF), AdaBoost, and gradient boosting decision tree (GBDT), for predicting EPIs in the GM12878 cell line, which is considered a binary classification problem. All three classifiers proved to be efficient in solving binary classification problems.

Random forest (Breiman, 2001) is an ensemble learning algorithm. It uses multiple decision trees to classify data by randomly selecting data and feature subsets, which helps to reduce the model's variance and overfitting risk. By voting or averaging the outputs of multiple decision trees, the model reduces the error rate and improves accuracy. In the experiment, a large amount of sample data was used, and setting the number of decision trees to 100 produced optimal performance.

AdaBoost (Schapire, 2013) assembles multiple weak classifiers to build a strong classifier, which applies to binary classification problems and has been shown to perform well on complex datasets. The algorithm assigns weights to each instance based on

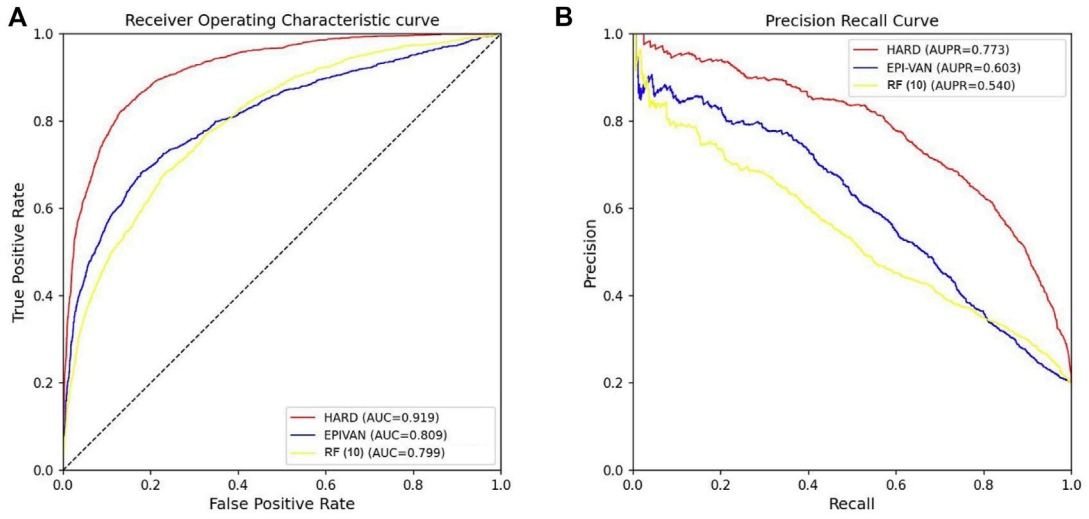


FIGURE 3
Comparison of the AUC and AUPRC performance of the three models tested independently in the GM12878 cell line. **(A)** The red curve is the ROC curve of the HARD model, the blue curve is the ROC curve of the EPIVAN model, and the yellow curve is the ROC curve of the RF (10) model; **(B)** The red curve is the PRC curve of the HARD model, the blue curve is the PRC curve of the EPIVAN model, and the yellow curve is the PRC curve of the RF (10) model.

TABLE 4 Comparison of HARD, EPIVAN and RF (10) models in the HeLa cell line.

Classifier	Sn	Sp	Precision	Acc	AUC	AUPRC
HARD	0.363	**0.953**	**0.660**	**0.836**	**0.831**	**0.601**
EPIVAN	**0.513**	0.890	0.539	0.815	0.795	0.564
RF (10)	0.144	0.949	0.402	0.786	0.572	0.296

The meaning of bold values is the highest value of a specific performance indicator under different classifiers.

its difficulty level and trains weak classifiers on the weighted data. Misclassified instances have increased weight, while correctly classified instances have decreased weight. This process is repeated multiple times until the ensemble classifier reaches a satisfactory level.

Gradient boosting decision tree (Friedman, 2001) builds a model by summing multiple decision trees. It optimizes the model iteratively by adding a new decision tree that reduces the prediction error of the previous trees. The model's accuracy improves with each iteration, making it suitable for binary classification problems. In the experiment, n_estimators, learning_rate, and subsample were set to 100, 0.1, and 1, respectively.

2.4 Performance evaluation

To evaluate the classification performance of the selected features and classifiers, we used six metrics: sensitivity (Sn) (Swift et al., 2020), specificity (Sp) (Swift et al., 2020), precision (Hong et al., 2020; Chen et al., 2023), accuracy (Shao et al., 2020; Yu et al., 2022), the area under the curve (AUC) (Myerson et al., 2001), and the area under the precision-recall curve (AUPRC) (Ozenne et al., 2015). These metrics serve as the basis for evaluation, and the relevant formulas for their calculation are shown below.

$$Sn = recall = TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$Sp = \frac{TN}{TN + FP} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

In binary classification, there are four possible outcomes: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). TP corresponds to the cases where the classifier correctly predicts the positive class, while FP corresponds to the instances where the classifier incorrectly predicts the positive class. Similarly, FN refers to the cases where the classifier incorrectly predicts the negative class, and TN refers to the instances where the classifier correctly predicts the negative class. Additionally, TPR (sensitivity/recall) is the ratio of correctly identified positive instances to the actual positive instances, while FPR is the proportion of falsely identified positive instances to the actual negative instances (Zeng et al., 2020). AUC is calculated by plotting TPR against FPR at different thresholds and represents the area under the resulting curve. AUPRC is calculated by plotting precision against recall at different thresholds and represents the area under the resulting curve.

3 Results and discussion

3.1 The features of HARD model are closely related with EPI

The accessibility of chromatin structural regions is associated with the regulation of gene expression. ATAC-seq is commonly

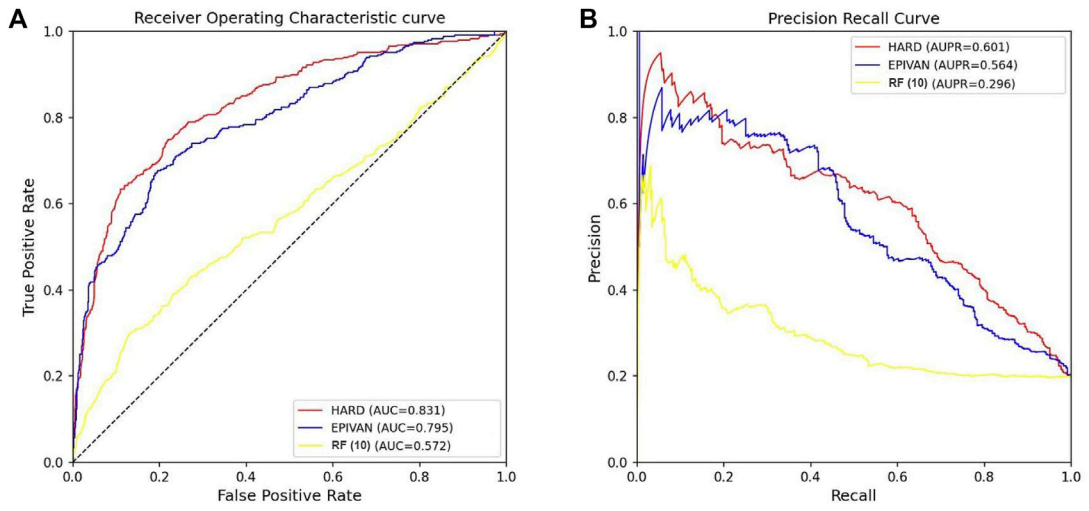


FIGURE 4
Comparison of AUC and AUPRC performance of the three models in the HeLa cell line. **(A)** The red curve is the ROC curve of the HARD model, the blue curve is the ROC curve of the EPIVAN model, and the yellow curve is the ROC curve of the RF (10) model; **(B)** The red curve is the PRC curve of the HARD model, the blue curve is the PRC curve of the EPIVAN model, and the yellow curve is the PRC curve of the RF (10) model.

used to detect open regions of chromatin across the genome. When combined with activated histone modification, such as H3K27ac, ATAC-seq can enable the identification of specific effects on gene expression (Bravo González-Blas et al., 2019). H3K27ac is primarily enriched in enhancer and promoter regions (Herrera-Uribe et al., 2020) and is associated with gene activation (Yan et al., 2019). RAD21 and the insulator-binding protein CTCF bind to highly conserved promoters and distal enhancers, contributing to transcriptional regulation (Whalen et al., 2016; Liu et al., 2021). Numerous studies have shown that distance is a useful factor for studying EPI (Bianco et al., 2018; Al Bkhetan et al., 2019). The distance feature has an essential contribution to many models (Moore et al., 2020; Ao et al., 2022b).

Figure 2 is an example that epigenomic modification influences the formation of EPI. The enhancer region (chr1: 116,919,153–116,921,153) interacts with the ATP1A1-AS1 promoter (chr1:116,959,158–116,961,658) and does not interact with the ATP1A1 promoter (chr1: 116,959,158–116,961,658), according to RNAPII ChIAPET data of GM12878 cell line. In the enhancer region, the signals of ATAC-seq, H3K27ac, and RAD21 are enriched, which indicates that the enhancer is highly activated. The promoter region of ATP1A1-AS1 is enriched in ATAC-seq, H3K27ac modifications, and RAD21 binding, whereas the promoter region of ATP1A1 is not.

3.2 Comparison and selection of classifiers

To select the most accurate classifier, we compared three classifiers, AdaBoost, GBDT, and RF. We trained the model using 31,256 GM12878 samples with ten-fold cross-validation and evaluated its performance on an independent test set of 7,814 GM12878 samples. The classifiers were trained and tested separately, and their performance was compared using different metrics. A comparison of the metrics of the test set is shown in Table 2. Results showed that the RF algorithm outperformed both GBDT and AdaBoost in all metrics. Specifically, the RF algorithm demonstrated higher Sn, Sp, precision, accuracy, AUC, and AUPRC values, at 0.578, 0.964, 0.799, 0.887, 0.919, and 0.773, respectively. Notably, the RF algorithm displayed superior performance in AUPRC and precision metrics. The RF algorithm merges the strengths of ensemble learning and tree models, and it is capable of balancing the error for an unbalanced set of classifiers, making it a suitable choice for the dataset at hand. Consequently, the HARD model was constructed using the RF algorithm.

3.3 Comparison with other models in GM12878 cell line

In order to verify the validity of the HARD model, we next compared the performance of HARD against the sequence-based and other epigenomic features-based models. EPIVAN is a typical representative of sequence-based models, which outperforms the majority of existing models. RIPPLE utilizes many epigenomic features to predict EPI. These epigenomic features include cohesin (RAD21), architectural proteins (CTCF), marks associated with active gene bodies and elongation (H3K36me3, H4K20me1), activating marks of transcription (H3K4me2, H3K27ac, and H3K9ac), open chromatin (DNase I), a repressive mark (H3K27me3), and a general transcription factor (TBP). Here, we used ten available features of RIPPLE to conduct a RF classification model, named RF (10). Then the HARD model was compared with RF (10) and EPIVAN in multiple aspects. We trained the models using 31,256 GM12878 sample data with ten-fold cross-validation and evaluated them using an independent test set of 7,814 GM12878 samples. The comparison results are shown in Table 3. The results indicated that RF (10) performed best in terms of Sn, while EPIVAN produced the best results for Sp. However, each model has its strengths and weaknesses in terms of Sn and Sp. HARD had shown significant improvement in all four performance metrics compared to other models. Specifically, compared to EPIVAN, HARD shows an improvement of 7.9% and 3.7% in precision and Acc, respectively, as well as an increase of 11% and 17% in AUC and AUPRC, respectively. Compared to the RF (10), HARD shows greater improvements, with increases of 40.3%, 16.1%, 12%, and 23.3% in precision, Acc, AUC, and AUPRC, respectively. The comparison of the AUC and ROC curves of the three models is shown in Figure 3.

3.4 Comparison of the HARD, EPIVAN and RF (10) model in cross-cell-lines

To verify the robustness of the models, we conducted a cross-cell-line analysis by training the models on the GM12878 cell line and testing them on the HeLa cell line. We used 39,070 GM12878 samples as the training set for ten-fold cross-validation, and 1,735 HeLa samples as the test set for evaluation. Experiments were implemented for the HARD, EPIVAN and RF (10) models, respectively. Among the three models, HARD achieved the best performance in terms of Sp, precision, accuracy, AUC, and AUPRC, followed by EPIVAN, with RF (10) showing the worst performance. In comparison to EPIVAN, the HARD model slightly improves five metrics, only lower than EPIVAN in Sn. The HARD model outperforms RF (10) by a significant margin (Table 4). The comparison of the AUC and ROC curves of the three models is shown in Figure 4. Results indicated that HARD outperformed EPIVAN and RF (10) in predicting EPIs in cross-cell-lines.

4 Conclusion

The interaction between enhancer and promoter is a complex process. Various genomic and epigenomic features are related to EPI. Many machine learning models have been developed to predict EPI based on a large number of genomic and epigenomic features. The redundancy of features leads to unsatisfactory experimental results and limits the application to more cell lines. In this paper, we developed the HARD model,

which employed a minimal number of epigenomic features to predict cell-line-specific EPIs. It is noteworthy that the HARD model is based on benchmark data from the BENGI database, which defined EPI strictly by integrating ChIA-PET, genetic interactions (cis-eQTLs), and CRISPR/Cas9 perturbations. By comparing with two other models, we found HARD outperformed them both in the same cell line and cross-cell-lines. Importantly, our model only used H3K27ac, ATAC-seq, RAD21, and Distance as input, which makes it possible to apply to more cell lines.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

LZ: Investigation, Methodology, Writing—Original draft preparation. LL: Conceptualization, Funding acquisition, Writing—Review & Editing. WZ: Conceptualization, Project administration, Funding acquisition. YD: Methodology, Writing—Review & Editing. FW: Investigation, Methodology.

Funding

This work was supported by the National Nature Science Foundation of China (Grant Nos 61863010, 11926205, 11926412, 61873076, and 61961031), National Key R and D Program of China (No. 2020YFB2104400), Natural Science Foundation of Hainan, China (Grant Nos 121RC538, 119MS036, and 120RC588), Key Laboratory of Computational Science and Application of Hainan Province (No. JSKX202201), and the Municipal Government of Quzhou (NO. 2022D017).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Al Bkhetan, Z., Kadlof, M., Kraft, A., and Plewczyński, D. (2019). Machine learning polymer models of three-dimensional chromatin organization in human lymphoblastoid cells. *Methods* 166, 83–90. doi:10.1016/j.ymeth.2019.03.002

Ao, C., Jiao, S., Wang, Y., Yu, L., and Zou, Q. (2022a). Biological sequence classification: A review on data and general methods. *Research* 24, 1198. doi:10.1093/bioinformatics/btn089

Ao, C., Zou, Q., and Yu, L. (2022b). NmRF: Identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences. *Briefings Bioinforma.* 23, bbab480. doi:10.1093/bib/bbab480

Bianco, S., Lupiáñez, D. G., Chiariello, A. M., Annunziatella, C., Kraft, K., Schöpflin, R., et al. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat. Genet.* 50, 662–667. doi:10.1038/s41588-018-0098-8

Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., et al. (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. methods* 16, 397–400. doi:10.1038/s41592-019-0367-1

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Chen, L., Yu, L., and Gao, L. (2023). Potent antibiotic design via guided search from antibacterial activity evaluations. *Bioinformatics* 39, btad059. doi:10.1093/bioinformatics/btad059

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal, R., et al. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* 24, 1–13. doi:10.1101/gr.164079.113

De Laat, W., and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* 502, 499–506. doi:10.1038/nature12753

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *science* 295, 1306–1311. doi:10.1126/science.1067799

Ecker, J. R., Bickmore, W. A., Barroso, I., Pritchard, J. K., Gilad, Y., and Segal, E. (2012). Genomics: ENCODE explained. *Nature* 489, 52–55. doi:10.1038/489052a

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. statistics* 29, 1189–1232. doi:10.1214/aos/1013203451

Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., et al. (2014). Genome-wide map of regulatory interactions in the human genome. *Genome Res.* 24, 1905–1917. doi:10.1101/gr.176586.114

Herrera-Uribe, J., Liu, H., Byrne, K. A., Bond, Z. F., Loving, C. L., and Tuggle, C. K. (2020). Changes in H3K27ac at gene regulatory regions in porcine alveolar macrophages following LPS or PolyIC exposure. *Front. Genet.* 11, 817. doi:10.3389/fgene.2020.00817

Hong, Z., Zeng, X., Wei, L., and Liu, X. (2020). Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinforma. Oxf. Engl.* 36, 1037–1043. doi:10.1093/bioinformatics/btz694

Krijger, P. H. L., and De Laat, W. (2016). Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* 17, 771–782. doi:10.1038/nrm.2016.138

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148, 84–98. doi:10.1016/j.cell.2011.12.014

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* 326, 289–293. doi:10.1126/science.1181369

Liu, L., Zhang, L.-R., Dao, F.-Y., Yang, Y.-C., and Lin, H. (2021). A computational framework for identifying the transcription factors involved in enhancer-promoter loop formation. *Mol. Therapy-Nucleic Acids* 23, 347–354. doi:10.1016/j.omtn.2020.11.011

Lv, H., Dao, F.-Y., Zulfiqar, H., Su, W., Ding, H., Liu, L., et al. (2021). A sequence-based deep learning approach to predict CTCF-mediated chromatin loop. *Briefings Bioinforma.* 22, bbab031. doi:10.1093/bib/bbab031

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., et al. (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat. methods* 12, 71–78. doi:10.1038/nmeth.3205

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* 7, 29–59. doi:10.1146/annurev.genom.7.080505.115623

Miele, A., and Dekker, J. (2008). Long-range chromosomal interactions and gene regulation. *Mol. Biosyst.* 4, 1046–1057. doi:10.1039/b803580f

Moore, J. E., Pratt, H. E., Purcaro, M. J., and Weng, Z. (2020). A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol.* 21, 17–16. doi:10.1186/s13059-019-1924-8

Myerson, J., Green, L., and Warusawitharana, M. (2001). Area under the curve as a measure of discounting. *J. Exp. analysis Behav.* 76, 235–243. doi:10.1901/jeab.2001.76-235

Ozenne, B., Subtil, F., and Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J. Clin. Epidemiol.* 68, 855–859. doi:10.1016/j.jclinepi.2015.02.010

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids Res.* 42, W187–W191. doi:10.1093/nar/gku365

Roy, S., Siahpirani, A. F., Chasman, D., Knaack, S., Ay, F., Stewart, R., et al. (2015). A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic acids Res.* 43, 8694–8712. doi:10.1093/nar/gkv865

Rubtsov, M. A., Polikanov, Y. S., Bondarenko, V. A., Wang, Y.-H., and Studitsky, V. M. (2006). Chromatin structure can strongly facilitate enhancer action over a distance. *Proc. Natl. Acad. Sci.* 103, 17690–17695. doi:10.1073/pnas.0603819103

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature* 489, 109–113. doi:10.1038/nature11279

Schapire, R. E. (2013). "Explaining adaboost," in *Empirical inference: Festschrift in honor of vladimir N. Vapnik* (Berlin, Germany: Springer), 37–52.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* 25, 582–597. doi:10.1101/gr.185272.114

Schöler, H. R., and Gruss, P. (1984). Specific interaction between enhancer-containing molecules and cellular components. *Cell* 36, 403–411. doi:10.1016/0092-8674(84)90233-2

Shao, J., Yan, K., and Liu, B. (2020). FoldRec-C2C: Protein fold recognition by combining cluster-to-cluster model and protein similarity network. *Briefings Bioinforma.* 22, bbaa144. doi:10.1093/bib/bbaa144

Singh, S., Yang, Y., Póczos, B., and Ma, J. (2019). Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quant. Biol.* 7, 122–137. doi:10.1007/s40484-019-0154-0

Splinter, E., Wit, E. D., Werken, H., Klous, P., and Laat, W. D. (2012). Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. *Methods* 58, 221–230. doi:10.1016/j.ymeth.2012.04.009

Swift, A., Heale, R., and Twycross, A. (2020). What are sensitivity and specificity? *Evidence-Based Nurs.* 23, 2–4. doi:10.1136/ebnurs-2019-103225

Whalen, S., Truty, R. M., and Pollard, K. S. (2016). Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* 48, 488–496. doi:10.1038/ng.3539

Yan, W., Chen, D., Schumacher, J., Durantini, D., Engelhorn, J., Chen, M., et al. (2019). Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. *Nat. Commun.* 10, 1705–1716. doi:10.1038/s41467-019-09513-2

Yu, L., Zheng, Y., and Gao, L. (2022). MiRNA–disease association prediction based on meta-paths. *Briefings Bioinforma.* 23, bbab571. doi:10.1093/bib/bbab571

Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., et al. (2020). Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* 11, 1775–1797. doi:10.1039/c9sc04336e

Zhuang, Z., Shen, X., and Pan, W. (2019). A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data. *Bioinformatics* 35, 2899–2906. doi:10.1093/bioinformatics/bty1050

frontiers | Frontiers in Genetics

TYPE Brief Research Report
PUBLISHED 27 June 2023
DOI 10.3389/fgene.2023.1217860

OPEN ACCESS

EDITED BY
Maria Luisa Chiusano,
University of Naples Federico II, Italy

REVIEWED BY
Ravi Madduri,
Argonne National Laboratory (DOE), United States
Dominik Grimm,
Weihenstephan-Triesdorf University of Applied Sciences, Germany

*CORRESPONDENCE
Dominik Heider,
dominik.heider@uni-marburg.de

RECEIVED 06 May 2023
ACCEPTED 13 June 2023
PUBLISHED 27 June 2023

CITATION
Klau JH, Maj C, Klinkhammer H, Krawitz PM, Mayr A, Hillmer AM, Schumacher J and Heider D (2023), AI-based multi-PRS models outperform classical single-PRS models.
Front. Genet. 14:1217860.
doi: 10.3389/fgene.2023.1217860

AI-based multi-PRS models outperform classical single-PRS models

Jan Henric Klau[1], Carlo Maj[2], Hannah Klinkhammer[3,4], Peter M. Krawitz[3], Andreas Mayr[4], Axel M. Hillmer[5], Johannes Schumacher[2] and Dominik Heider[1]*

[1]Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany, [2]Center for Human Genetics, University of Marburg, Marburg, Germany, [3]Institute for Genomic Statistics and Bioinformatics, Medical Faculty, University Bonn, Bonn, Germany, [4]Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University Bonn, Bonn, Germany, [5]Institute of Pathology, Faculty of Medicine, University of Cologne, Cologne, Germany

Polygenic risk scores (PRS) calculate the risk for a specific disease based on the weighted sum of associated alleles from different genetic loci in the germline estimated by regression models. Recent advances in genetics made it possible to create polygenic predictors of complex human traits, including risks for many important complex diseases, such as cancer, diabetes, or cardiovascular diseases, typically influenced by many genetic variants, each of which has a negligible effect on overall risk. In the current study, we analyzed whether adding additional PRS from other diseases to the prediction models and replacing the regressions with machine learning models can improve overall predictive performance. Results showed that multi-PRS models outperform single-PRS models significantly on different diseases. Moreover, replacing regression models with machine learning models, i.e., deep learning, can also improve overall accuracy.

KEYWORDS

polygenic risk score, machine learning, deep learning, breast cancer, regression

1 Introduction

Disease prevention is a crucial part of medical care. It reduces the costs for the healthcare system and reduces the number of hospitalization and deaths (Kahn et al., 2008). For targeted preventive measures, it is necessary to determine the individual risks for certain diseases. In addition to age, sex, and lifestyle, genetic factors play an important role in determining the individual risk. Polygenic risk scores (PRS) are used to take multivariate genomic information into consideration and can be used for the selection of a targeted treatment in personalized medicine (Lambert et al., 2019; Lewis and Vassos, 2020; Schröder et al., 2022).

PRS are typically modeled as a regression task by calculating a weighted sum of all genotypes and their corresponding estimated effect size. Relevant single nucleotide polymorphisms are discovered by genome-wide association studies (GWAS). For individual risk prediction, another regression model is built based on the previously calculated PRS and other covariates, such as age, sex, and lifestyle (e.g., smoking and alcohol consumption) (Choi et al., 2020).

In recent years, machine learning (ML) has led to numerous advances in medicine (MacEachern and Forkert, 2021) due to the ability to train models on complex problems and being able to handle large amounts of data. These models have been used in various

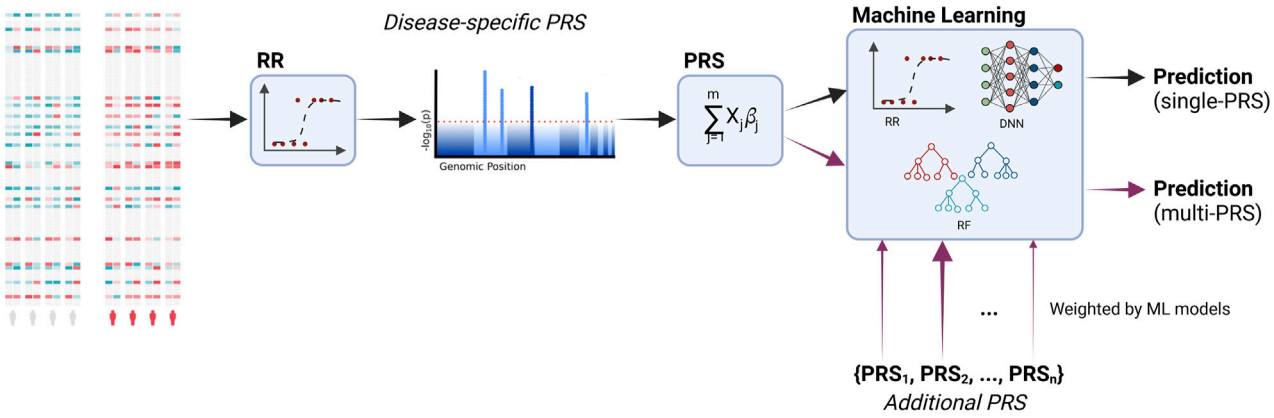


FIGURE 1
Workflow of the study. PRS are calculated based on the associated genetic loci (i.e., SNPs, single-nucleotide polymorphisms). Significant loci are identified via a regression model. These loci are then used to calculate the PRS based on a linear combination. Additional PRS for other diseases are incorporated into the final predictive model. During training, the models learn to distinguish between relevant and irrelevant features, including the additional PRS. Moreover, we compare the typically used ridge regression with machine learning models, namely, deep neural networks and random forests. Created with BioRender.com.

applications, e.g., oncology (Bibault et al., 2016), pathology (Madabhushi and Lee, 2016; Coudray et al., 2018), diabetes (Spänig et al., 2019), human genetics (Libbrecht and Noble, 2015), and infectious diseases (Riemenschneider et al., 2016b; Ren et al., 2021) as part of a growing trend toward personalized/precision medicine.

In this study, we trained multiple models, i.e., ridge regression (RR), random forests (RFs), and deep neural networks (DNNs), to predict an individual's phenotype for the following diseases: breast cancer (BC), coronary artery disease (CAD), and type 2 diabetes (T2D). We selected those three common chronic diseases to demonstrate the usefulness of our approach for different diseases. For instance, breast cancer is diagnosed in approximately 2.3 million women yearly. Cardiovascular diseases are the leading cause of death globally. Coronary artery disease affects approximately 126 million individuals, with 7.2 million deaths each year. Diabetes affects approximately 425 million people worldwide.

The inclusion of additional PRS has been shown to improve the prediction of traits and diseases (Krapohl et al., 2017) (Sinnott-Armstrong et al., 2021) (Abraham et al., 2019), psychological diseases, such as schizophrenia, bipolar disorder, or depression (Rodriguez et al., 2022), the risk of exposure to bullying (Schoeler et al., 2019), and hazard ratios (Meisner et al., 2020). Thus, we further evaluated the inclusion of 139 additional PRS in a multi-PRS approach to the prediction of the previously mentioned diseases. The additionally used PRS do not have to be directly associated with the investigated disease (Sinnott-Armstrong et al., 2021). Including these PRS, even if the phenotypes appear to be unrelated, may be beneficial as similar underlying biological mechanisms may be involved.

2 Materials and methods

The workflow of the current study is shown in Figure 1. We incorporated additional PRS into the predictive models and, additionally, compared different machine learning models to the regression models that are typically used in PRS.

2.1 Data

This research has been conducted using the UK Biobank resource (Bycroft et al., 2018) under application number 81202. The UK Biobank is a large-scale cohort study covering a huge prospective sample (n > 500,000) of the British general population, including both genotype and phenotype (health-related outcomes) data. We used the imputed UK Biobank data which include ~96 million variants.

We excluded available genotype data outliers for heterozygosity (F within three standard deviations (SD) from the mean), sample genotype missing rates (>2%), and discordant reported sex vs. genotypic sex. Allele frequency MAF < 0.1% was removed. Variants not in the Hardy–Weinberg equilibrium (p-value $<10^{-6}$) were excluded.

In total, 139 PRS (Supplementary Table S1) for different phenotypes, e.g., lung cancer (PGS000078), venous thromboembolism (PGS000043), and fasting glucose (PGS000305), were computed using PLINK (Chang et al., 2015) score function, and the corresponding effect alleles and beta coefficients were retrieved from the PGS Catalog (https://www.pgscatalog.org/). The PRS are therefore based on a linear additive combination of effect alleles and are characterized by a normal distribution. Due to the great abundance of SNPs in the imputed UK Biobank, adequate coverage was ensured.

The additional 139 PRS were added as additional input features without any pre-selection to enable a data-driven approach without any subject-matter knowledge. Therefore, we included all PRS that were available in the PGS Catalog at the time we started the project. The underlying idea is that different diseases can share different pathways, e.g., inflammatory pathways, or even comorbidities. Selection of PRS according to phenotype association with the

TABLE 1 Number of individuals in the case and control groups.

	BC (female only)	CAD	T2D
Cases	13,679	23,033	24,241
Controls	232,424	406,433	405,225

investigated disease, though more interpretable, can potentially miss relevant information. By using multiple risk scores, we were able to capture the interdependencies in a data-driven approach by machine learning models. PRS that were calculated on the same UK Biobank cohort for one of our target diseases could induce overfitting or circularity. For PRS that were calculated on the UK Biobank cohort, but for different diseases, this would only affect the control group. Therefore, these effects are, if at all, of very little impact.

From the phenotypic data, we derived the case/control status for three diseases, namely, BC, CAD, and T2D. BC cases were women based on self-report in an interview with a trained nurse and/or BC-related ICD-9 codes (174 or 174.9) or ICD-10 codes (C50.X) in hospitalization records. CAD cases were individuals with myocardial infarction based on self-report or hospital admission diagnosis according to ICD-9 codes of 410.X, 411.0, 412.X, or 429.79 or ICD-10 codes of I21.X, I22.X, I23.X, I24.1, or I25.2 in hospitalization records and/or with coronary artery bypass grafting (K40.1–40.4, K41.1–41.4, or K45.1–45.5) or coronary angioplasty with or without stenting (K49.1–49.2, K49.8–49.9, K50.2, K75.1–75.4, or K75.8–75.9). T2D cases were samples based on self-report in an interview with a trained nurse or an ICD-10 code of E11.X in hospitalization records. For controls, all individuals without the phenotype were considered (for BC, the analysis was restricted only to women).

In order to limit the confounding due to the genetic background, the analysis was restricted only to individuals with White British origin (Field 21000) and with European genetic ancestry according to the principal components provided by UK Biobank (Field 22006), and among the remaining samples, to account for the residual population stratification, we considered the principal components (PCs) as computed in UK Biobank (Field 22009). The total number of individuals in the data set amounts to 429,466, while the number of patients for the three diseases, BC, CAD, and T2D, are 13,679, 23,033, and 24,241, respectively (Table 1).

2.2 Data preparation

We included the following features into the model training: corresponding PRS (i.e., BC-PRS (PGS000015), CAD-PRS (PGS000013), or T2D-PRS (PGS000014), respectively), first 10 PCs, age, sex, and the genotyping array. Categorical features such as sex and genotyping array were one-hot encoded, while all other features were normalized to values between 0 and 1. For the prediction of BC, only female individuals were included, and sex was removed as an input feature. For the multi-PRS approach, 139 additional PRS (e.g., lung cancer (PGS000078), venous thromboembolism (PGS000043), and fasting glucose (PGS000305)) were included in the data set.

2.3 Model development

The data sets were split for each individual disease into training and test sets (75:25) using a stratified approach to preserve a disease's prevalence within each data set. This was repeated three times with different seeds to assert the robustness of the model's prediction on previously unseen data sets. The training set was then used in a stratified 10-fold nested cross-validation. Due to the class imbalance in the data, the training data set was upsampled within the nested cross-validation (Beinecke and Heider, 2021). We compared multiple methods in our study: RR, RF, and DNN.

2.3.1 Ridge regression

Ridge regression (RR) is a statistical method that includes a penalty parameter, rendering it more stable when input features are correlated compared to other regression models. RR is typically used in calculating PRS. For the RR, we used the scikit-learn library version 0.23.2 (Pedregosa et al., 2011).

2.3.2 Random forests

Random forests (RFs) are proven non-linear classifiers that have been shown to produce good results even in small-*n*-large-*p* scenarios in biomedical classification (Riemenschneider et al., 2016a; Anastasiou et al., 2017). They are based on multiple decision trees that are combined via a majority vote (Breiman, 2001). We used the implementation of the scikit-learn library version 0.23.2 (Pedregosa et al., 2011).

2.3.3 Deep neural networks

Deep neural networks (DNNs) are modeled after biological neurons and consist of multiple layers of artificial neurons. In our study, we used only deep feed-forward networks, where each of these neurons has multiple inputs via weighted connections to previous neurons and calculates an output on the sum of all inputs and with a given activation function. The first layer is called the input layer and is fed with the training features, while the last layer is called the output layer and provides the prediction of the network. These two layers are connected by several so-called hidden layers. All DNNs were implemented using the PyTorch library version 1.7.1 (Paszke et al., 2019).

2.3.4 Hyperparameter optimization

Hyperparameter optimization of all models was carried out within the nested cross-validation. For the DNNs, we evaluated different topologies, ranging from 3 to 6 layers and 2 to 512 neurons per layer. Learning rates of 1×10^{-5}, 1×10^{-4}, and 1×10^{-3} were tested. The loss function used was BCELoss. RFs were optimized with regard to the number of trees (100, 250, 500, and 1,000) and the maximum depth per tree (default, 10, 25, and 50). For RR models, the number of iterations (default, 100, 500, 1,000, and 5,000) was optimized.

After optimizing the hyperparameters in the 10-fold nested cross-validation, models were trained on the full training set using the optimal hyperparameters and then used to predict the test set. Models were evaluated based on the area under the receiver operating characteristic curve (AUC) and accuracy on the test set averaged over three random seeds.

TABLE 2 Comparison of DNN, RF, and RR on the three data sets, BC, CAD, and T2D, for single- and multi-PRS approaches. Evaluation based on AUC and accuracy according to Khera et al. (2018). Values are shown as mean ± SD.

Method	Disease	PRS mode	Accuracy	AUC
DNN	BC	Single-PRS	0.613 ± 0.021	0.653 ± 0.004
DNN	BC	Multi-PRS	0.628 ± 0.024	0.668 ± 0.001
RF	BC	Single-PRS	0.592 ± 0.015	0.626 ± 0.005
RF	BC	Multi-PRS	0.609 ± 0.009	0.648 ± 0.002
RR	BC	Single-PRS	0.598 ± 0.007	0.652 ± 0.004
RR	BC	Multi-PRS	0.612 ± 0.011	0.670 ± 0.002
DNN	CAD	Single-PRS	0.694 ± 0.009	0.785 ± 0.002
DNN	CAD	Multi-PRS	0.698 ± 0.012	0.790 ± 0.002
RF	CAD	Single-PRS	0.674 ± 0.002	0.765 ± 0.003
RF	CAD	Multi-PRS	0.683 ± 0.004	0.768 ± 0.002
RR	CAD	Single-PRS	0.696 ± 0.004	0.785 ± 0.002
RR	CAD	Multi-PRS	0.693 ± 0.004	0.790 ± 0.002
DNN	T2D	Single-PRS	0.626 ± 0.017	0.703 ± 0.002
DNN	T2D	Multi-PRS	0.653 ± 0.010	0.716 ± 0.003
RF	T2D	Single-PRS	0.607 ± 0 014	0.675 ± 0.001
RF	T2D	Multi-PRS	0.610 ± 0.001	0.686 ± 0.002
RR	T2D	Single-PRS	0.636 ± 0.007	0.703 ± 0.002
RR	T2D	Multi-PRS	0.636 ± 0.008	0.716 ± 0.002

3 Results

For the DNNs, no single best topology for all tasks was found (Table 2). The best learning rate for all DNN models was 1×10^{-4}. The best topology for the single-PRS approach for all data sets is 16-8-4-1, while the best topology for the multi-PRS approach is 8-4-4-1 for CAD and T2D and 16-8-4-1 for BC. The rectified linear unit (ReLU) was used as an activation function after all layers, except for the output layer, where the sigmoid function was used. The models performed best after 100 epochs of training. The training of single-PRS models took approximately 8 min, while multi-PRS trainings took approximately 10 min, resulting in a total training time of approximately 80 and 100 min, respectively, for a 10-fold cross-validation. Due to the lower amount of samples for BC, training times were halved for these models.

For the RF models, the best predictions were obtained with 500 trees, while all other parameters were left at the default value. For the RR models, all parameters were left at the default value.

It turned out that the DNNs performed equally well or outperformed RR in all data sets, in particular for the multi-PRS approach. RF did not outperform RR in any data set, neither as single-PRS nor as multi-PRS. In fact, RF performed significantly worse for all data sets and PRS modes with approximately 2% lower AUC and accuracy values than RR and DNNs.

For instance, the DNNs reached an accuracy of 0.653 ± 0.010 compared to 0.636 ± 0.008 for RR for the T2D data set using the multi-PRS approach. For the BC data set, the DNN reached an accuracy of 0.628 ± 0.024 for the multi-PRS approach, while the RR reached only an accuracy of 0.612 ± 0.011. For the single-PRS, the DNN reached an accuracy of 0.613 ± 0.021 and the RR reached an accuracy of 0.598 ± 0.007. For the CAD data set, the DNN reached an accuracy of 0.698 ± 0.012 with the multi-PRS approach, while the RR reached 0.693 ± 0.004. For the single-PRS approach, there were no differences between RR and DNN. Interestingly, using the multi-PRS approach instead of the typically used single-PRS approach generally leads to higher accuracy of the resulting model, irrespective of the underlying prediction model, i.e., RF, RR, or DNN.

4 Discussion

We showed that the inclusion of additional PRS improves the prediction quality of PRS models for predicting an individual's phenotype for BC, CAD, and T2D. The improved prediction quality by including additional PRS can be attributed to the fact that disease susceptibility can be characterized by different risk factors for which at least a partially independent underlying genetic liability exists. For instance, the risk for CAD (coronary artery disease) can be associated with high LDL-cholesterol, high body mass index, smoking, etc., which is also influenced by genetics. Therefore, more comprehensive genetic risk models can be obtained by using a multi-PRS modeling approach. Moreover, by replacing the typically used RR with DNNs, prediction performance could also be improved. DNNs are non-linear classifiers able to capture non-linearity in the underlying data. By not selecting additional PRS manually, we ensured that no information is lost and left it to the algorithms to identify important features. The effect of different PRS on the prediction is likely to be very different. Approaches from explainable AI could be used to identify the relevant PRS.

Although these differences are rather small, the improvement in overall accuracy implies that there are non-linear relationships in the genomics data, as expected from other studies. Improvements in accuracy of up to 1.5%–2% are rather small, but they can have strong implications for patients. For instance, in Europe, there are approximately 355,000 BC cases per year, accounting for more than 90,000 deaths; however, incidences are increasing. Currently, one out of 11 women will develop BC in Europe. In the US, the number is even higher, with approximately 13%, and BC is the second leading cause of death among women. Using prediction models to detect high-risk patients for screening of BC can improve early detection and thus increase life expectancy. An improvement of 1.5% corresponds to more than 5,000 cases that can be detected only in Europe. If we consider T2D, one in 11 adults has diabetes, i.e., 425 million people worldwide. In the United States of America, approximately 11% of people aged between 20 and 79 years have diabetes, while in Europe, it is approximately 6.8%. Approximately 90% of those affected have type 2 diabetes. Every 8 seconds, a person dies as a result of diabetes. It is estimated that almost 700 million people will have diabetes in 2045. Moreover, it has been estimated that a very high number (almost half) of cases are unreported. By improving the risk prediction by 2% solely by incorporating the available data and novel AI models, approximately 7 million more cases could be identified in risk screenings.

From a translational point of view, better prediction performance will improve disease risk stratification. So far, multi-PRS approaches have been rarely applied, mainly due to the limited availability of large-population-based cohorts with deep-phenotyping data to train the model and for the computational issues to deal with high-dimensional data. With the availability of population-based cohorts (such as UK Biobank) and the parallel improvement of computational algorithms for big-data processing, the training of multi-PRS models is feasible on standard HPC infrastructure. Instead, the final application of the models on independent test data is not computationally demanding and therefore can be run locally and potentially integrated into a clinical setting. Additional PRS can be calculated on imputed SNPs based on reference haplotypes if they were not included in the original SNP array.

Our study presents different limitations. In particular, we focused on the genetic predictions of complex traits, including only sex and age as non-genetic factors. However, it is well known that genetic predictors explain only a relatively small proportion of the heritability of complex traits (Gusev et al., 2013). Therefore, in translational settings, different non-genetic risk factors should be included in the prediction models in order to obtain an optimized risk stratification [e.g., the BOADICEA model for breast cancer (Lee et al., 2019)]. Since the multi-PRS model is based on multiple PRS, general limitations of PRS also apply to our model. Some SNPs associated with the diseases may be undiscovered by GWAS, and effect sizes are imprecise (Lewis and Vassos, 2020). Additionally, PRS suffer from a portability problem. PRS calculated on one genetic ancestry perform worse on groups of different ancestry (Martin et al., 2019). In our work, the data set is mainly composed of samples with European genetic backgrounds. Given the different allele frequencies across populations and the limited sample size of non-European individuals, overfitting with respect to the target European population can affect the generalizability of the model. Family-based GWAS are more robust to the effects of population stratification but generally lack power in comparison to non-family-based GWAS (Laird and Lange, 2009). Furthermore, the interpretation of PRS can be difficult and lead to overdiagnosis, resulting in inappropriate treatment (Polygenic Risk Score Task Force of the International Common Disease Alliance et al., 2021).

In the future, we aim to incorporate not only genomics information and PRS but also other clinical data and questionnaires to further improve the risk predictions. As the number of scores in the PGS Catalog constantly grows, those new PRS can be used to update and potentially improve the multi-PRS model. Furthermore, tools other than PLINK (Chang et al., 2015) [e.g., LDpred2 (Privé et al., 2021), PRSice-2 (Choi and O'Reilly, 2019), PRS-CSx (Ruan et al., 2022), or PRSMix (Truong et al., 2023)] can be used to calculate the input PRS.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found at: UK Biobank.

Author contributions

Conceptualization: DH and JK; methodology: JK; software: JK; validation: JK; formal analysis: JK; investigation: JK; resources: DH, CM, HK, PK, AM, AH, and JS; data curation: CM; writing—original draft preparation: JK; writing—review and editing: all authors; visualization: DH and JK; supervision: DH; project administration: DH; funding acquisition: DH, AH, and JS. All authors contributed to the article and approved the submitted version.

Funding

This work was financially supported by the German Federal Ministry of Education and Research (BMBF) [031L0267A] (Deep Insight).

Acknowledgments

Figure 1 was created using BioRender.com.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1217860/full#supplementary-material

References

Abraham, G., Malik, R., Yonova-Doing, E., Salim, A., Wang, T., Danesh, J., et al. (2019). Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat. Commun.* 10, 5819. doi:10.1038/s41467-019-13848-1

Polygenic Risk Score Task Force of the International Common Disease AllianceAdeyemo, A., Balaconis, M. K., Darnes, D. R., Fatumo, S., Granados Moreno, P., et al. (2021). Responsible use of polygenic risk scores in the clinic:

Potential benefits, risks and gaps. *Nat. Med.* 27, 1876–1884. doi:10.1038/s41591-021-01549-6

Anastasiou, O. E., Kälsch, J., Hakmouni, M., Kucukoglu, O., Heider, D., Korth, J., et al. (2017). Low transferrin and high ferritin concentrations are associated with worse outcome in acute liver failure. *Liver Int. Official J. Int. Assoc. Study Liver* 37, 1032–1041. doi:10.1111/liv.13369

Beinecke, J., and Heider, D. (2021). Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making. *BioData Min.* 14, 49. doi:10.1186/s13040-021-00283-6

Bibault, J.-E., Giraud, P., and Burgun, A. (2016). Big data and machine learning in radiation oncology: State of the art and future prospects. *Cancer Lett.* 382, 110–117. doi:10.1016/j.canlet.2016.05.033

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi:10.1038/s41586-018-0579-z

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* 4, 7. doi:10.1186/s13742-015-0047-8

Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. *Nat. Protoc.* 15, 2759–2772. doi:10.1038/s41596-020-0353-1

Choi, S. W., and O'Reilly, P. F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* 8, 082. giz082. doi:10.1093/gigascience/giz082

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., et al. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* 24, 1559–1567. doi:10.1038/s41591-018-0177-5

Gusev, A., Bhatia, G., Zaitlen, N., Vilhjalmsson, B. J., Diogo, D., Stahl, E. A., et al. (2013). Quantifying missing heritability at known GWAS loci. *PLoS Genet.* 9, 1003993. doi:10.1371/journal.pgen.1003993

Kahn, R., Robertson, R. M., Smith, R., and Eddy, D. (2008). The impact of prevention on reducing the burden of cardiovascular disease. *Diabetes Care* 31, 1686–1696. doi:10.2337/dc08-9022

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50, 1219–1224. doi:10.1038/s41588-018-0183-z

Krapohl, E., Patel, H., Newhouse, S., Curtis, C. J., von Stumm, S., Dale, P. S., et al. (2017). Multi-polygenic score approach to trait prediction. *Mol. Psychiatry* 23, 1368–1374. doi:10.1038/mp.2017.163

Laird, N. M., and Lange, C. (2009). The role of family-based designs in genome-wide association studies. *Stat. Sci.* 24. doi:10.1214/08-STS280

Lambert, S. A., Abraham, G., and Inouye, M. (2019). Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* 28, R133–R142. doi:10.1093/hmg/ddz187

Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., et al. (2019). Boadicea: A comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet. Med.* 21, 1708–1718. doi:10.1038/s41436-018-0406-9

Lewis, C. M., and Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. *Genome Med.* 12, 44. doi:10.1186/s13073-020-00742-5

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332. doi:10.1038/nrg3920

MacEachern, S. J., and Forkert, N. D. (2021). Machine learning for precision medicine. *Genome* 64, 416–425. doi:10.1139/gen-2020-0131

Madabhushi, A., and Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med. Image Anal.* 33, 170–175. doi:10.1016/j.media.2016.06.037

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51, 584–591. doi:10.1038/s41588-019-0379-x

Meisner, A., Kundu, P., Zhang, Y. D., Lan, L. V., Kim, S., Ghandwani, D., et al. (2020). Combined utility of 25 disease and risk factor polygenic risk scores for stratifying risk of all-cause mortality. *Am. J. Hum. Genet.* 107, 418–431. doi:10.1016/j.ajhg.2020.07.002

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems 32*. Editors H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Red Hook, New York: Curran Associates, Inc), 8024–8035.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Privé, F., Arbel, J., and Vilhjálmsson, B. J. (2021). LDpred2: Better, faster, stronger. *Bioinformatics* 36, 5424–5431. doi:10.1093/bioinformatics/btaa1029

Ren, Y., Chakraborty, T., Doijad, S., Falgenhauer, L., Falgenhauer, J., Goesmann, A., et al. (2021). Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics* 38, 325–334. doi:10.1093/bioinformatics/btab681

Riemenschneider, M., Cashin, K. Y., Budeus, B., Sierra, S., Shirvani-Dastgerdi, E., Bayanolhagh, S., et al. (2016a). Genotypic prediction of co-receptor tropism of HIV-1 subtypes a and c. *Sci. Rep.* 6, 24883. doi:10.1038/srep24883

Riemenschneider, M., Hummel, T., and Heider, D. (2016b). Shiva - a web application for drug resistance and tropism testing in HIV. *BMC Bioinforma.* 17, 314. doi:10.1186/s12859-016-1179-2

Rodriguez, V., Alameda, L., Quattrone, D., Tripoli, G., Gayer-Anderson, C., Spinazzola, E., et al. (2022). Use of multiple polygenic risk scores for distinguishing schizophrenia-spectrum disorder and affective psychosis categories in a first-episode sample; the eu-gei study. *Psychol. Med.* 1, 1–10. doi:10.1017/S0033291721005456

Ruan, Y., Lin, Y.-F., Feng, Y.-C. A., Chen, C.-Y., Lam, M., Guo, Z., et al. (2022). Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* 54, 573–580. doi:10.1038/s41588-022-01054-7

Schoeler, T., Choi, S. W., Dudbridge, F., Baldwin, J., Duncan, L., Cecil, C. M., et al. (2019). Multi–polygenic score approach to identifying individual vulnerabilities associated with the risk of exposure to bullying. *JAMA Psychiatry* 76, 730–738. doi:10.1001/jamapsychiatry.2019.0310

Schröder, J., Chegwidden, L., Maj, C., Gehlen, J., Speller, J., Böhmer, A. C., et al. (2022). GWAS meta-analysis of 16 790 patients with barrett's oesophagus and oesophageal adenocarcinoma identifies 16 novel genetic risk loci and provides insights into disease aetiology beyond the single marker level. Gut , *gutjnl*– 72, 612–623. 326698. doi:10.1136/gutjnl-2021-326698

Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., et al. (2021). Genetics of 35 blood and urine biomarkers in the UK biobank. *Nat. Genet.* 53, 185–194. doi:10.1038/s41588-020-00757-z

Spänig, S., Emberger-Klein, A., Sowa, J.-P., Canbay, A., Menrad, K., and Heider, D. (2019). The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes. *Artif. Intell. Med.* 100, 101706. doi:10.1016/j.artmed.2019.101706

Truong, B., Hull, L. E., Ruan, Y., Huang, Q. Q., Hornsby, W., Martin, H., et al. (2023). Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. *Prepr. Genet. Genomic Med.*, 23286110. doi:10.1101/2023.02.21.23286110

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 13 July 2023
DOI 10.3389/fgene.2023.1232038

OPEN ACCESS

EDITED BY
Margherita Mutarelli,
National Research Council (CNR), Italy

REVIEWED BY
Hilal Tayara,
Jeonbuk National University, Republic of Korea
Kunqi Chen,
Fujian Medical University, China

*CORRESPONDENCE
Jianhua Jia,
jjh163yx@163.com
Zhangying Wei,
weizy5003@163.com

RECEIVED 31 May 2023
ACCEPTED 29 June 2023
PUBLISHED 13 July 2023

CITATION
Jia J, Wei Z and Cao X (2023), EMDL-ac4C: identifying N4-acetylcytidine based on ensemble two-branch residual connection DenseNet and attention. *Front. Genet.* 14:1232038. doi: 10.3389/fgene.2023.1232038

EMDL-ac4C: identifying N4-acetylcytidine based on ensemble two-branch residual connection DenseNet and attention

Jianhua Jia*, Zhangying Wei* and Xiaojing Cao

School of Information Engineering, Jingdezhen Ceramic University, Jingdezhen, China

Introduction: N4-acetylcytidine (ac4C) is a critical acetylation modification that has an essential function in protein translation and is associated with a number of human diseases.

Methods: The process of identifying ac4C sites by biological experiments is too cumbersome and costly. And the performance of several existing computational models needs to be improved. Therefore, we propose a new deep learning tool EMDL-ac4C to predict ac4C sites, which uses a simple one-hot encoding for a unbalanced dataset using a downsampled ensemble deep learning network to extract important features to identify ac4C sites. The base learner of this ensemble model consists of a modified DenseNet and Squeeze-and-Excitation Networks. In addition, we innovatively add a convolutional residual structure in parallel with the dense block to achieve the effect of two-layer feature extraction.

Results: The average accuracy (Acc), mathews correlation coefficient (MCC), and area under the curve Area under curve of EMDL-ac4C on ten independent testing sets are 80.84%, 61.77%, and 87.94%, respectively.

Discussion: Multiple experimental comparisons indicate that EMDL-ac4C outperforms existing predictors and it greatly improved the predictive performance of the ac4C sites. At the same time, EMDL-ac4C could provide a valuable reference for the next part of the study. The source code and experimental data are available at: https://github.com/13133989982/EMDLac4C.

KEYWORDS

ac4C site identification, ensemble deep learning, DenseNet, attention mechanism, residual structure

1 Introduction

RNAs from both eukaryotic and prokaryotic cells may include a broad range of nucleoside modifications (Tardu et al., 2019), and it is statistically known that there are more than 170 types (Boccaletto et al., 2018). Song et al. (2023) have developed the RMDisease database and identified a large number of disease-associated variants to elucidate the important regulatory role of RNA modifications. Among them, N4-acetylcytidine (ac4C) is a highly conserved RNA modification, and at the same time, he is the sole acetylation modification of eukaryotic RNA that has been identified (Zhao et al., 2019a; Jin et al., 2020). Ac4C plays an important role in biology, and it has different functions on different RNAs. On tRNA, ac4C helps to improve the accuracy of protein

translation and maintain the heat resistance of the organism (Kumbhar et al., 2013); the role of ac4C on rRNA likewise includes maintaining high fidelity of protein translation (Sharma et al., 2015), while it is also a marker of thermophilic organisms, which is significant; on mRNA, ac4C is required to safeguard the stability of mRNA while increasing the efficiency of protein translation (Arango et al., 2018; Dominissini and Rechavi, 2018). Meanwhile, Chen et al. (2022) demonstrated that the only known ac4C writer, N-acetyltransferase 10 (NAT10), has an important effect in male reproduction. In addition, ac4C has regulatory effects on viruses (Tsai et al., 2020; Hao et al., 2022) and has been associated with several human diseases, including: osteoporosis (Yang et al., 2021), pancreatic cancer (Feng et al., 2022), *etc.*

The recognition of ac4C sites has gradually become a popular topic in biology and computer research. In the context of biological experiments, multiple testing methods exist for ac4C studies. Previously, partially enzymatic hydrolysis and two-dimensional paper chromatography were commonly used to identify ac4C modifications in RNA (Thomas et al., 2019). In the past few years, researchers had found that the combination of LC-MS and HPLC-MS analyses is more efficient in isolating partially modified nucleic acids, including ac4C (Ito et al., 2014; Sharma et al., 2015; Sharma et al., 2017). In addition, RPHPLC (Mezzar et al., 2014) is widely used by the biological community as it requires only a small number of samples and does not rely on expensive mass spectrometry assays or the use of radioactive substrates. There are also specific ac4C sequencing approaches, including ac4C-seq (Gamage et al., 2021) and RedaC:T-seq (Sturgill et al., 2022), both of which sequence ac4C through a series of experiments under certain conditions. Nevertheless, these experimental methods always have several problems, such as time-consuming and high expensive.

Several machine learning-based predictive models (Zhou et al., 2016; Wei et al., 2019; Basith et al., 2019; Lv et al., 2020; Hasan et al., 2021) for the predictive identification of RNA post-translational modification sites have been developed by researchers in the past several years. Among them, there are two predictors used to identify ac4C sites, firstly, Zhao et al. (2019a) developed PACES based on position-specific dinucleotide sequences as well as K-nucleotide frequency coding, which was trained using two random forest classifiers. Secondly, Alam et al. (2020b) proposed XG-ac4C predictor based on this, which used multiple encoding methods, including one-hot, nucleotide chemistry and density, Kmer, *etc.*, and used extreme gradient boosting (XGboost) to train the dataset to identify ac4C loci. Nonetheless, neither of these two predictors' ability in making predictions is sufficient.

With the widely use of deep learning, researchers have introduced different deep learning models to the field of RNA post-transcriptional modification site prediction (Yu and Dai, 2019; Liu et al., 2020; Hasan et al., 2021; Rehman et al., 2021; Hasan et al., 2022; Tsukiyama et al., 2022; Yang et al., 2023), and a lot of experimental results show that deep learning models perform better than machine learning models for datasets with a large number of samples. Muhammad et al. (Iqbal et al., 2022) used a deep learning network-based CNN model using an encoding approach similar to the XG-ac4C predictor DL-ac4C predictor was proposed to identify ac4C sites, and compared with machine learning methods, the results showed that DL-ac4C has better prediction performance. Recently, a new deep learning model DeepAc4C (Wang et al., 2021a) had also been developed to increase the efficiency of ac4C locus recognition in mRNA. It also used CNN to extract information and classify the feature maps, and encoded a combination of physicochemical features and nucleotide semantic information. Yet, the classification performance of these two predictors still needs to be improved. As a result, a more analytically precise model to anticipate ac4C sites is urgently required.

In order to more accurately predict ac4C sites in mRNAs, an effective prediction model EMDL-ac4C was developed in this paper, and the contributions of this paper were various: 1) EMDL-ac4C used only the simplest encoding method one-hot to represent nucleotides. This encoding showed the distribution probability of each nucleotide, making it easier to calculate the distance between nucleotides. 2) It proposed a powerful deep learning that used DenseNet in combination with convolutional residuals to form two-branch residual connection DenseNet, and the effectiveness of feature extraction was increased by this way. 3) For cases like this paper, where unbalanced datasets were processed into multiple balanced datasets, downsampling integration was used to achieve significantly superior generalization performance than a single learner. 4) The attention mechanism was carried out throughout the network structure to give greater attention to the important information at each stage. 5) We compared the performance of various encoding techniques, different numbers of dense blocks, multiple model architectures, and several predictors, respectively, to confirm the efficacy of the EMDL-ac4C model.

Therefore, the final predictor is called "EMDL-ac4C," in which "EM" stands for "ensemble," "DL" represents "deep learning," and "ac4C" means "N4-acetylcytidine."

TABLE 1 Distribution of data set D1.

Dataset	Positive	Negative
Training	1,148	1,148
Testing	467	467

2 Materials and methods

2.1 Benchmark dataset

The base dataset for this study was extracted from 2,134 genes offered by Zhao et al. (2019a) from a highly throughput dataset previously presented. In the training set, there were a total of 1,160 positive and 10,855 negative samples, and in the test set, the counts of positive and negative samples were: 469, 4,343, respectively. Alam et al. (2020b) also built the XG-ac4C predictor from this dataset. Wang et al. (2021b) performed a de-sampling redundancy with a threshold of 0.4 using CD-HIT (Weizhong et al., 2006) software in order to remove redundant sequences from these datasets, resulting in 1,615 positive and 7,590 negative samples. These samples were separated into a set for training and one for testing, with 1,148 positive samples and 5,439 negative samples in the former. On the other hand, there were 467 positive samples and 2,151 negative samples in the independent test set. Finally,

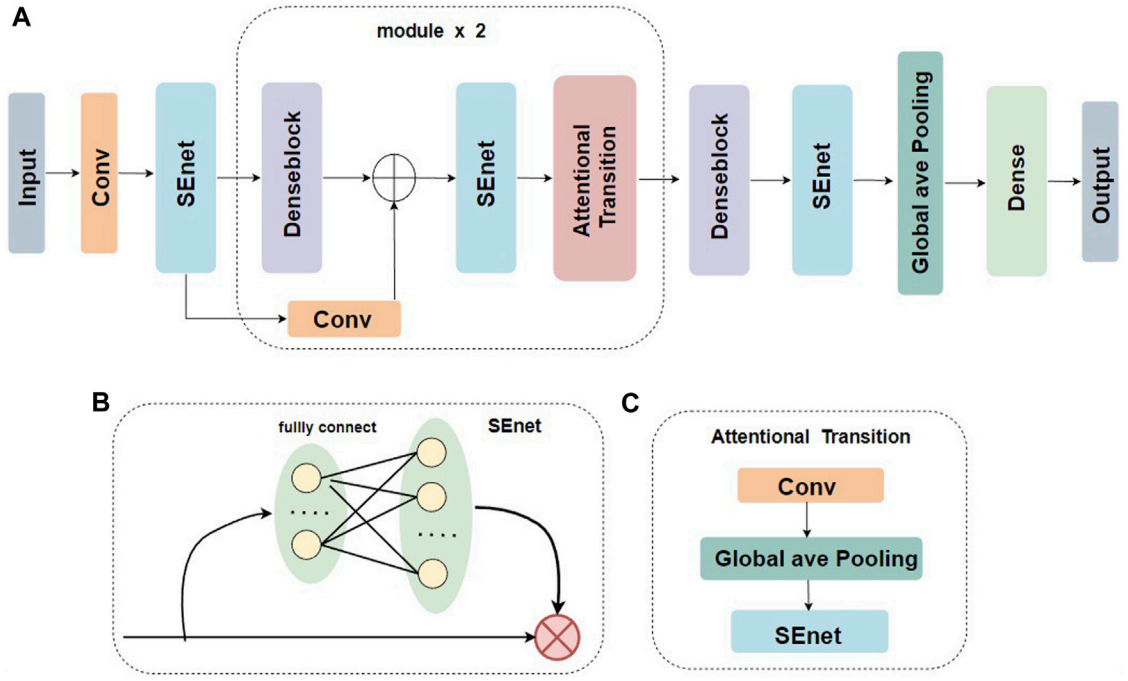


FIGURE 1
Base classifier for EMDL-ac4C. **(A)** Schematic graph of the two-branch residual connection DenseNet model. **(B)** Schematic diagram of the SEnet module. **(C)** Schematic diagram of the Attentional Transition module.

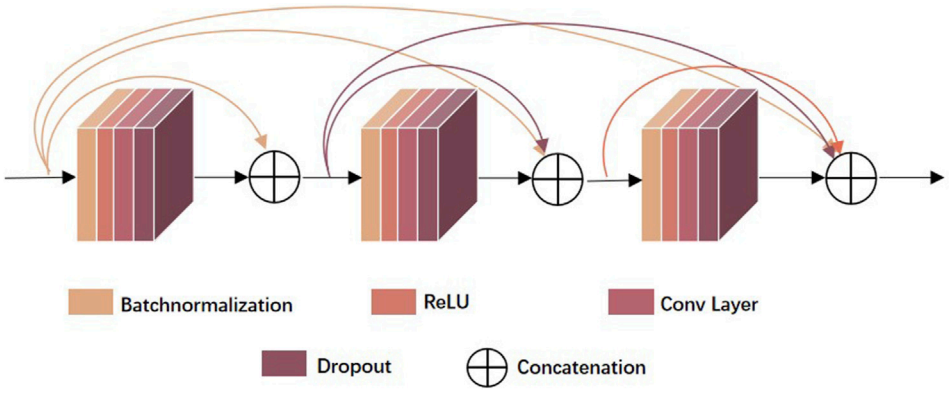


FIGURE 2
The structure of denseblock.

they created ten balanced datasets using the unbalanced training and test sets, respectively, to make it easier to train and test the model. The number of positive samples in each sub-training set and test set remains the same, and negative samples are randomly selected from the corresponding negative dataset, and the number is consistent with the positive sample size. The ten samples are denoted as D1, D2, , D10. Table 1 shows an example of the distribution of data set D1: (the distribution of D2, D3, , D10 is the same).

As in Eq. 1, nucleotide sequences containing potential N4-acetylcytidine sites can normally be read as:

$$f_{\delta}(K) = R_{-\delta}R_{-(\delta-1)}\cdots R_{-2}R_{-1}KR_{+1}R_{+2}\cdots R_{+(\delta-1)}R_{+\delta} \quad (1)$$

where the center K denotes "N4-acetylcytidine," $R_{-\delta}$ means the δth upstream nucleotide from the center K, while $R_{+\delta}$ stands for the δth downstream nucleotide from the center K. In this study, δ is 207, that is, the length of a nucleotide sequence is $(2\delta + 1)$.

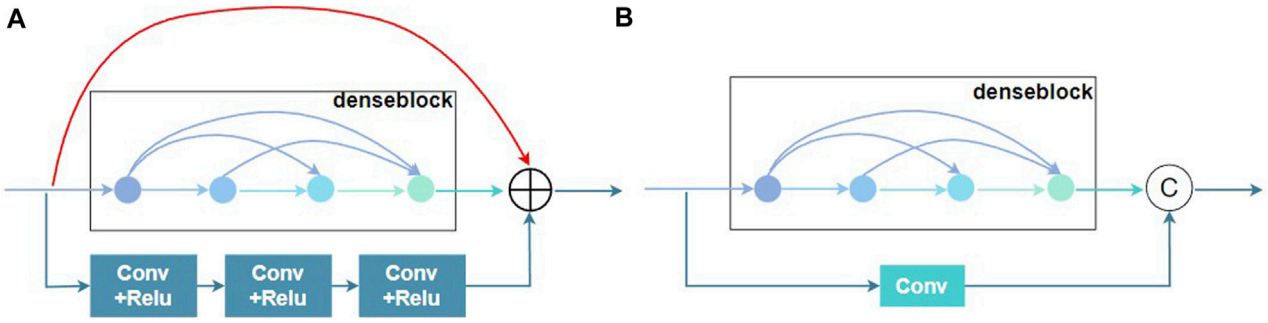


FIGURE 3
Residual connection schematic of the denseblock module. **(A)** RXDNFuse's multi-branch Residual denseblock. **(B)** Two-branch Residual denseblock.

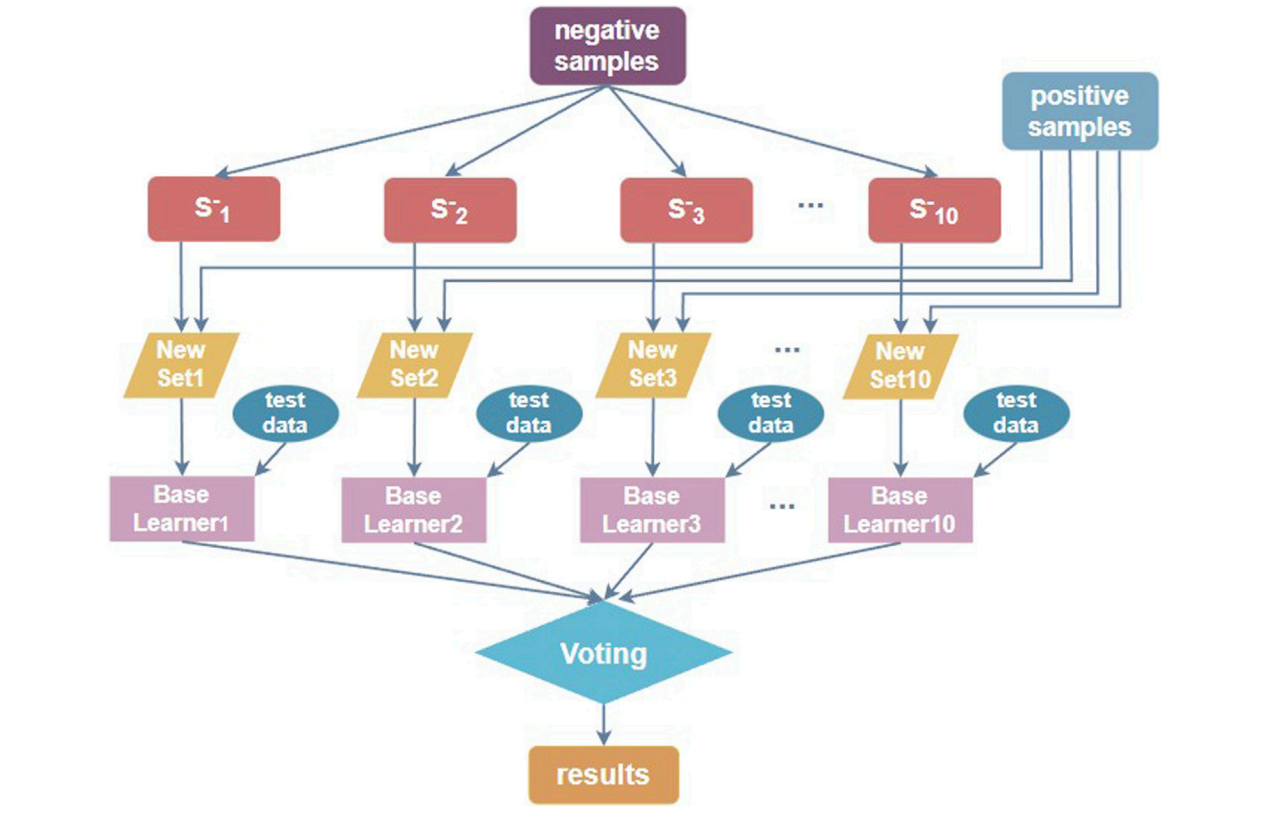


FIGURE 4
Downsampling ensemble classifier construction process.

2.2 Feature coding methods

2.2.1 One-hot coding

The RNA sequence used in this study consists of four nucleotides and a "-," where the "-" represents a missing value or an undetected nucleotide in the RNA sequence. One-hot encoding, also known as binary encoding, converts each nucleotide into a numeric vector of 0 and 1, encoding "A" as [1, 0, 0, 0, 0], "C" as [0, 1, 0, 0, 0], "G" as [0, 0, 1, 0, 0], "T" as [0, 0, 0, 1, 0], and "-" is assigned as [0, 0, 0, 0, 0, 1]. One-hot encoding not only simply converts sequence information into digital information for computer processing, but more importantly, it makes the calculation of distances between nucleotides more reasonable. For example, if nucleotides are represented in sequential encoding: 1:A, 2:C, 3:G, 4:T, then the distance between A (adenine ribonucleotide) and C (cytosine ribonucleotide) is smaller than the distance between A (adenine ribonucleotide) and G (guanine ribonucleotide), which is not reasonable. At the same time, one-hot coding in fact means the

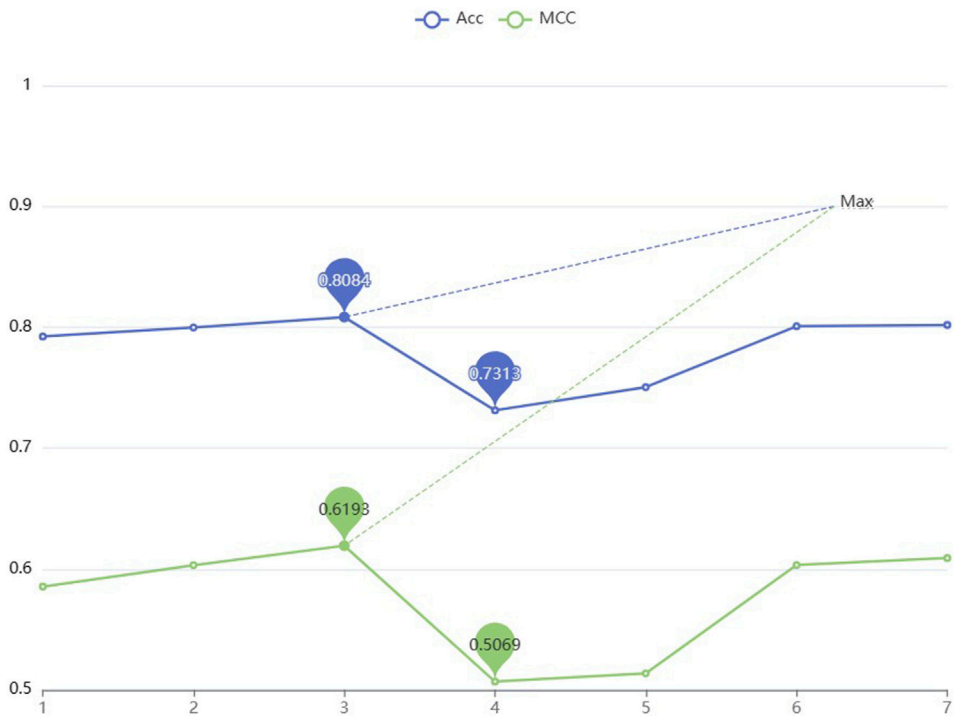


FIGURE 5
The comparison of Acc and MCC values for models with various amounts of denseblocks.

probability distribution of nucleotides, that is, one nucleotide has probability 1 and the others are all 0.

2.3 Classification model

2.3.1 Base learning model

Figure 1 illustrates the model structure of the base learner two-branch residual connection DenseNet for constructing EMDL-ac4C. A): The basic structure of two-branch residual connection DenseNet consists of SENet, denseblock, attentional-transition, residual convolution layer and fully connected layer. B): SENet is composed of two fully connected layers with different number of neurons to fulfill the aim of first dimensionality reduction and then dimensionality increase of the feature map. C): The attentional-transition is composed of 1 × 1 convolution, global average pooling layer and SENet, where the role of convolution and global average pooling is to condense the output feature map of two-branch denseblock and convolutional residual module, decrease the size and dimension of the feature map, and simultaneously can successfully decrease the quantity of denseblock parameters and stop the network to excessive fitting. The purpose of SENet is to extract important features. The final fully connected layer is used as the classification prediction of the model.

Below are explanations of each step in greater detail.

2.3.1.1 DenseNet

Huang et al. (2017) proposed DenseNet in 2017 for the target recognition task to alleviate the gradient disappearance issue that often occurs in deep networks, while feature reuse also enhanced feature propagation with fewer parameters in a network of equal layer depth. As shown in Eq. 2, ResNet only adds features to the input of the latter layer and connects them in a summation manner. DenseNet is an improvement of ResNet in that it combines the features of each layer by concatenation. As shown in Eq. 3, DenseNet connects all the previous layers as the input of the next layer, obtaining better performance than ResNet with fewer parameters and computational cost.

$$x_l = H_l(x_{l-1}) + x_{l-1} \tag{2}$$

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{3}$$

The $H_l(\bullet)$ in Eqs 2, 3 represents the non-linear transformation function, which is a combined operation that may include a series of BN (Batch Normalization), ReLU(Rectified Linear Unit), POOL (Pooling) and Conv (Convolution) operations. The non-linear transformation function in this paper adopted the structure of BN+ReLU+1*1 Conv +3*3 Conv, which were created through a preactivation strategy to facilitate network training as well as enhanced the efficiency of generalization, and 1 × 1 Conv served to reduce the number of features, thus reduced computational workload and improved computational efficiency. In addition, 3 × 3 Conv offered a larger receptive field.

As shown in Figure 2, the denseblock is a module containing many layers, each layer has the same feature map size, and the layers are closely connected to each other, while the Transition module connects two neighboring denseblocks and reduces the feature map size by Pooling. In this paper, we used a new approach Attentional

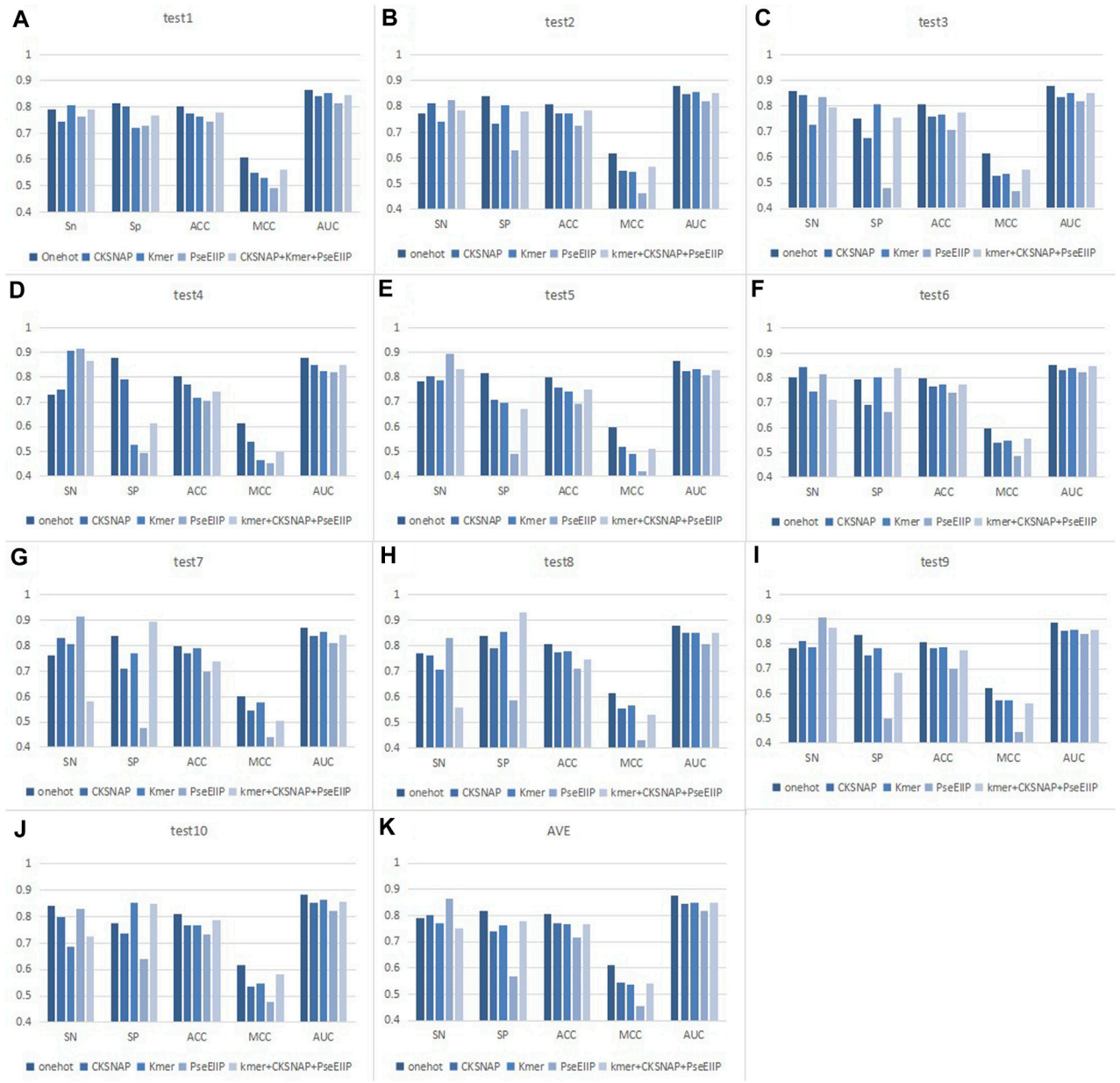


FIGURE 6
Performance comparison of different encoding methods on ten test sets. **(A–J)** Results for different test sets. **(K)** Average results of 10 test sets.

Transition instead of transition, whose construct is depicted in Figure 1C.

2.3.1.2 Residual connection

In order to effectively spread shallow information to deep layers, the output of residual connection is represented by a linear superposition of the inputs and their nonlinear transformations. Deep learning propagates the gradient (derivative) of the loss function step by step from the back to the front in the back propagation process, and the gradient is less than 1 at each level. Due to the cumulative multiplicative effect, the gradient may be too small and cause the network to stop training optimization. Therefore, the addition of residual connections to the network can address the issue of network degradation and enhance network functionality.

Long et al. (2021) proposed the aggregated residual dense network (RXDNFuse) for the mix problem of IR and visible images, which combined the residual and convolutional residuals into parallel dense blocks to extract multi-level features, as depicted in Figure 3A. The results also demonstrated that RXDNFuse can effectively retain the important thermal radiation information in the feature map, from which this work was inspired to propose Two-branch residual connection dense network, which consisted of a denseblock and a convolutional residual connection to form two channels. As illustrated in Figure 3B, compared with RXDNFuse, two-

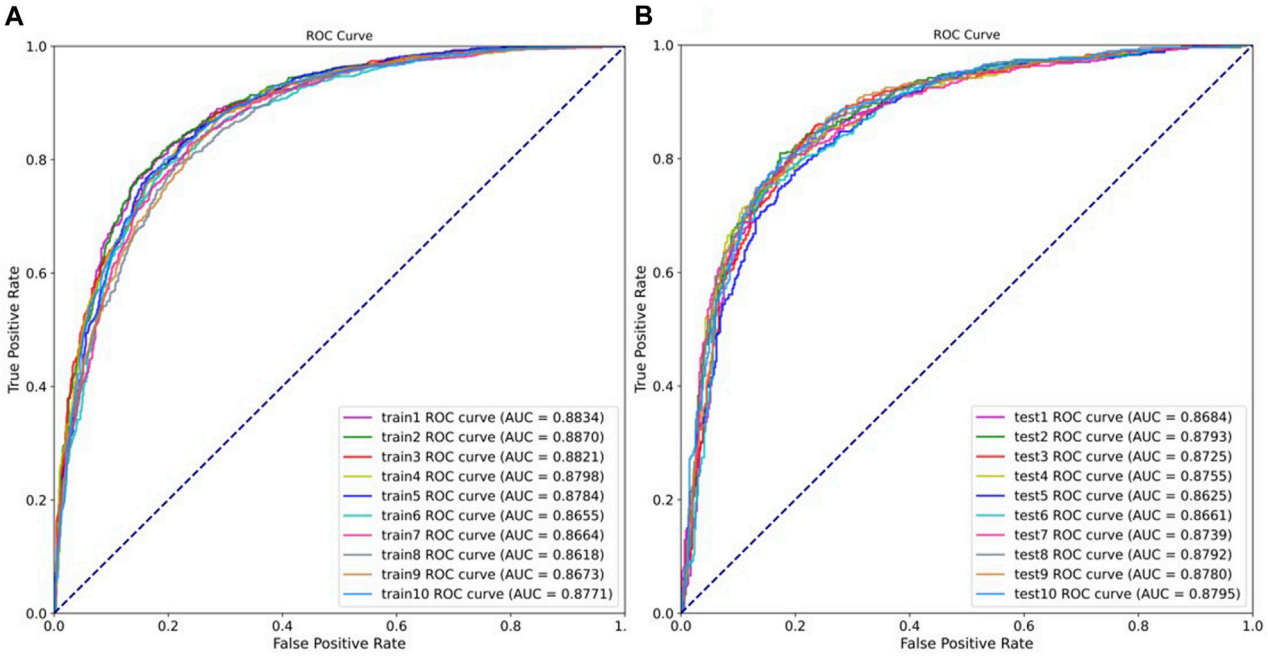


FIGURE 7
ROC curve of two-branch residual connection DenseNet on each dataset. **(A)** Results on training sets. **(B)** Results on testing sets.

branch residual connection DenseNet has three differences, firstly, it reduces one layer of residual connection, secondly, it changes convolutional residual connection part from 3 layers of Conv+ReLU to one layer of Conv to extract features, and for the last, it changes the residual connection method from summation to concatenation. In this way, we can improve the diversity of feature extraction and achieve the effect of two-branch feature extraction, while reducing the complexity of the model.

2.3.1.3 SENet

Introducing attention mechanism in the predictor can make model more efficient in learning the interrelationships between feature information (Vaswani et al., 2017), while focusing on useful information. (Wang et al., 2020; Wang et al., 2021; Jia et al., 2022; Jia et al., 2023). The use of the attention mechanism in both image processing and natural language learning has demonstrated its value in enhancing the model's capacity for recognition, and our study again confirms this point that the attention mechanism can help suppress useless information, pay attention to critical information, and improve the model performance.

Researchers have developed various attention mechanisms, commonly used including self-attention mechanisms, spatial attention mechanism, channel attention mechanism, and so on. Squeeze-and-Excitation Networks (SENet) (Hu et al., 2018) is one of the channel attention mechanisms. We investigated the use of attention mechanisms in our model in two ways, one was to choose the attention mechanism that best matches the model and the data characteristics, and try to use the four attention mechanisms individually or in combination. The second was where to place the attention mechanism in the model. There are various options for where to insert attention into the model, including introducing it in densecells (Bastings and Filippova, 2020), adding it in denseblocks (Wei et al., 2019; Zhou et al., 2019; Wang et al., 2021), inserting it between denseblocks and transitions (Jia et al., 2022; Jia et al., 2023), bring in the attention mechanism in the transition layer (Song et al., 2021), or attaching it before the data enters DenseNet or at the end of the model prediction.

The combined use of SENet and DenseNet has been repeatedly shown to boost network detection and site prediction performance (Yan et al., 2019; Wang et al., 2021; Shi et al., 2021; Jia et al., 2023). We had also found after numerous ablation experiments that SENet alone works best, while it was advisable to place SENet before DenseNet, between denseblock and Attentional Transition, and before the final global average pooling, as seen in Figure 1A. Adding the attentional layer before the initial feature map enters the DenseNet helps the model not miss important information in the original feature map, while the attentional layer behind denseblock aims to repress redundant features and strengthen propagation of important features.

The construction of SENet is to first perform a global pooling operation on the feature map with input h*w*c, which is a spatial compression process that makes the feature map 1*1*c in size. Next there are two fully connected layers. The first full connection has c/16 neurons, which is a dimensionality reduction procedure, and the second fully connected is ascending to c neurons. The significance of dimensionality reduction and then dimensionality increase is to discover the correlation between channels. The final step is to multiply the original h*w*c feature map with the 1*1*c feature map after dimensionality down and dimensionality up to obtain a feature map with the importance levels of different channels.

It is worth noting that the SENet used in this paper removes the global pooling operation at the beginning, the reason is that the global pooling operation will lose some location information.

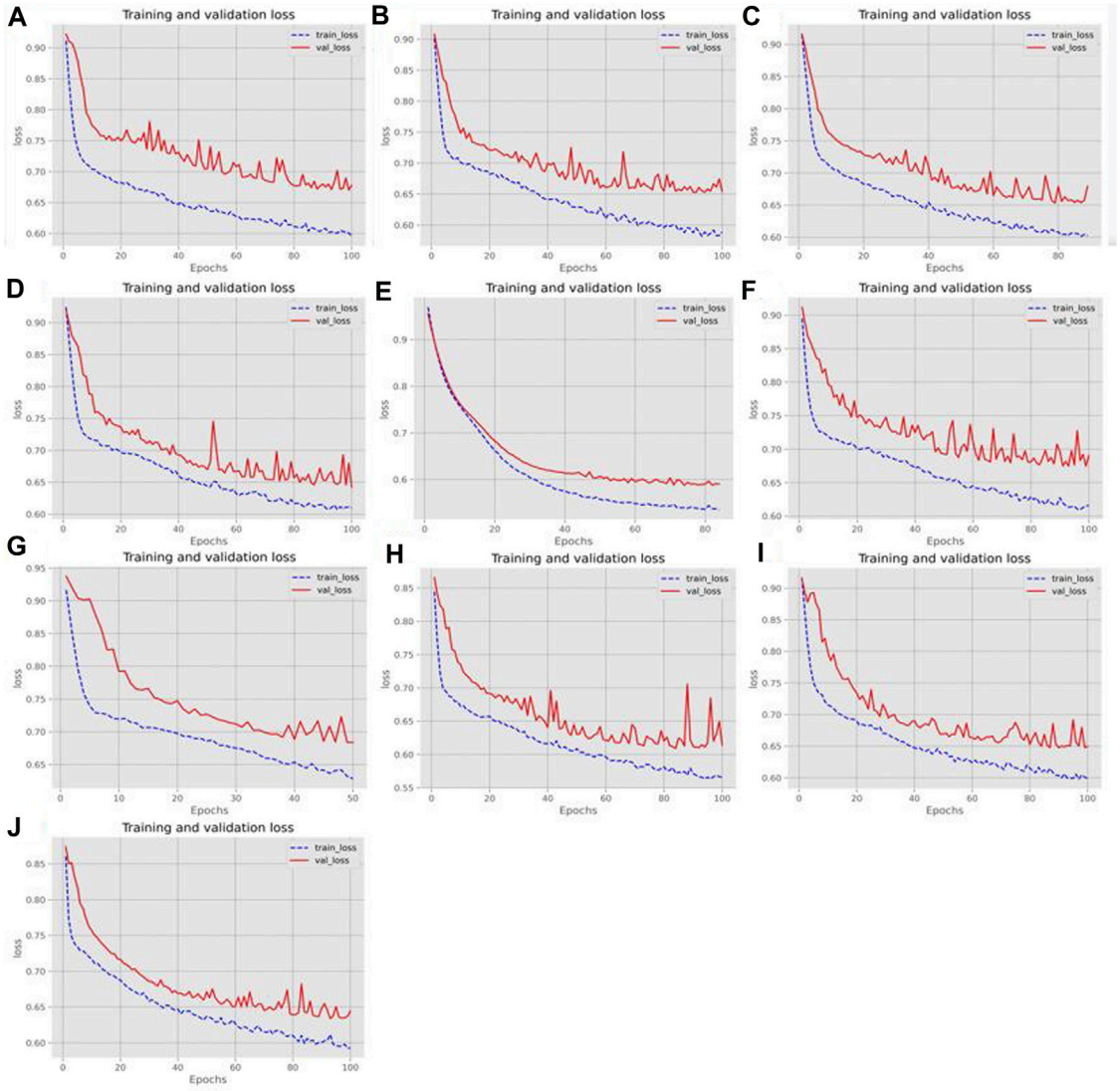


FIGURE 8
The loss change of the training and validation sets. **(A–J)** Loss changes in ten different training and validation sets.

2.3.1.4 Attentional transition

For the transition layer, it mostly connects two neighboring denseblocks and decreases the size of the feature map. Yang et al. (Yang et al., 2018) proposed CliqueNet to alleviate the training challenge of deep networks, which introduced a channel-based attentional mechanism in the transition layer. Following their approach, we also introduce the attentional mechanism in transition layers to ensure high-quality propagation of features between dense blocks. As Figure 1C shows, after convolution, the feature map is then brought into SENet, which means that the feature map is first global average pooled, followed by two fully connected layers (FC) to complete the descending and ascending operations.

2.3.2 Ensemble learning

When the number of positive samples in the benchmark dataset is significantly lower than the number of negative sample data, the dataset is unbalanced. For such a non-equilibrium dataset, using a simple model is not friendly to identify positive samples, while for us, the information of N4-acetylcytidine loci is the most critical and the most necessary to be identified. The ensemble learning method can be used to downsample the non-balanced data, meaning that for the majority of negative samples, downsampling is performed each time, and the same number of subsets as positive samples are extracted, and the two constitute a balanced dataset for training. Multiple sub-classifiers are thus constructed, and then the training models are validated by cross-validation and the model training effect is verified by independent test sets, respectively. Such use of the ensemble classifier dramatically improves the accuracy of loci prediction (Jia et al., 2016). The process of constructing the downsampling ensemble classifier is shown in Figure 4.

Wang et al. (Wang et al., 2021) constructed the dataset in a similar manner to Figure 4, with ten random downsampling of negative samples to build ten balanced datasets. Therefore, we used 10 balanced training sets to build 10 sub-classifiers, trained the model by cross-validation, and then validated the model training effect using ten independent test sets

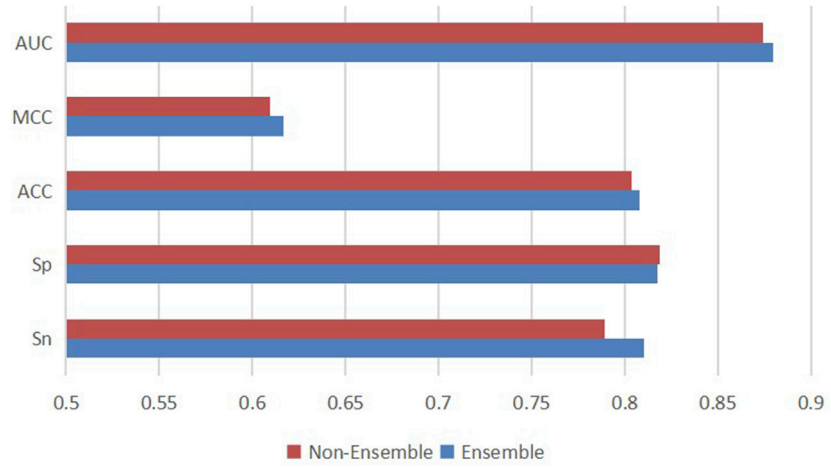


FIGURE 9
Performance comparison of single two-branch residual connection DenseNet and downsampling ensemble model EMDL-ac4C.

one by one. The prediction results obtained from multiple sub-classifiers were soft-voted to obtain the final ensemble results. The base sub-classifiers are implemented with the two-branch residual connection DenseNet shown in Figure 1A.

2.3.3 Performance evaluation

Five metrics are typically used to evaluate models in such studies: Accuracy (Acc), Sensitivity (Sn), Specificity (Sp), Area under curve (AUC) and Matthew's correlation coefficient (MCC), which are computed as follows in Eq. 4:

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FN) \times (TP + FP) \times (TN + FP)}} \quad (4)$$

TP is a correctly identified positive ac4C site, FN is a misidentified positive ac4C site, and TN and FP are correctly and incorrectly predicted negative ac4C sites, respectively. Researchers often use ROC curves to indicate the performance of classifiers, and the area under the ROC curve is measured as the AUC value, and a bigger value indicates greater performance.

In this study, we used these five prevalent evaluation metrics to assess the performance of EMDL-ac4C.

3 Results and discussion

3.1 Comparison of models with different denseblocks

To improve the performance of the predictor, the parameters are optimized in two-branch residual connection DenseNet. In this section, different numbers of denseblocks are set and the Acc and MCC values are used to compare the model performance for different numbers of denseblocks. According to Figure 5, selecting a model with three denseblocks yields the highest performance. The trough is reached when the number is 4. After the number is 5, the Acc and MCC values increase as the amount of blocks grows, but considering that when the amount of denseblocks reaches 8, the maximum feature map scale in the model run has reached [211464,768], which is easy to cause insufficient server memory. Therefore, based on the consideration of computational complexity, the denseblock = 3 is selected.

3.2 Comparison of models with various encoding methods

For the choice of feature coding methods, we considered four traditional coding methods, namely,: One-hot, composition of k-spaced nucleic acid pairs (CKSNAP), Kmer, and electron–ion interaction pseudopotentials of trinucleotide (PseEIIP). Also included: this coding scheme of CKSNAP + Kmer + PseEIIP combination. These coding methods have been applied in many studies (Li et al., 2020; Wang et al., 2021; Chen et al., 2021; El Allali et al., 2021; Le et al., 2022; Wang et al., 2023) and are not described in detail here. In addition to these traditional coding approaches, there are some new coding approaches including: Gene2vec (Zou et al., 2019), Geo2vec (Huang et al., 2022), Genomics features (Chen et al., 2019), Chemical property (Chen et al., 2017), Heuristic nucleotide physical-chemical properties reduction (Dou et al., 2020) also gave us a lot of inspiration on sequence encoding. We compared these coding schemes on ten balanced test datasets, as indicated in Figure 6. In the ten experiments, we tested univariate the encoding style suitable for the model using a unified classifier: two-branch residual connection DenseNet.

As seen in Figure 6, we can find that for each test set, the values of the indicators corresponding to the same coding method do not differ much. For example, the AUC value of the model with one-hot

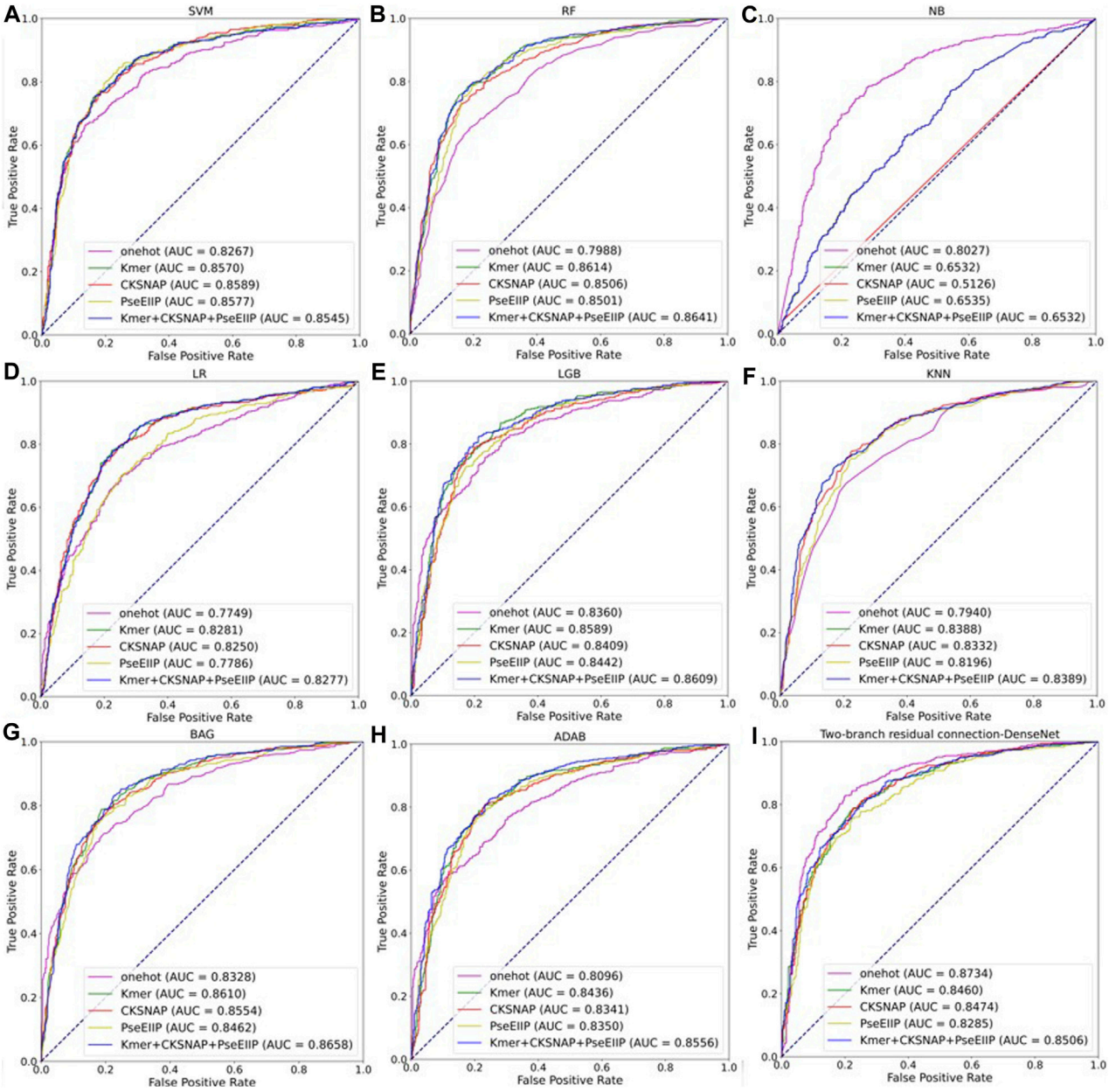


FIGURE 10
Multiple model ROC curves based on test sets with different encoding methods. **(A–H)** Results of machine learning. **(I)** Results of two-branch residual connection DenseNet.

encoding in each dataset is about 0.87. The detailed performance of the different encoding methods of EMDL-ac4C on ten independent test sets is shown in Supplementary Tables S1–S10. This indicates that two-branch residual connection DenseNet has good generalizability and does not show excessive differences depending on the dataset. In addition, it is also obvious from Figure 6 that the model with one-hot coding has significantly higher Acc, MCC and AUC values than the other coding methods. After comprehensive consideration, we concluded that one-hot encoding is more suitable for two-branch residual connection DenseNet model for predicting ac4C sites.

3.3 Comparison of ensemble and non-ensemble models

To illustrate the effectiveness of the two-branch residual connection DenseNet more intuitively, we further tested its ROC curves on ten training and test sets, as shown in Figure 7. Also, Figure 8 showed the loss changes on the training and validation sets. The model demonstrated excellent performance and balanced results on ten training sets and ten testing sets. The difference in indicators on each dataset done not exceed 2.52%, indicating that the two branch residual connection DenseNet can stably predict the ac4C sites.

TABLE 2 Average performance comparison of EMDL-ac4C and additional machine learning models on ten independent test sets.

Model	Sn	Sp	MCC	AUC	Auprc
SVM	0.7610	0.7668	0.5279	0.8267	0.8301
RF	0.7248	0.7602	0.4854	0.7988	0.6824
NB	0.8046	0.6634	0.4956	0.8027	0.7904
LR	0.6998	0.6893	0.3892	0.7749	0.7626
LGB	0.7030	0.8081	0.5141	0.8360	0.8538
KNN	0.7786	0.6294	0.5028	0.7940	0.7873
DT	0.6998	0.6893	0.3892	0.625040	0.5783
BAG	0.7717	0.740814	0.5427	0.8328	0.8557
ADAB	0.6833	0.7636	0.4486	0.8096	0.8453
EMDL-ac4C	**0.8104**	**0.8173**	**0.6169**	**0.8734**	**0.8643**

The best outcomes are in bold.

In addition, to research whether downsampling ensemble has better prediction ability, we made a comparison between the downsampling ensemble model and single two-branch residual connection DenseNet. On ten independent test sets, we used the ensemble classifier and the non-ensemble classifier to predict ac4C sites based on one-hot coding, and the average results were shown in Figure 9. The ensemble model EMDL-ac4C outperforms the non-ensemble model in Sn, Acc, MCC and AUC by 2.13%, 0.42%, 0.74% and 0.53%, respectively. In contrast, Sp is 0.14% lower, and collectively, the performance of the ensemble model is better than that of the non-ensemble model.

TABLE 3 Average performance of several advanced models on ten test sets.

Model	Sn	Sp	MCC	Acc	AUC
VGG-16	**0.9227**	0.2921	0.2533	0.6074	0.6850
ResNet	0.6673	0.6285	0.3008	0.6479	0.7086
CSPNet	0.8267	0.7349	0.5683	0.7808	0.8576
VGG-19	0.7537	0.6417	0.4304	0.6977	0.8231
Inception V3	0.7934	0.3488	0.1770	0.5711	0.6596
EMDL-ac4C	0.8104	**0.8173**	**0.6169**	**0.8080**	**0.8794**

The best outcomes are in bold.

3.4 Comparison with other machine learning models

To further evaluate the performance of the two-branch residual connection DenseNet model with that of other machine learning

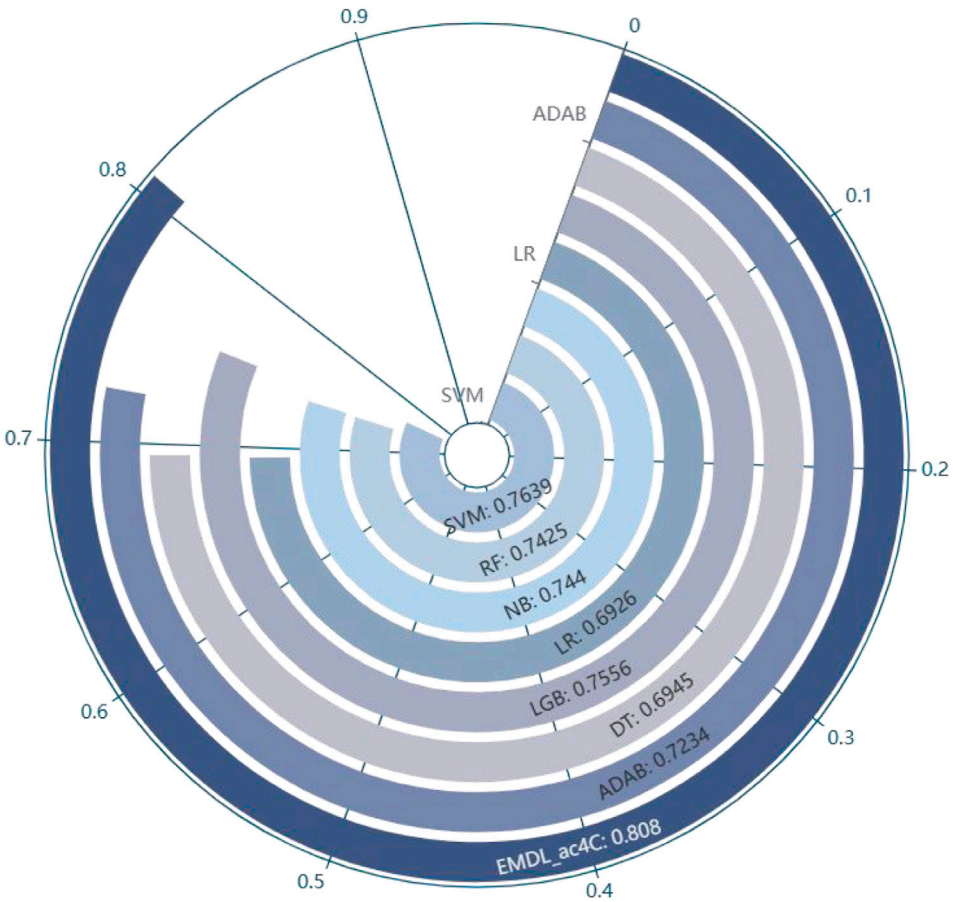


FIGURE 11
Comparison of Average Accuracy of EMDL-ac4C and additional machine learning models on ten independent test sets.

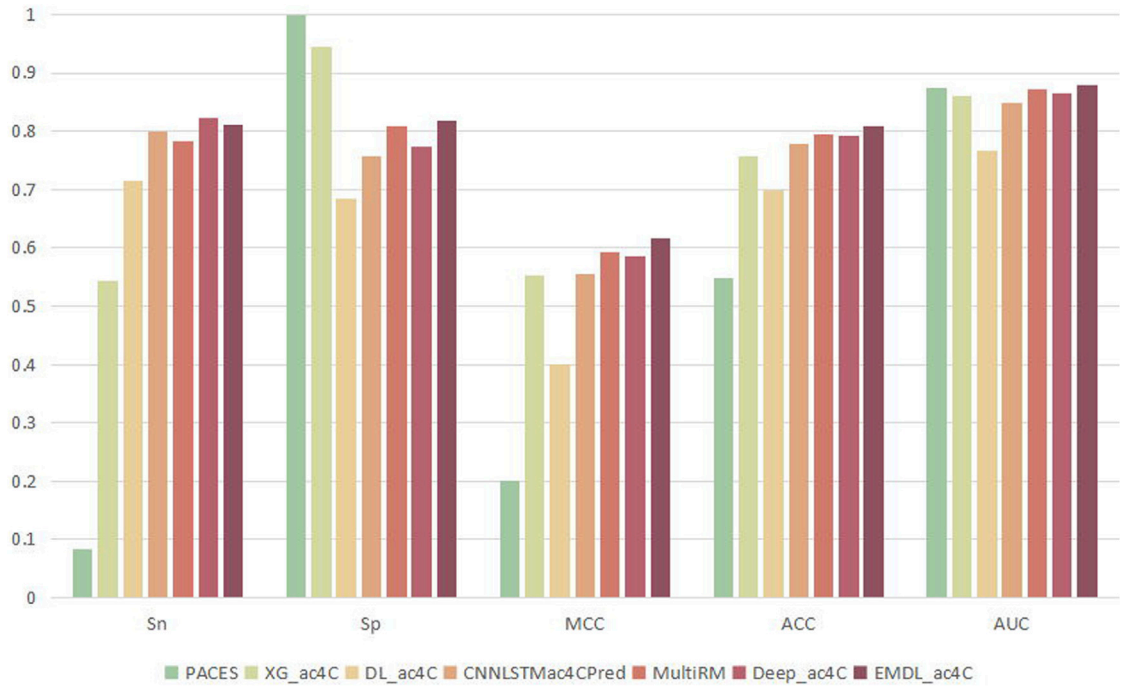


FIGURE 12
Comparison of the average results of ten test sets on various predictors.

models, we performed a comparison of the average results of various coding methods and distinct classifiers on ten test datasets. The ROC curves comparison are shown in Figure 10, where, A-H are the machine learning models: support vector machine (SVM), random forest (RF), naive Bayesian (NB), logistic regression (LR), light gradient boosting machine (LGB), k-nearest neighbor (KNN), bagging (BAG) and adaboost classifier (ADAB). I is the result of two-branch residual connection DenseNet. It is clear from Figure 10 that among the machine learning models, SVM, LGB, and BAG perform well and NB performs the worst. Of course, the combination of the two-branch residual connection DenseNet and one-hot coding has the best performance. Table 2 shows the average performance metrics of EMDL-ac4C and additional machine learning models using one-hot coding on ten independent test sets. All four metrics of EMDL-ac4C return higher values than the other traditional machine learning models.

In balanced datasets, accuracy (Acc) is one of the important metrics to evaluate the classifier performance. Figure 11 visualizes that EMDL-ac4C is more effective than other machine learning models.

3.5 Comparison with other advanced models

To assess the prediction performance of EMDL-ac4C, we compared EMDL-ac4C with several advanced models for analysis, including: Cross Stage Partial DenseNet (CSPNet) (Wang et al., 2020), VGG-16 (Guan et al., 2019), ResNet (He et al., 2016), VGG-19 (Xiao et al., 2020), Inception V3 (Yu et al., 2017). These models performed differently on the ten balanced test sets, as shown in Supplementary Tables S12–S17, while the average prediction results on the ten datasets are shown in Table 3. VGG-16 scored the highest in Sn, but the score of Sp was too low, and the balance between Sn and Sp was lost, with too much deviation, and the prediction accuracy was not high. In contrast, our model EMDL-ac4C reached a balance between Sn and Sp with less than 1% deviation, and obtained the highest Sp, MCC, Acc, and AUC scores among several models, getting the greatest prediction results.

To further evaluate the predictive performance of EMDL-ac4C, we collected three ac4C sites identified by Oxford Nanopore Technology (ONT) from the large public database DirectRMDB (Zhang. et al., 2022) as an additional test set, these three ac4C loci are located at positions 453630, 455959, and 456452 of chromosome NC_001144.5, respectively. After testing, all three ac4C loci were correctly predicted, and the probabilities of predicting positive samples were 0.7923848, 0.9826068, and 0.9666126, respectively. Therefore, we can consider EMDL-ac4C as a high-performance ac4C classifier.

3.6 Comparison of different classifiers

To prove the validity of EMDL-ac4C, we found six models that can be used to predict ac4C loci for comparison, including PACES (Zhao et al., 2019a), XG_ac4C (Alam et al., 2020b), DL_ac4C (Iqbal et al., 2022), CNNLSTMac4CPred(Zhang et al., 2022), MultiRM (Song et al., 2021) and DeepAc4C (Wang et al., 2021). Among these six predictors, PACES and XG_ac4C used machine learning approaches: random forest, XGboost. While DL_ac4C,

TABLE 4 The performance of DeepAc4C and EMDL-ac4C.

Dataset	Model	Training Acc	Validation Acc	Test Acc	Test MCC	Test AUC
D1	DeepAc4C	0.8093	0.8043	0.7934	0.5871	0.8620
	EMDL-ac4C	**0.8911**	**0.8092**	**0.8104**	**0.6217**	**0.8821**
D2	DeepAc4C	0.8195	0.8000	0.7943	0.5921	0.8660
	EMDL-ac4C	**0.8967**	**0.8075**	**0.8126**	**0.6267**	**0.8801**
D3	DeepAc4C	0.8209	0.8043	0.7874	0.5777	0.8641
	EMDL-ac4C	**0.8985**	**0.8097**	**0.8115**	**0.6244**	**0.8832**
D4	DeepAc4C	0.8490	**0.8348**	0.7969	0.5945	0.8658
	EMDL-ac4C	**0.8954**	0.8096	**0.8040**	**0.6102**	**0.8819**
D5	DeepAc4C	0.8403	0.8043	0.7950	0.5902	0.8646
	EMDL-ac4C	**0.8919**	**0.8063**	**0.8019**	**0.6041**	**0.8647**
D6	DeepAc4C	0.7996	0.7913	0.7938	0.5897	0.8645
	EMDL-ac4C	**0.8871**	**0.8076**	**0.7997**	**0.6000**	**0.8665**
D7	DeepAc4C	0.8209	**0.8609**	0.7911	0.5836	0.8671
	EMDL-ac4C	**0.8915**	0.8179	**0.7976**	**0.5956**	**0.8749**
D8	DeepAc4C	0.8475	0.8043	0.7978	0.5959	0.8657
	EMDL-ac4C	**0.8880**	**0.8056**	**0.8147**	**0.6306**	**0.8809**
D9	DeepAc4C	0.8078	0.8130	0.7846	0.5703	0.8615
	EMDL-ac4C	**0.8308**	**0.8177**	**0.8169**	**0.6340**	**0.8908**
D10	DeepAc4C	0.8277	**0.8217**	0.7850	0.5725	0.8680
	EMDL-ac4C	**0.8924**	0.8201	**0.8147**	**0.6295**	**0.8913**
Average	DeepAc4C	0.8242	**0.8139**	0.7919	0.5857	0.8649
	EMDL-ac4C	**0.8927**	0.8111	**0.8084**	**0.6177**	**0.8794**

The best outcomes are in bold.

TABLE 5 The AUC values and Aupr values of PACES, XG-ac4C and Two-branch residual connection DenseNet.

Dataset	Methods	AUC	Aupr
Cross-validation	PACES	0.885	0.559
	XG-ac4C	**0.910**	**0.653**
	Two-branch residual connection DenseNet	0.904	0.615
Independent-test	PACES	0.874	0.485
	XG-ac4C	0.889	0.581
	Two-branch residual connection DenseNet	**0.901**	**0.594**

The best outcomes are in bold.

CNNLSTMac4CPred, MultiRMand DeepAc4C used the deep learning approach. For a fair comparison, all seven predictors were tested using the same training and testing sets, and the predicted results were compared to determine their performance. Figure 12 had shown the average result of the seven predictors on ten test sets. Among them, EMDL-ac4C had the best comprehensive performance, followed by the deep learning model MultiRM, and then DeepAc4C, CNNLSTMac4CPred, XG_ac4C and DL_ac4C. PACES had the worst performance. Compared to MultiRM, EMDL-ac4C was 2.88%, 0.79%, 2.33%, 1.25% and 0.85% higher for Sn, Sp, MCC, ACC, and AUC, respectively. Meanwhile, compared to DeepAc4C, EMDL-ac4C was 4.39%, 3.12%, 1.61%, and 1.45% higher in Sp, MCC, Acc, and AUC, respectively. In addition, EMDL-ac4C was also higher than CNNLSTMac4CPred in all metrics, and the average is 3.93% higher. For Sp, PACES and XG_ac4C had higher return values than EMDL-ac4C. Nevertheless, in particular, the difference between the Sn and Sp values of PACES and XG_ac4C was too large, even reaching 91.52% for PACES, and too low Sn indicated that few positive samples were identified, which was not a good phenomenon. For the balanced dataset, we pay more attention to the return value of Acc, and the larger the Acc, the better the model performance. Our model EMDL-ac4C had higher Acc metrics than DL_ac4C, XG_ac4C and PACES: 10.89%, 5.08% and 26.08% higher, respectively. DeepAc4C is the most advanced model at present, so the detailed analysis of EMDL-ac4C and DeepAc4C is compared, see Table 4.

Some of the classification performance metrics of EMDL-ac4C and DeepAc4C on the training set, validation set, and test set were listed in Table 4 with comparative analysis of the two predictors. EMDL-ac4C

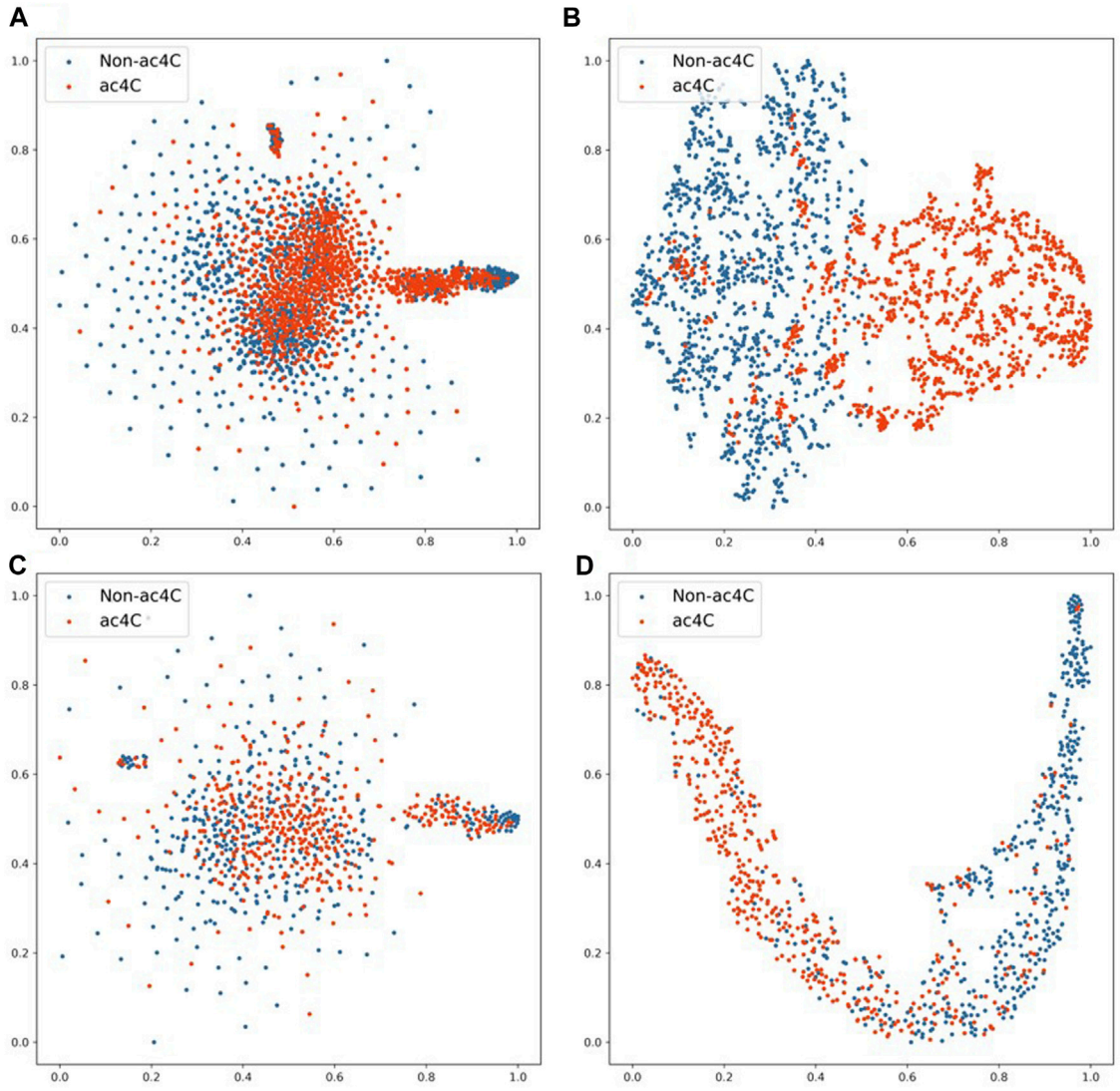


FIGURE 13
2D t-SNE visualization of the training and testing sets. **(A)** 2D t-SNE visualization of the training set with one-hot encoding. **(B)** 2D t-SNE visualization of features learned from the training set by EMDL-ac4C. **(C)** 2D t-SNE visualization of the testing set with one-hot encoding. **(D)** 2D t-SNE visualization of features learned from the testing set by EMDL-ac4C.

had a greater accuracy (Acc) than DeepAc4C in all 10 balanced datasets used for training. Among them, the best performance was on D6, which was 8.75% higher. On the validation dataset, EMDL-ac4C performed relatively poorly on the D4, D7 datasets, and it also performed slightly worse than DeepAc4C on D10. In this regard, we believe that EMDL-ac4C is unable to learn enough information and features from the data due to the small number of data in the validation set, which affects the performance of the model. For the datasets in this paper, the samples used for training in each balanced dataset were 2066, while the amounts of samples used for validation and testing were 230 and 934, correspondingly.

For the ten balanced test sets, EMDL-ac4C obtained good prediction results, with the Acc, MCC and AUC values of each test subset exceeding the corresponding metrics of DeepAc4C. The average Acc of the ten test sets of EMDL-ac4C was 1.61% higher than the average Acc of DeepAc4C, and the average MCC and average AUC values were 3.12% and 1.45% higher, respectively. This showed that the EMDL-ac4C model is effective and also has good generalization. In addition, it can also be seen from Table 4 that the difference of each metric corresponding to the ten balanced data sets is small, which proven that EMDL-ac4C is a stable and reliable model. By comparison, see Supplementary Tables S14, S15, the MCC values of VGG16 and VGG 19 models on different equilibrium datasets vary widely. VGG 16 had a MCC value of 0.99% on D7, and the MCC value on D1 reaches 52.75%, with a variance of 51.76%. The maximum difference in MCC values for VGG19 was also as high as 42.63%, which indicated that the VGG model is not stable on the data sets of this paper.

To evaluate the performance of our model on the unbalanced dataset, we use the dataset downloaded from the PACES website (http://www.rnanut.net/paces/) for testing. In this case, the training set contains 1,160 positive and 10,855 negative samples, respectively, while the test set contains 469 positive and 4,343 negative samples.

The results of the 5-fold cross-validation and independent tests of our base model Two-branch residual connection DenseNet and two other predictors PACES (Zhao et al., 2019b) and XG-ac4C (Alam et al., 2020a) are shown in Table 5.

On cross-validation, the AUC value of Two-branch residual connection DenseNet is slightly lower than XG-ac4C, but exceeds PACES, and the same is true for the Aupr value, which may be due to the fact that Two-branch residual connection DenseNet is only the base model, without unbalanced preprocessing of the dataset and without using the ensemble method. In the independent test, Two-branch residual connection DenseNet is better than the other two models, and the difference between the independent test and cross-validation results is small, not more than 0.021, and there is no overfitting, which just shows that Two-branch residual connection DenseNet has a strong generalization ability.

3.7 Visualization of the classification ability of EMDL-ac4C

To test the classification performance of EMDL-ac4C, we selected D1 in ten balanced datasets for validation. After encoding the sequence data with one-hot, We reduced the encoding vector to two-dimensional using t-distribution random neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) method, as shown in Figures 13A, C, where (A) and (C) were the classification effects of the training and testing sets after one-hot encoding using t-SNE downscaling, respectively. In order to compare with (A) (C) in Figure 13, we first extracted the important features from the training and testing data after one-hot encoding using EMDL-ac4C, and then downscaled them by t-SNE, and finally displayed the classification effects as shown in Figures 13B, D. The red circles in Figure 13 stand for the positive class samples, whereas the blue circles for the negative class samples. From Figure 13, we can clearly see that there are more overlapping clusters generated after one-hot coding, which indicates that the quality of one-hot coding is imperfect. In contrast, after EMDL-ac4C extracts the features, fewer overlapping clusters are generated, especially the two classes of clusters generated in the testing set have reached a highly disjoint classification effect, which proves the efficiency of EMDL-ac4C in terms of extracting features, that is, EMDL-ac4C has powerful classification ability.

4 Conclusion

In this work, we built a downsampling ensemble learning model called EMDL-ac4C, which aimed to predict ac4C sites from sequence fragments of RNA. To effectively identify the ac4C locus, we had done a lot of work at both sequence encoding and feature extraction levels. Firstly, we had compared five commonly used feature encoding schemes and found that the combination of simple one-hot encoding and deep learning models can identify ac4C loci more efficiently. Second, we proposed the ensemble learning model EMDL-ac4C to extract features and predict sites, whose underlying learner was two-branch residual connection DenseNet. Compared with other advanced models and predictors for identifying ac4C, EMDL-ac4C obtained superior performance in independent tests, which proved EMDL-ac4C's powerful feature learning capability and predictive power. We will develop the model and increase its prediction power in subsequent studies. For instance, we will be able to anticipate multi-class sites simultaneously, such as 6ma, 4mc, *etc.*

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

The studies were created and planned by JJ and ZW. ZW carried out the feature extraction, model building, deep learning, and performance assessment. The manuscript was written by ZW and revised by JJ and XC. This work was supervised by JJ and XC. All authors contributed to the article and approved the submitted version.

Funding

This work was partially supported by the National Natural Science Foundation of China (Nos. 61761023, 62162032, and 31760315), the Natural Science Foundation of Jiangxi Province, China (Nos. 20202BABL202004 and 20202BAB202007), the Scientific Research Plan of the Department of Education of Jiangxi Province, China (GJJ190695 and GJJ212419).

Acknowledgments

The authors are grateful for the constructive comments and suggestions made by the reviewers.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1232038/full#supplementary-material

References

Alam, W., Tayara, H., and Chong, K. T. (2020). XG-ac4C: Identification of N4-acetylcytidine (ac4C) in mRNA using eXtreme gradient boosting with electron-ion interaction pseudopotentials. *Sci. Rep.* 10, 20942. doi:10.1038/s41598-020-77824-2

Arango, D., Sturgill, D., Alhusaini, N., Dillman, A. A., Sweet, T. J., Hanson, G., et al. (2018). Acetylation of cytidine in mRNA promotes translation efficiency. *Cell.* 175, 1872–1886 e24. doi:10.1016/j.cell.2018.10.030

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2019). SDM6A: A web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Nucleic acids.* 18, 131–141. doi:10.1016/j.omtn.2019.08.011

Bastings, J., and Filippova, K. (2020). "The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? " in Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, November 2020, South Carolina, pp. 149–155.

Boccaletto, P., Machnicka, M. A., Purta, E., Piątkowski, P., Bagiński, B., Wirecki, T. K., et al. (2018). Modomics: A database of RNA modification pathways 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi:10.1093/nar/gkx1030

Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., et al. (2019). Whistle: A high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res.* 47, e41. doi:10.1093/nar/gkz074

Chen, L., Wang, W. J., Liu, Q., Wu, Y. K., Wu, Y. W., Jiang, Y., et al. (2022). NAT10-mediated N4-acetylcytidine modification is required for meiosis entry and progression in male germ cells. *Nucleic Acids Res.* 50, 10896–10913. doi:10.1093/nar/gkac594

Chen, W., Tang, H., and Lin, H. (2017). MethyRNA: A web server for identification of N6-methyladenosine sites. *J. Biomol. Struct. Dyn.* 35, 683–687. doi:10.1080/07391102.2016.1157761

Chen, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y.-Z., et al. (2021). iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res.* 49, e60. doi:10.1093/nar/gkab122

Dominissini, D., and Rechavi, G. (2018). N(4)-acetylation of cytidine in mRNA by NAT10 regulates stability and translation. *Cell.* 175, 1725–1727. doi:10.1016/j.cell.2018.11.037

Dou, L., Li, X., Ding, H., Xu, L., and Xiang, H. (2020). Prediction of m5C modifications in RNA sequences by combining multiple sequence features. *Nucleic acids.* 21, 332–342. doi:10.1016/j.omtn.2020.06.004

El Allali, A., Elhamraoui, Z., and Daoud, R. (2021). Machine learning applications in RNA modification sites prediction. *Comput. Struct. Biotechnol. J.* 19, 5510–5524. doi:10.1016/j.csbj.2021.09.025

Feng, Z., Li, K., Qin, K., Liang, J., Shi, M., Ma, Y., et al. (2022). The LINC00623/NAT10 signaling axis promotes pancreatic cancer progression by remodeling ac4C modification of mRNA. *J. Hematol. Oncol.* 15, 112. doi:10.1186/s13045-022-01338-9

Gamage, S. T., Sas-Chen, A., Schwartz, S., and Meier, J. L. (2021). Quantitative nucleotide resolution profiling of RNA cytidine acetylation by ac4C-seq. *Nat. Protoc.* 16, 2286–2307. doi:10.1038/s41596-021-00501-9

Guan, Q., Wang, Y., Ping, B., Li, D., Du, J., Qin, Y., et al. (2019). Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: A pilot study. *J. Cancer* 10, 4876–4882. doi:10.7150/jca.28769

Hao, H., Liu, W., Miao, Y., Ma, L., Yu, B., Liu, L., et al. (2022). N4-acetylcytidine regulates the replication and pathogenicity of enterovirus 71. *Nucleic Acids Res.* 50, 9339–9354. doi:10.1093/nar/gkac675

Hasan, M. M., Basith, S., Khatun, M. S., Lee, G., Manavalan, B., and Kurata, H. (2021). Meta-i6mA: An interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Briefings Bioinforma.* 22, bbaa202. doi:10.1093/bib/bbaa202

Hasan, M. M., Tsukiyama, S., Cho, J. Y., Kurata, H., Alam, M. A., Liu, X., et al. (2022). Deepm5C: A deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. *Mol. Ther. J. Am. Soc. Gene Ther.* 30, 2856–2867. doi:10.1016/j.ymthe.2022.05.001

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition." in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 2016, pp. 770–778.

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-Excitation networks." in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018, pp. 7132–7141.

Huang, D., Chen, K., Song, B., Wei, Z., Su, J., Coenen, F., et al. (2022). Geographic encoding of transcripts enabled high-accuracy and isoform-aware deep learning of RNA methylation. *Nucleic Acids Res.* 50, 10290–10310. doi:10.1093/nar/gkac830

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks." in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, June 2017, pp. 4700–4708.

Iqbal, M. S., Abbasi, R., Bin Heyat, M. B., Akhtar, F., Abdelgeliel, A. S., Albogami, S., et al. (2022). Recognition of mRNA N4 acetylcytidine (ac4C) by using non-deep vs. Deep learning. *Deep Learn. Appl. Sci.* 12, 1344. doi:10.3390/app12031344

Ito, S., Akamatsu, Y., Noma, A., Kimura, S., Miyauchi, K., Ikeuchi, Y., et al. (2014). A single acetylation of 18 S rRNA is essential for biogenesis of the small ribosomal subunit in *Saccharomyces cerevisiae*. *J. Biol. Chem.* 289, 26201–26212. doi:10.1074/jbc.M114.593996

Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K. C. (2016). pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.* 394, 223–230. doi:10.1016/j.jtbi.2016.01.020

Jia, J., Sun, M., , et al.Jia, J., Sun, M., Wu, G., Qiu, W. (2023). DeepDN_iGlu: Prediction of lysine glutarylation sites based on attention residual learning method and DenseNet. *Math. Biosci. Eng.* 20, 2815–2830. doi:10.3934/mbe.2023132

Jia, J., Wu, G., Li, M., and Qiu, W. (2022). pSuc-EDBAM: Predicting lysine succinylation sites in proteins based on ensemble dense blocks and an attention module. *BMC Bioinforma.* 23 (1), 450. doi:10.1186/s12859-022-05001-5

Jin, G., Xu, M., Zou, M., and Duan, S. (2020). The processing, gene regulation, biological functions, and clinical relevance of N4-acetylcytidine on RNA: A systematic review. *Mol. Ther. Nucleic Acids* 20, 13–24. doi:10.1016/j.omtn.2020.01.037

Kumbhar, B. V., Kamble, A. D., and Sonawane, K. D. (2013). Conformational preferences of modified nucleoside N(4)-acetylcytidine, ac4C occur at "wobble" 34th position in the anticodon loop of tRNA. *Cell. Biochem. Biophys.* 66, 797–816. doi:10.1007/s12013-013-9525-8

Le, N. Q. K., Ho, Q.-T., Nguyen, V.-N., and Chang, J.-S. (2022). BERT-Promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection. *Comput. Biol. Chem.* 99, 107732. doi:10.1016/j.compbiolchem.2022.107732

Li, Q., Xu, L., Li, Q., and Zhang, L. (2020). Identification and classification of enhancers using dimension reduction technique and recurrent neural network. *Comput. Math. Methods Med.* 2020, 8852258. doi:10.1155/2020/8852258

Liu, Q., Chen, J., Wang, Y., Li, S., Jia, C., Song, J., et al. (2020). DeepTorrent: A deep learning-based approach for predicting DNA N4-methylcytosine sites. *Briefings Bioinforma.* 22, bbaa124. doi:10.1093/bib/bbaa124

Long, Y., Jia, H., Zhong, Y., Jiang, Y., and Jia, Y. (2021). RXDNFuse: A aggregated residual dense network for infrared and visible image fusion. *Inf. Fusion* 69, 128–141. doi:10.1016/j.inffus.2020.11.009

Lv, H., Zhang, Z.-M., Li, S.-H., Tan, J.-X., Chen, W., and Lin, H. (2020). Evaluation of different computational methods on 5-methylcytosine sites identification. *Briefings Bioinforma.* 21, 982–995. doi:10.1093/bib/bbz048

Mezzar, S., de Schryver, E., and Van Veldhoven, P. P. (2014). RP-HPLC-fluorescence analysis of aliphatic aldehydes: Application to aldehyde-generating enzymes HACL1 and SGPL1. *J. Lipid Res.* 55, 573–582. doi:10.1194/jlr.D044230

Rehman, M. U., Tayara, H., and Chong, K. T. (2021). DCNN-4mC: Densely connected neural network based N4-methylcytosine site prediction in multiple species. *Comput. Struct. Biotechnol. J.* 19, 6009–6019. doi:10.1016/j.csbj.2021.10.034

Sharma, S., Langhendries, J. L., Watzinger, P., Kotter, P., Entian, K. D., and Lafontaine, D. L. (2015). Yeast Kre33 and human NAT10 are conserved 18S rRNA cytosine acetyltransferases that modify tRNAs assisted by the adaptor Tan1/THUMPD1. *Nucleic Acids Res.* 43, 2242–2258. doi:10.1093/nar/gkv075

Sharma, S., Marchand, V., Motorin, Y., and Lafontaine, D. L. J. (2017). Identification of sites of 2′-O-methylation vulnerability in human ribosomal RNAs by systematic mapping. *Sci. Rep.* 7, 11490. doi:10.1038/s41598-017-09734-9

Shi, Z., Wang, T., Huang, Z., Xie, F., and Song, G. (2021). A method for the automatic detection of myopia in Optos fundus images based on deep learning. *Int. J. Numer. Methods Biomed. Eng.* 37, e3460. doi:10.1002/cnm.3460

Song, B., Wang, X., Liang, Z., Ma, J., Huang, D., Wang, Y., et al. (2023). RMDisease V2.0: An updated database of genetic variants that affect RNA modifications with disease and trait implication. *Nucleic Acids Res.* 51, D1388–D1396. doi:10.1093/nar/gkac750

Song, Z., Huang, D., Song, B., Chen, K., Song, Y., Liu, G., et al. (2021). Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. *Nat. Commun.* 12, 4011. doi:10.1038/s41467-021-24313-3

Sturgill, D., Arango, D., and Oberdoerffer, S. (2022). Protocol for base resolution mapping of ac4C using RedaC:T-seq. *Star. Protoc.* 3, 101858. doi:10.1016/j.xpro.2022.101858

Tardu, M., Jones, J. D., Kennedy, R. T., Lin, Q., and Koutmou, K. S. (2019). Identification and quantification of modified nucleosides in *Saccharomyces cerevisiae* mRNAs. *ACS Chem. Biol.* 14, 1403–1409. doi:10.1021/acschembio.9b00369

Thomas, J. M., Bryson, K. M., and Meier, J. L. (2019). Nucleotide resolution sequencing of N4-acetylcytidine in RNA. *Methods Enzym.* 621, 31–51. doi:10.1016/bs.mie.2019.02.022

Tsai, K., Jaguva Vasudevan, A. A., Martinez Campos, C., Emery, A., Swanstrom, R., and Cullen, B. R. (2020). Acetylation of cytidine residues boosts HIV-1 gene expression

by increasing viral RNA stability. *Cell. Host Microbe* 28, 306–312 e6. doi:10.1016/j.chom.2020.05.011

Tsukiyama, S., Hasan, M. M., Deng, H.-W., and Kurata, H. (2022). BERT6mA: Prediction of DNA N6-methyladenine site using deep learning-based approaches. *Briefings Bioinforma.* 23, bbac053. doi:10.1093/bib/bbac053

van der Maaten, L., and Hinton, G. (2008). Viualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach, CA, December 2017.

Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., and Yeh, I.-H. (2020a). "CSPNet: A new backbone that can enhance learning capability of CNN." in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, June 2020, pp. 390–391.

Wang, C., Ju, Y., Zou, Q., and Lin, C. (2021a). DeepAc4C: A convolutional neural network model with hybrid features composed of physicochemical patterns and distributed representation information for identification of N4-acetylcytidine in mRNA. *Bioinformatics* 38, 52–57. doi:10.1093/bioinformatics/btab611

Wang, C., Zou, Q., Ju, Y., and Shi, H. (2023). Enhancer-FRL: Improved and robust identification of enhancers and their activities using feature representation learning. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20, 967–975. doi:10.1109/TCBB.2022.3204365

Wang, H., Yan, Z., Liu, D., Zhao, H., and Zhao, J. (2020b). MDC-kace: A model for predicting lysine acetylation sites based on modular densely connected convolutional networks. *IEEE Access* 8, 214469–214480. doi:10.1109/ACCESS.2020.3041044

Wang, H., Zhao, H., Yan, Z., Zhao, J., and Han, J. (2021b). MDCAN-lys: A model for predicting succinylation sites based on multilane dense convolutional attention network. *Biomolecules* 11, 872. doi:10.3390/biom11060872

Wang, L., Li, Y., Peng, S., Tang, X., and Yin, B. (2021c). Multi-level feature fusion network for crowd counting. *IET Comput. Vis.* 15, 60–72. doi:10.1049/cvi2.12012

Wei, L., Luan, S., Nagai, L. A. E., Su, R., Zou, Q., and Hancock, J. (2019a). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* 35, 1326–1333. doi:10.1093/bioinformatics/bty824

Wei, Z., Song, H., Chen, L., Li, Q., and Han, G. (2019b). Attention-based DenseUnet network with adversarial training for skin lesion segmentation. *IEEE Access* 7, 136616–136629. doi:10.1109/ACCESS.2019.2940794

Weizhong, L., and Godzik, A. (2006). Cd-Hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22 (13), 1658–1659. doi:10.1093/bioinformatics/btl158

Xiao, J., Wang, J., Cao, S., and Li, B. (2020). Application of a novel and improved VGG-19 network in the detection of workers wearing masks. *J. Phys. Conf. Ser.* 1518, 012041. doi:10.1088/1742-6596/1518/1/012041

Yan, Y., Ni, B., Liu, J., and Yang, X. (2019). Multi-level attention model for person re-identification. *Pattern Recognit. Lett.* 127, 156–164. doi:10.1016/j.patrec.2018.08.024

Yang, S., Yang, Z., and Yang, J. (2023). 4mCBERT: A computing tool for the identification of dna N4-methylcytosine sites by sequence- and chemical-derived information based on ensemble learning strategies. *Int. J. Biol. Macromol.* 231, 123180. doi:10.1016/j.ijbiomac.2023.123180

Yang, W., Li, H. Y., Wu, Y. F., Mi, R. J., Liu, W. Z., Shen, X., et al. (2021). ac4C acetylation of RUNX2 catalyzed by NAT10 spurs osteogenesis of BMSCs and prevents ovariectomy-induced bone loss. *Mol. Ther. Nucleic Acids* 26, 135–147. doi:10.1016/j.omtn.2021.06.022

Yang, Y., Zhong, Z., Shen, T., and Lin, Z. (2018). "Convolutional neural networks with alternately updated clique." in Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, June 2018, pp. 2413–2422.

Yu, H., and Dai, Z. (2019). SNNRice6mA: A deep learning method for predicting dna N6-methyladenine sites in rice genome. *Front. Genet.* 10, 1071. doi:10.3389/fgene.2019.01071

Yu, W., Chang, J., Yang, C., Zhang, L., Shen, H., Xia, Y., et al. (2017). "Automatic classification of leukocytes using deep neural network." in Proceedings of the 2017 IEEE 12th International Conference on ASIC (ASICON), Guiyang, China, October 2017, pp. 1041–1044.

Zhang, G., Luo, W., Lyu, J., Yu, Z. G., and Huang, G. (2022a). CNNLSTMac4CPred: A hybrid model for N4-acetylcytidine prediction. *Interdiscip. Sci.* 14, 439–451. doi:10.1007/s12539-021-00500-0

Zhang, Y., Jiang, J., Ma, J., Wei, Z., Wang, Y., Song, B., et al. (2022b). DirectRMDB: A database of post-transcriptional RNA modifications unveiled from direct RNA sequencing technology. *Nucleic Acids Res.* 51, D106–D116. doi:10.1093/nar/gkac1061

Zhao, W., Zhou, Y., Cui, Q., and Zhou, Y. (2019). Paces: Prediction of N4-acetylcytidine (ac4C) modification sites in mRNA. *Sci. Rep.* 9, 11112. doi:10.1038/s41598-019-47594-7

Zhou, Q., Zhou, Z., Chen, C., Fan, G., Chen, G., Heng, H., et al. (2019). Grading of hepatocellular carcinoma using 3D SE-DenseNet in dynamic enhanced MR images. *Comput. Biol. Med.* 107, 47–57. doi:10.1016/j.compbiomed.2019.01.026

Zhou, Y., Zeng, P., Li, Y.-H., Zhang, Z., and Cui, Q. (2016). Sramp: Prediction of mammalian N ^{6} -methyladenosine (m ^{6} A) sites based on sequence-derived features. *Nucleic Acids Res.* 44, e91. doi:10.1093/nar/gkw104

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: Gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA (New York, N.Y.)* 25, 205–218. doi:10.1261/rna.069112.118

frontiers | Frontiers in Genetics

TYPE Brief Research Report
PUBLISHED 26 July 2023
DOI 10.3389/fgene.2023.1213829

OPEN ACCESS

EDITED BY
Nathan Olson,
National Institute of Standards and Technology (NIST), United States

REVIEWED BY
Dieter Maurice Tourlousse,
National Institute of Advanced Industrial Science and Technology (AIST), Japan
Hassan Ghazal,
National Center for Scientific and Technical Research (CNRST), Morocco

*CORRESPONDENCE
Christian Waechter,
christian.waechter@staff.uni-marburg.de
Volker Ruppert,
ruppert@med.uni-marburg.de

RECEIVED 28 April 2023
ACCEPTED 10 July 2023
PUBLISHED 26 July 2023

CITATION
Waechter C, Fehse L, Welzel M, Heider D, Babalija L, Cheko J, Mueller J, Pöling J, Braun T, Pankuweit S, Weihe E, Kinscherf R, Schieffer B, Luesebrink U, Soufi M and Ruppert V (2023), Comparative analysis of full-length 16s ribosomal RNA genome sequencing in human fecal samples using primer sets with different degrees of degeneracy. *Front. Genet.* 14:1213829. doi: 10.3389/fgene.2023.1213829

Comparative analysis of full-length 16s ribosomal RNA genome sequencing in human fecal samples using primer sets with different degrees of degeneracy

Christian Waechter[1,2]*, Leon Fehse[3], Marius Welzel[3], Dominik Heider[3], Lek Babalija[1], Juan Cheko[1], Julian Mueller[1], Jochen Pöling[2], Thomas Braun[2], Sabine Pankuweit[1], Eberhard Weihe[4], Ralf Kinscherf[4], Bernhard Schieffer[1], Ulrich Luesebrink[1], Muhidien Soufi[1,5] and Volker Ruppert[1,5]*

[1]Department of Cardiology, University Hospital Marburg, Philipps University Marburg, Marburg, Germany, [2]Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany, [3]Department of Mathematics and Computer Science, Philipps University Marburg, Marburg, Germany, [4]Institute of Anatomy and Cell Biology, Medical Faculty, Philipps University Marburg, Marburg, Germany, [5]Center for Undiagnosed and Rare Diseases, University Hospital Marburg, Philipps University Marburg, Marburg, Germany

Next-generation sequencing has revolutionized the field of microbiology research and greatly expanded our knowledge of complex bacterial communities. Nanopore sequencing provides distinct advantages, combining cost-effectiveness, ease of use, high throughput, and high taxonomic resolution through its ability to process long amplicons, such as the entire 16s rRNA genome. We examine the performance of the conventional 27F primer (27F-I) included in the 16S Barcoding Kit distributed by Oxford Nanopore Technologies (ONT) and that of a more degenerate 27F primer (27F-II) in the context of highly complex bacterial communities in 73 human fecal samples. The results show striking differences in both taxonomic diversity and relative abundance of a substantial number of taxa between the two primer sets. Primer 27F-I reveals a significantly lower biodiversity and, for example, at the taxonomic level of the phyla, a dominance of *Firmicutes* and *Proteobacteria* as determined by relative abundances, as well as an unusually high ratio of *Firmicutes*/*Bacteriodetes* when compared to the more degenerate primer set (27F-II). Considering the findings in the context of the gut microbiomes common in Western industrial societies, as reported in the American Gut Project, the more degenerate primer set (27F-II) reflects the composition and diversity of the fecal microbiome significantly better than the 27F-I primer. This study provides a fundamentally relevant comparative analysis of the *in situ* performance of two primer sets designed for sequencing of the entire 16s rRNA genome and suggests that the more degenerate primer set (27F-II) should be preferred for nanopore sequencing-based analyses of the human fecal microbiome.

KEYWORDS

16S rRNA, gut microbiome, human fecal microbiome, next-generation sequencing (NGS), nanopore sequencing, oxford nanopore technologies (ONT), MinION Mk1C, primer degeneracy

Introduction

The study of complex bacterial communities associated with the human body, known as the microbiome, has experienced unprecedented growth over the past two decades and is currently one of the most intensively studied areas in biomedical science (Jones, 2013). The gut microbiome has been a particular focus of interest, as alterations in its complex and highly diverse composition are emerging as potential diagnostic biomarkers or pathogenetic factors for a plethora of disease (Gomaa, 2020). Accordingly, researchers, physicians and patients have high hopes for the further deciphering of microbial signatures and expect great therapeutic potential from specific modulation of the microbiome. The door-opener for this development has been the advent of next-generation sequencing (NGS) technologies, which has enabled a large number of laboratories to analyze complex microbiological communities by allowing rapid, accurate, and comparatively inexpensive sequencing of large amounts of DNA (Malla et al., 2019). The available sequencing platforms can be distinguished, for example, according to their ability to deliver different read lengths, and can thus be classified into short- and long-read technologies. The most widely used technology to date in microbiome research, including the American Gut Project and the Human Microbiome Project, is the Illumina platform, which delivers short reads with a maximum length of 2 × 300 base pairs (bp) using its latest version of the MiSeq® system (Huttenhower et al., 2012; McDonalda et al., 2018; Ravi et al., 2018). As targeted sequencing of 16S ribosomal RNA (16s rRNA) gene has become the established method for amplicon-based identification of bacterial taxa in complex microbiological communities, one drawback of the Illumina technique has become apparent: Its short read length. The ~1,500 bp bacterial 16S rRNA gene contains nine hypervariable regions (V1-V9) interrupted by highly conserved segments, which are suitable as anchor sequences for PCR primers. Illumina-based sequencing can therefore only target short fragments of the 16s rRNA gene, which in the majority of studies are amplified with primers targeting the V3 and V4 regions, which limits taxonomic resolution to the genus level, at best (Johnson et al., 2019; Matsuo et al., 2021; Szoboszlay et al., 2023). Third-generation sequencing platforms, such as the nanopore sequencing technology from Oxford Nanopore Technologies (ONT), have overcome these read-length limitations; the newest version of ONT's nanopore sequencing provides an average read length of approximately 15 kbp and is thus easily capable of covering the entire 16s rRNA genome (Wang and Qian, 2009). Compared to the Illumina platform, however, the long read length comes at the expense of lower sequencing accuracy. Since the commercial launch of the ONT system in 2015, continuous improvements in device design, as well as its chemistry and bioinformatics, have reduced the initially very high error rate of about 6% down to well below 2% when using the very recently introduced nanopore sequencing kit Q20+ (LSK112) and the flow cell R10.4 (Delahaye and Nicolas, 2021; Luo et al., 2022). For comparison, Illumina's error rate is between 0.1% and 1% (Stoler and Nekrutenko, 2021). However, even with this comparatively high error rate, the ONT system provides higher taxonomic resolution than short-read sequencing techniques due to its complete coverage of the 16s rRNA gene, which recent studies have shown extends down to the species level (Benítez-Páez et al., 2016; Shin et al., 2016; Nygaard et al., 2020; Matsuo et al., 2021). This impressive evolution of nanopore technology, combined with its convenient workflow and high cost-effectiveness render it an increasingly attractive and promising approach for the analysis of complex microbial communities, such as the human fecal microbiome. In this context, the selection of appropriate primer sets for 16s rRNA amplification is crucial, as it carries a major risk of biasing the detection of microbial signatures detected. Thus, in artificial microbial communities of known composition, it has been demonstrated that the selection of the 16s rRNA region can substantially influence the detected taxonomic diversity (Klindworth et al., 2013; Yang et al., 2016). In contrast, data on this topic are very limited for complex biological samples, especially for full-length 16s rRNA genome sequencing using the ONT platform. First results in this direction were provided by Matsuo et al. describing a primer-associated bias at the species level for *Bifidobacteria* in the context of 16s full-length rRNA sequencing with the nanopore technology (Matsuo et al., 2021). The present study elucidated the effects of primer selection on the microbial signature by systematically comparing the primers included in the very commonly used kit distributed by ONT with a primer set optimized according to the approach of Matsuo et al. (2021) in a large sample of complex human fecal samples.

Materials and methods

Sample collection and DNA extraction

Fecal samples from German donors without a history of relevant digestive tract disease such as chronic inflammatory bowel disease, cancers of the digestive tract or acute systemic or intestinal inflammation were collected using a special paper (#R1101-1-10, Zymo Research, Irvine, CA, United States) placed over the toilet seat to provide a low-germ environment and transferred into tubes containing DNA/RNA shielding buffer (#R1101, Zymo Research). After collection, samples were stored at room temperature and further processed within 3 days. Nucleic acid was extracted using the *Quick*-DNA© HMW MagBead Kit (#D6060, Zymo Research) according to the manufacturer's protocol. DNA purity and quantity were determined using NanoDrop© (ThermoFisher Scientific,

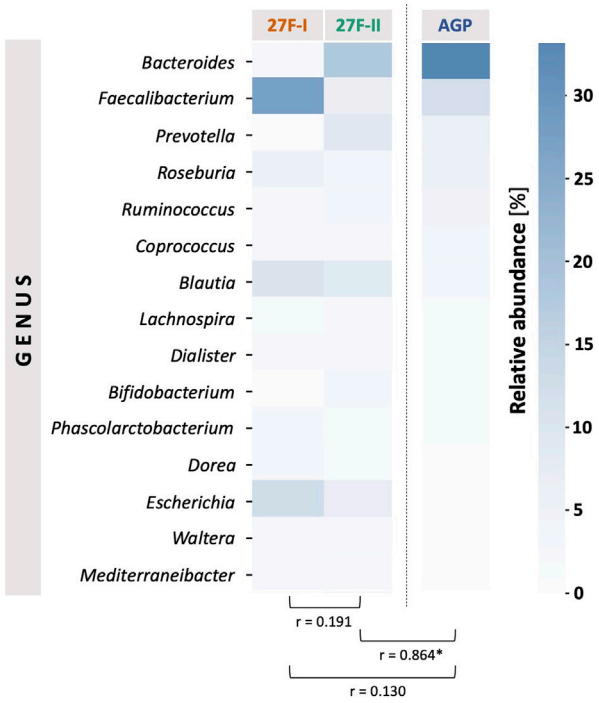


FIGURE 1
Comparison of the mean values of relative genus abundance for the 15 most abundant taxa across samples between the two primer sets using a heatmap and Pearson's correlation (*r*). The American Gut Project (AGP) dataset of samples from individuals without intestinal disease was referenced for a normal taxonomic profile of a human fecal sample. The asterisk indicates statistical significance.

Waltham, MA, United States) and a Quantus© Fluorometer (Promega, Madison, WI, United States), respectively, then stored at −20°C until further use.

PCR amplification and nanopore 16s rRNA gene sequencing

From the DNA extracted as described above, two libraries were prepared, each with a different primer set.

For the construction of the first library (hereafter referred to as 27F-I library), 50 ng of whole genomic DNA was used and processed with the 16S barcoding kit containing the 16s rDNA primers 27F (5′- AGAGTTTGATC**M**TGGCTCAG -3′) and 1492R (5′- CGG TTACCTTGTTACGACTT -3'; numbered according to the *Escherichia coli* rRNA; SQK-RAB204, ONT, Oxford, United Kingdom) according to the manufacturer's protocol.

The second library (hereafter referred to as 27F-II library) was constructed using 50 ng of whole genomic DNA for the first PCR performed (see below) using a comparatively more degenerate 16s rDNA primer set [S-D-Bact-0008-c-S-20 and S-D-Bact-1492-a-A-22, (Klindworth et al., 2013; Matsuo et al., 2021)] with the anchor sequence 5′-TTTCTGTTGGTGCTGATATTGCAG**R**GTT**Y**GAT**Y** **M**TGGCTCAG-3′ plus its reverse primer and with the anchor sequence 5′-ACTTGCCTGTCGCTCTATCTTCCGG**Y**TACCTTG TTACGACTT-3′ plus an appended barcode PCR according to the ONT protocol "Ligation sequencing amplicons - PCR barcoding (SQK-LSK110 with EXP-PBC096)":

1. Preparation of 16s-PCR: 50 ng DNA in 11.5 µL nuclease-free water, 0.5 µL Primer 27F-II, 0.5 µL Primer1492R-II, 12.5 µL LongAMP® Taq 2x Master Mix (New England Biolabs, Ipswich, MA, United States of America). Cycle program: 1 min 95°C; 25 cycles 20 s 95°C, 30 s 51°C, 2 min 65°C and a 5 min final elongation at 65°C.
2. Preparation of barcoding-PCR: 100 fmol 16S-PCR amplicons in 12.0 µL nuclease-free water, 0.5 µL barcode primer, 12.5 µL LongAMP® Taq 2x Master Mix. Cycle program: 1 min 95°C; 15 cycles 20 s 95°C, 30 s 62°C, 2 min 65°C and a 5 min final elongation at 65°C.

After barcoding PCR, the DNA content of each amplicon was determined using a Quantus Fluorometer and adjusted to an equal volume. The amplicons were pooled, and 1 µg was used for library preparation. The library preparation was performed according to the protocol "Ligation sequencing amplicons–PCR barcoding (SQK-LSK110 with EXP-PBC096)" by ONT.

The bold characters in the primer sequences above indicate the degenerate bases, according to the code of the International Union of Biochemistry (IUB). This results in three different sequences for the 27F-I primer approach and 18 for the 27F-II primer approach (16 for the forward primer, 2 for the reverse primer). All sequence variants of the two primer approaches are listed in Supplementary Table S1.

The bar-coded libraries (27F-I and 27F-II libraries) were loaded and subsequently each sequenced on a separate flow cell (FLO-MIN106D, type R.9.4.1, ONT) using the MinION Mk1C device (ONT). MinKNOW version 22.03.4 (ONT) and Guppy 6.0.7 were used for data acquisition. Both libraries were prepared from DNA obtained by the same extraction procedure.

A total of 1,328,830 reads were generated for the library using the 27F-I primer approach (mean 18,203 reads, SEM 1,201 reads) and 1,578,822 reads were generated for the library using the 27F-II primer approach (mean 21,628 reads, SEM 1,991 reads, $p = 0.14$).

Bioinformatics processing and analysis

Raw data processing was carried out with the Nanopore branch of Natrix (Welzel et al., 2020). Natrix is a modular sequencing read processing pipeline written in the workflow management engine Snakemake (Köster and Rahmann, 2012). The pipeline contains rules for the demultiplexing of raw sequencing data, quality control, removal of additional subsequences such as primer or barcodes, read assembly, dereplication, chimera detection and removal, abundance filtering, identification of representative sequences (either operational taxonomic units (OTU) or amplicon sequence variants), taxonomic assignment and additional assignment of meta information (for example, their functional roles or common habitats). A typical workflow used for this study is provided in the Supplementary Figure S1). Natrix processes a set of three file groups: A configuration file with user-configurable parameters (the choice of parameters used in this study are available in

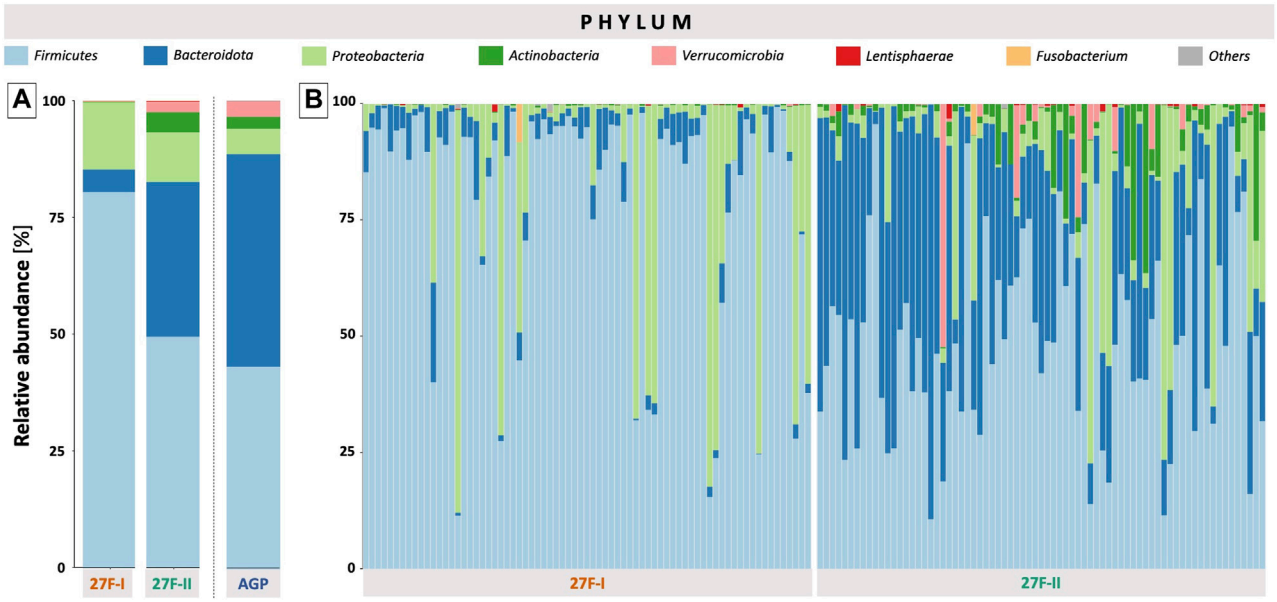


FIGURE 2
Overview of the relative abundance of the different phyla averaged over all samples **(A)** and at the individual sample level **(B)**. In Panel **(A)**, the American Gut Project (AGP) dataset of samples from individuals without intestinal disease was used as a reference for a normal phyla-level taxonomic profile of a human fecal sample.

Supplementary Presentation S1), a primer table, containing information of additional subsequences for each sample, and the raw sequence files in FASTQ format. Natrix supports the filtering out of sequences lower than a user-defined quality score. For the initial quality filtering in this study, we used PRINSEQ (Schmieder and Edwards, 2011) with a mean quality threshold of 15, which corresponds to a maximum mean probability of a wrongly called base of around three percent. Every sequence read, that had a lower mean quality value below 15, was removed from further processing. We chose a more stringent quality threshold then the commonly used thresholds of 7–10 (Delahaye and Nicolas, 2021; Lee et al., 2021) for Nanopore data to reduce the probability of erroneous reads distorting the downstream analysis.

The removal of primers was carried out using a customized Porechop (https://github.com/rrwick/Porechop) version (https://github.com/MW55/Porechop). Porechop is a tool for the removal of adapter sequences in Nanopore reads. While the original Porechop version only searches for a hardcoded set of commonly used primers and can only remove a fixed number of bases from the end of a read, the customized version allows the definition of the primers by the user, and the removal of a fixed number of bases from both ends of the read. The minimal read length for a read to not be discarded by Porechop was set to 1,000 bases, while the maximal read length was set to 2000 bases. Additionally, Porechop was used to remove the first 100 bases from both ends to account for the decrease in read quality of Nanopore reads at both ends (Delahaye and Nicolas, 2021).

As the reads generated in this study were not paired end, no read assembly was carried out. The dereplication was carried out using the CD-HIT-EST algorithm (Fu et al., 2012) with a identity threshold of 1, to only combine reads that are 100% identical. The chimera detection utilized the uchime3_*denovo* algorithm of VSEARCH (Rognes et al., 2016), with the parameters beta 8.0, abskew 16 and pseudo_count 1.2. Natrix supports both the generation of OTUs and ASVs, but, as the ASV generation is carried out using DADA2, which uses a statistical model of Illumina error profiles (Callahan et al., 2016), the OTU modules of the pipeline were chosen for the generation of sequence clusters. OTUs were identified using VSEARCH (Rognes et al., 2016), using a similarity threshold of 85% and a minimal cluster size of 10 sequences. Compared to the more stringend 97% similarity threshold commonly used for OTU generation (Welzel et al., 2020), the lower similarity threshold was chosen to account for the increased error rates of Nanopore sequencing, compared to Illumina sequencing. Taxonomic information was assigned to the OTUs using the National Center for Biotechnology Information (NCBI) nucleotide (nt) database (Sayers et al., 2021) (latest as of 06/22), that contains sequences from Genbank (Benson et al., 2013), Refseq (O'Leary et al., 2016), TPA (Benson et al., 2015) and PDB (Berman et al., 2000), with the use of the nucleotide-nucleotide BLAST (BLASTn) algorithm of the BLAST+ (Camacho et al., 2009) toolkit. The taxonomic assignment was performed using a minimal identity overlap between target and query sequence of 90% and an E-value threshold of 10^{-51}. The parameter max_target_seqs, which corresponds to the amount of hits per query that are returned by BLAST (Shah et al., 2018), was set to 10, with a subsequent filtering step that assigned the target sequence with the highest percentage identity times the logarithm of the alignment length to the query sequence.

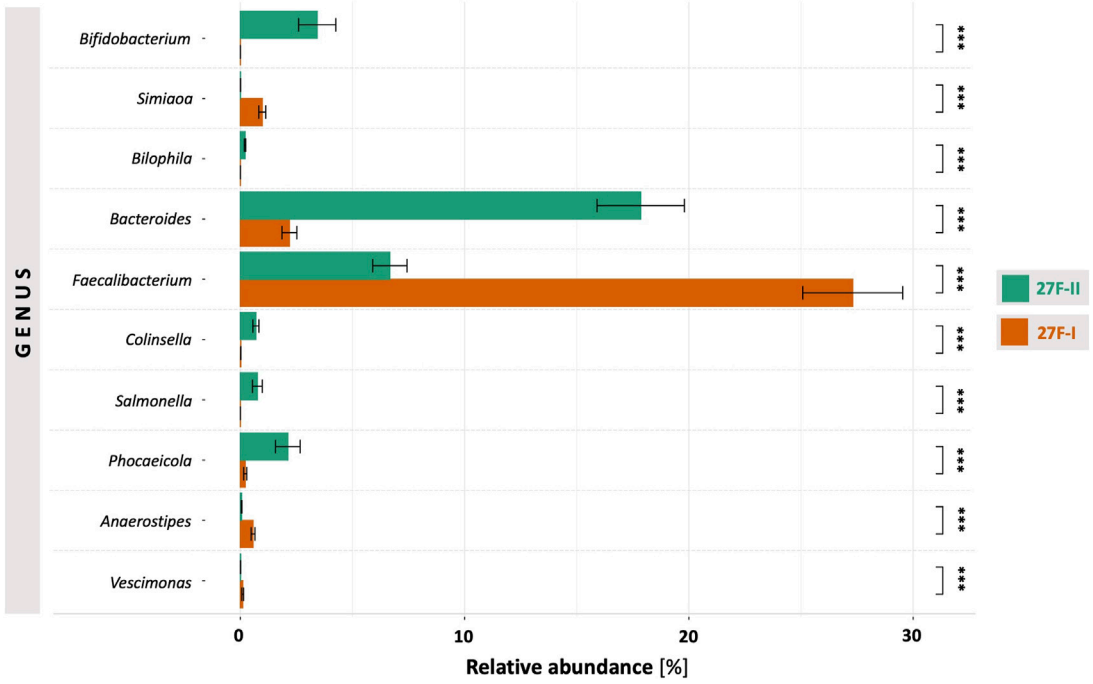


FIGURE 3
Comparison of genera with the most significant differences in abundance between the two primer approaches. ***–p-value < 0.001.

Statistical analyses

All statistical analyses and visualizations were performed using either the statistical programming language R with the *microeco* package (Liu et al., 2020), or the programming language Python. To compare the taxonomic composition via relative abundance on genus level between the two datasets acquired by sequencing using the two different primer sets 27F-I and 27F-II, a Pearson's correlation test was performed. For the further statistical comparisons between the two datasets, the relative abundance on all different taxonomic levels, as well as the results for the alpha biodiversity measured via a Shannon Index, Wilcoxon singed-rank tests were performed and resulting p-values were corrected with the Benjamini-Hochberg method. All statistical tests performed accounted for the nature of paired analyses. A two-tailed p-value <0.05 was considered statistically significant.

Results

Using a full-length 16S rRNA gene amplicon sequencing approach on the nanopore platform, we investigated the performance of the conventional 27F primer (hereafter referred to as 27F-I) included in the 16S Barcoding Kit (SQK-16S024) distributed by ONT, and that of a more degenerate 27F primer (hereafter referred to as 27F-II) covering possible polymorphisms in the conserved regions of the 16s rRNA genome in the context of highly complex bacterial communities derived from 73 human fecal samples. The comparative primer approach used is based on the four-primer PCR method described by Matsuo et al. (Klindworth et al., 2013; Matsuo et al., 2021), consisting of a PCR step with a more degenerate 27F and 1492R primer pair (S-D-Bact-0008-c-S-20 and S-D-Bact-1492-a-A-22, (Klindworth et al., 2013)) followed by barcoding PCR. OTUs were generated from the classifiable reads of the respective primer sets via alignment with the NCBI reference database and systematically compared.

For a global comparison of the taxonomic profiles of human gut microbiota between the two primer sets, the Pearson correlation coefficient (r) was computed based on the mean values for the relative genera abundances across the samples for each primer approach. This revealed only a weak, statistically insignificant correlation (r = 0.191, p = 0.495) between the genera determined for the respective primer sets. To provide an estimate of which of the two primers more accurately maps the fecal microbiome, the taxonomic data generated by primers 27F-I and 27F-II were compared to an American Gut Project (AGP) dataset containing 3,560 samples from subjects without intestinal disease (McDonalda et al., 2018). This showed a statistically significant correlation between the taxonomic profile of fecal samples generated with primer 27F-II and the AGP dataset (r = 0.864, p = 3.29e-05). In contrast, there was only a weak, statistically insignificant correlation between the taxonomic profiles of the fecal samples generated with primer 27F-I and the AGP dataset (r = 0.130, p = 0.638). Figure 1 illustrates the comparison of relative genus abundance for the 15 most abundant taxa between the two primer sets using a heatmap. On further consideration, a clear discrepancy in the relative abundance is already evident at the taxonomic level of the phyla. The mean of all analyzed samples shows that the use of primer 27F-I results in a significantly higher abundance of *Firmicutes* (80.4% vs. 49.4%, p < 0.001) and a lower abundance

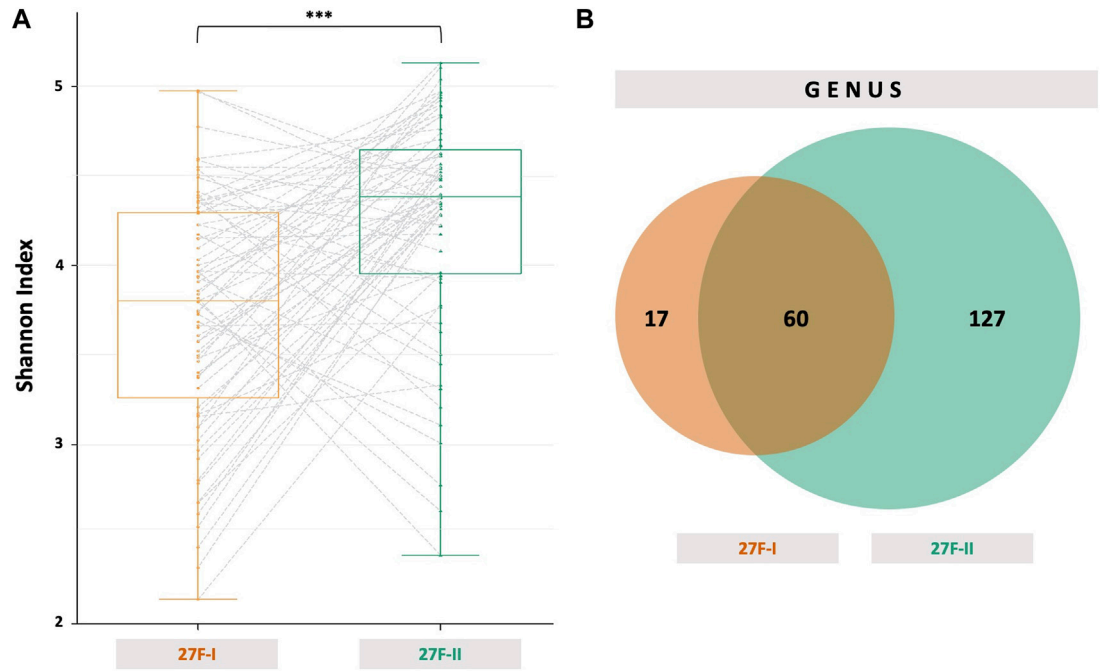


FIGURE 4
Alpha diversity represented as Shannon index **(A)** for the two primer approaches and a Venn diagram **(B)** showing the common and specific taxonomic units at the genus level between the two primer sets used. The dashed gray lines in **(A)** link the results of a sample after analysis with 27F-I and 27F-II primer set, respectively. ***–p-value < 0.001.

of *Bacteroidota* (4.8% vs. 33.2%, $p < 0.001$), *Actinobacteria* (0.1% vs. 4.3%, $p < 0.001$), *Verrucomicrobia* (0.01% vs. 2.2%, $p < 0.001$) compared to primer 27F-II. Consequently, this leads to a significant divergence in the *Firmicutes/Bacteroidota* ratio (16.7 vs. 1.5, $p < 0.001$), discussed as a marker for dysbiosis (Ley et al., 2006), between the two primer sets. Figure 2 provides an overview of the relative abundance of the different phyla averaged over all samples and at the individual sample level. Supplementary Table S2 reports quantitative data on all bacterial phyla for the two sets of primers. At the taxonomic level of genera, statistically significant differences in relative abundance can be observed for a total of 125 distinct genera. When restricted to the ten genera with the most significant differences in abundance, the use of primer 27F-I results in a higher abundance of *Faecalibacterium* (7.306% vs. 6.666%, $p < 0.001$), *Simiaoa* (0.977% vs. 0.003%, $p < 0.001$), *Anaerostipes* (0.565% vs. 0.057%, $p < 0.001$) and *Vescimonas* (0.106% vs. 0.005%, $p < 0.001$) and a lower abundance of *Bacteroides* (2.189% vs. 17.853%, $p < 0.001$), *Bifidobacterium* (0.000% vs. 3.427%, $p < 0.001$), *Phocaeicola* (0.216% vs. 2.118%, $p < 0.001$), *Salmonella* (0.000% vs. 0.759%, $p < 0.001$), *Clinsella* (0.021% vs. 0.693%, $p < 0.001$) and *Bilophila* (0.000% vs. 0.210%, $p < 0.001$) compared to primer 27F-II as shown in Figure 3. Complete quantitative data for all genera are provided in Supplementary Table S3. As the 16S Barcoding Kit (SQK-16S024) containing primer 27F-I has only been validated for genera-level resolution, species-level resolution was not performed.

In addition to the outlined discrepancies in the taxonomic profiles of the human gut microbiome, significant differences in taxonomic diversity are also apparent, depending on the primer set used. Primer 27F-I detects notably fewer distinct OTUs in the human fecal samples than primer set 27F-II, as reflected by a statistically significant lower Shannon index (3.733 vs. 4.271, $p < 0.001$, Figure 4) expressing alpha diversity.

Discussion

The introduction of next-generation sequencing has revolutionized the field of microbiology research and greatly expanded our knowledge of complex human enteric bacterial communities. In this context, nanopore sequencing stands out as combining the advantages of cost-effectiveness, simplicity of use, high throughput, and high taxonomic resolution through its ability to read long amplicons. Recent substantial advances in sequencing accuracy significantly mitigate what has been a significant weakness of this technology and represent a culmination of this impressively rapid technical evolution of the nanopore platform, which has allowed it to eclipse the performance of short-read sequencing techniques. Not least, the 16S Barcoding Kit (SQK-16S024) offered by ONT and widely used in the community contributes notably to the simplicity, speed, and high cost-effectiveness of nanopore 16S rRNA gene sequencing (Santos et al., 2020).

As the choice of primers is known to have a decisive impact on qualitative and quantitative taxonomic signatures (Armougom, 2009), we tested the performance of the primer set included in the commercial 16S Barcoding Kit (referred to here as 27F-I) and compared it to a more degenerate primer set (referred to here as 27F-II) on complex microbial communities. Our analyses demonstrate striking differences in both taxonomic

diversity and relative abundance of a high number of different taxa between the two primer approaches in a large sample of human fecal specimens: The primer set included in the commercial kit (27F-I) results in significantly lower bacterial biodiversity, as measured by the Shannon index and, for example, at the taxonomic level of phyla, a dominance of *Firmicutes* and *Proteobacteria* as measured by relative abundances as well as an unusually high ratio of *Firmicutes/Bacteriodetes* compared to the degenerate primer set (27F-II). These substantial differences in relative abundances of taxa are detectable at all taxonomic levels and result, for example, in a lower relative abundance of the genera *Bacteriodes, Bifidobacterium* and *Phocaeicola* and a higher relative abundance of *Faecalibacterium* when using primer 27F-I compared to using primer 27F-II. Also, Pearson's correlation shows only a weak, statistically non-significant correlation between the taxonomic signatures generated by the two primer sets. Comparing our microbiome data with commonly observed fecal microbiome signatures in Western industrial societies, such as those derived from the American Gut Project (AGP), indicates that the 27F-II primer more reliably reflects the fecal microbial composition and diversity than the primer 27F-I (Huttenhower et al., 2012). Despite a comparable Western lifestyle and level of urbanization of our population and the AGP population, as well as a sequencing approach that is also 16s rRNA gene amplicon-based targeting V4 region, such a comparison can only be indicative and is subject to several limitations. Apart from the fundamentally different sampling, both sample collection and DNA extraction were performed according to different protocols. The main difference between the AGP dataset and the present data is the use of the Illumina short-read sequencing platform with all its differences, e.g., selection of the conserved region of the 16s rRNA genome for amplification, choice of corresponding primer sets, and the subsequent bioinformatic processing.

Analysis of the 27F primer binding sites by Frank et al. may explain the differences between the 27F-I and 27F-II approaches. The commonly used 27F primer formulations, including the 27F-I primer set, do not cover several sequence variations involving contiguous phylogenetic clusters (Frank et al., 2008). The exclusion of such sequence variations explains the striking underrepresentation of several essential phylogenetic groups when the 27F-I primers or the "standards", e.g., from the AGP or Human Microbiome Project (HMP), are used. For particular taxa such as *Bifidobacterium* or *Bacteroides,* several base mismatches with the 27F-I forward primer were identified, consistent with the comparative underrepresentation of these genera in the samples analyzed with primer 27F-I (Frank et al., 2008; Matsuo et al., 2021). In contrast, the optimal coverage of the taxon *Faecalibacterium* by the 27F-I primer would explain its higher relative abundance compared to the 27F-II based analysis. Another potential, albeit unlikely, explanation for the reduced relative abundance of the taxon *Faecalibacterium* in the 27F-II primer compared to the 27F-I primer approach may involve dilution effects reported when using degenerate primers (Linhart and Shamir, 2005; Frank et al., 2008).

In addition to the more faithful representation of taxonomic abundancies, the higher degree of degeneracy of the 27F-II primer allows superior mapping of fecal microbiome diversity compared to 27F-I primers. This is a particular advantage when analyzing complex microbial samples, which have very high genetic diversity, requiring amplification of numerous unknown target sequences (Frank et al., 2008; Klindworth et al., 2013).

The disadvantage of using a degenerate primer is the greater potential for non-specific amplification or primer dimer formation, which can reduce PCR efficiency and accuracy (Dieffenbach et al., 1993). However, this problem can be effectively addressed by appropriate processing of sequenced reads, as the length of the expected amplicon can be accurately predicted and accounted for in the filter settings.

Conclusion

Recent advances in sequencing chemistry and base-calling algorithms have improved accuracy of ONT, which provides higher taxonomic resolution of full-length 16S rRNA gene sequencing compared to short-read sequencing (Shin et al., 2016; Johnson et al., 2019; Wei et al., 2020; Matsuo et al., 2021). It is highly likely that the time- and cost-efficient Nanopore platform will play an increasingly important role in the expanding field of microbiome research in the future. Our study provides a relevant comparative analysis of the performance of two different primer sets designed for full 16s rRNA genome sequencing of complex *in situ* samples. We demonstrate limitations of the universal 27F primer set (here referred to as 27F-I) for reliable detection of microbiome signatures in complex samples, such as human feces. In contrast, the more degenerate 27-F primer set (here referred to as 27F-II) uncovers microbial signatures much more faithfully and should be preferred for nanopore sequencing-based analyses of the human fecal microbiome.

The present study provides novel and important implications for both scientific and clinical applications when conducting microbial community analysis.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics committee, Medical Faculty, Philipps University Marburg, Germany. The patients/participants provided their written informed consent to participate in this study.

Author contributions

Conceptualization, CW and VR; Methodology, VR, MS, SP, LB; Software, LF, MW, DH; Validation, EW, RK, JP, UL; Formal analysis, CW, LF, MW, LB, VR; Investigation, CW, VR, JC; Resources, JM, JP, BS, TB; Data curation, CW, VR, LB; Writing–original draft preparation, CW, EW, VR, LF, MW; Writing–review and editing, BS, JP, TB, DH, BS; Visualization,

CW, LF, MW; Supervision, RK, SP, BS, EW, UL; Project administration, CW, VR; Funding acquisition, CW. All authors contributed to the article and approved the submitted version.

Funding

The present study was supported by a grant from the German Heart Foundation (Deutsche Herzstiftung, Frankfurt, Germany) and the P. E. Kempkes Foundation (Stiftung P. E. Kempkes, Marburg, Germany). CW also received support from the Clinician Scientist Program (SUCCESS) of the Medical Faculty of the Philipps University Marburg, Germany. Open Access funding provided by the Open Acess Publishing Fund of Philipps-Universität Marburg with support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

Acknowledgments

Parts of the study are included in the dissertation of LB. We thank Erik Wade for helpful comments on the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1213829/full#supplementary-material

References

Armougom, F. (2009). Exploring microbial diversity using 16S RRNA high-throughput methods. *J. Comput. Sci. Syst. Biol.* 2009. doi:10.4172/jcsb.1000019

Benítez-Páez, A., Portune, K. J., and Sanz, Y. (2016). Species-level resolution of 16S RRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *Gigascience* 5, 4. doi:10.1186/s13742-016-0111-z

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2013). GenBank. *Nucleic Acids Res.* 41, D36–D42. doi:10.1093/nar/gks1195

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2015). GenBank. *Nucleic Acids Res.* 43, D30–D35. doi:10.1093/nar/gku1216

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. D. A. D. A2 (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi:10.1038/nmeth.3869

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinform* 10, 421. doi:10.1186/1471-2105-10-421

Delahaye, C., and Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *Plos One* 16, e0257521. doi:10.1371/journal.pone.0257521

Dieffenbach, C. W., Lowe, T. M., and Dveksler, G. S. (1993). General concepts for PCR primer design. *Genome Res.* 3, S30–S37. doi:10.1101/gr.3.3.s30

Frank, J. A., Reich, C. I., Sharma, S., Weisbaum, J. S., Wilson, B. A., and Olsen, G. J. (2008). Critical evaluation of two primers commonly used for amplification of bacterial 16S RRNA genes. *Appl. Environ. Microb.* 74, 2461–2470. doi:10.1128/aem.02272-07

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. C. D-H. I. T. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi:10.1093/bioinformatics/bts565

Gomaa, E. Z. (2020). Human gut microbiota/microbiome in health and diseases: A Review. *Ant. Van Leeuwenhoek* 113, 2019–2040. doi:10.1007/s10482-020-01474-7

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi:10.1038/nature11234

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., et al. (2019). Evaluation of 16S RRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* 10, 5029. doi:10.1038/s41467-019-13036-1

Jones, S. (2013). Trends in microbiome research. *Nat. Biotechnol.* 31, 277. doi:10.1038/nbt.2546

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 41, e1. doi:10.1093/nar/gks808

Köster, J., and Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522. doi:10.1093/bioinformatics/bts480

Lee, S., Nguyen, L. T., Hayes, B. J., and Ross, E. M. (2021). Prowler: A novel trimming algorithm for Oxford nanopore sequence data. *Bioinformatics* 37, 3936–3937. doi:10.1093/bioinformatics/btab630

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: Human gut microbes associated with obesity. *Nature* 444, 1022–1023. doi:10.1038/4441022a

Linhart, C., and Shamir, R. (2005). The degenerate primer design problem: Theory and applications. *J. Comput. Biol.* 12, 431–456. doi:10.1089/cmb.2005.12.431

Liu, C., Cui, Y., Li, X., and Yao, M. (2020). Microeco: An R package for data mining in microbial community ecology. *Fems Microbiol. Ecol.* 97, fiaa255. doi:10.1093/femsec/fiaa255

Luo, J., Meng, Z., Xu, X., Wang, L., Zhao, K., Zhu, X., et al. (2022). Systematic benchmarking of nanopore Q20+ kit in SARS-CoV-2 whole genome sequencing. *Front. Microbiol.* 13, 973367. doi:10.3389/fmicb.2022.973367

Malla, M. A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., and Abd_Allah, E. F. (2019). Exploring the human microbiome: The potential future role of next-generation sequencing in disease diagnosis and treatment. *Front. Immunol.* 9, 2868. doi:10.3389/fimmu.2018.02868

Matsuo, Y., Komiya, S., Yasumizu, Y., Yasuoka, Y., Mizushima, K., Takagi, T., et al. (2021). Full-length 16S RRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. *Bmc Microbiol.* 21, 35. doi:10.1186/s12866-021-02094-5

McDonalda, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., et al. (2018). American gut: An open platform for citizen-science microbiome research. *Biorxiv*, 277970. doi:10.1101/277970

Nygaard, A. B., Tunsjø, H. S., Meisal, R., and Charnock, C. (2020). A preliminary study on the potential of nanopore MinION and Illumina MiSeq 16S RRNA gene sequencing to characterize building-dust microbiomes. *Sci. Rep-uk* 10, 3209. doi:10.1038/s41598-020-59771-0

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi:10.1093/nar/gkv1189

Ravi, R. K., Walton, K., and Khosroheidari, M. (2018). MiSeq: A next generation sequencing platform for genomic analysis. *Methods Mol. Biol.* 1706, 223–232. doi:10.1007/978-1-4939-7471-9_12

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). Vsearch: A versatile open source tool for metagenomics. *Peerj* 4, e2584. doi:10.7717/peerj.2584

Santos, A., Aerle, R. V., Barrientos, L., and Martinez-Urtaza, J. (2020). Computational methods for 16S metabarcoding studies using nanopore sequencing data. *Comput. Struct. Biotechnol. J.* 18, 296–305. doi:10.1016/j.csbj.2020.01.005

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., et al. (2021). Database Resources of the national center for Biotechnology information. *Nucleic Acids Res.* 50, D20–D26. doi:10.1093/nar/gkab1112

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi:10.1093/bioinformatics/btr026

Shah, N., Nute, M. G., Warnow, T., and Pop, M. (2018). Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics* 35, 1613–1614. doi:10.1093/bioinformatics/bty833

Shin, J., Lee, S., Go, M.-J., Lee, S. Y., Kim, S. C., Lee, C.-H., et al. (2016). Analysis of the mouse gut microbiome using full-length 16S RRNA amplicon sequencing. *Sci. Rep-uk* 6, 29681. doi:10.1038/srep29681

Stoler, N., and Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *Nar. Genom Bioinform* 3, lqab019. doi:10.1093/nargab/lqab019

Szoboszlay, M., Schramm, L., Pinzauti, D., Scerri, J., Sandionigi, A., and Biazzo, M. (2023). Nanopore is preferable over Illumina for 16S amplicon sequencing of the gut microbiota when species-level taxonomic classification, accurate estimation of richness, or focus on rare taxa is required. *Microorganisms* 11, 804. doi:10.3390/microorganisms11030804

Wang, Y., and Qian, P.-Y. (2009). Conservative fragments in bacterial 16S RRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *Plos One* 4, e7401. doi:10.1371/journal.pone.0007401

Wei, P.-L., Hung, C.-S., Kao, Y.-W., Lin, Y.-C., Lee, C.-Y., Chang, T.-H., et al. (2020). Characterization of fecal microbiota with clinical specimen using long-read and short-read sequencing platform. *Int. J. Mol. Sci.* 21, 7110. doi:10.3390/ijms21197110

Welzel, M., Lange, A., Heider, D., Schwarz, M., Freisleben, B., Jensen, M., et al. (2020). Natrix: A snakemake-based workflow for processing, clustering, and taxonomically assigning amplicon sequencing reads. *Bmc Bioinforma.* 21, 526. doi:10.1186/s12859-020-03852-4

Yang, B., Wang, Y., and Qian, P.-Y. (2016). Sensitivity and correlation of hypervariable regions in 16S RRNA genes in phylogenetic analysis. *Bmc Bioinforma.* 17, 135. doi:10.1186/s12859-016-0992-y

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 07 September 2023
DOI 10.3389/fgene.2023.1233657

OPEN ACCESS

EDITED BY
Margherita Mutarelli,
National Research Council (CNR), Italy

REVIEWED BY
Ze Zhang,
Dartmouth College, United States
Nguyen Quoc Khanh Le,
Taipei Medical University, Taiwan

*CORRESPONDENCE
Vijayachitra Modhukur,
modhukur@ut.ee

RECEIVED 02 June 2023
ACCEPTED 24 August 2023
PUBLISHED 07 September 2023

CITATION
Sharif Rahmani E, Lawarde A, Lingasamy P, Moreno SV, Salumets A and Modhukur V (2023), MBMethPred: a computational framework for the accurate classification of childhood medulloblastoma subgroups using data integration and AI-based approaches. *Front. Genet.* 14:1233657. doi: 10.3389/fgene.2023.1233657

MBMethPred: a computational framework for the accurate classification of childhood medulloblastoma subgroups using data integration and AI-based approaches

Edris Sharif Rahmani[1], Ankita Lawarde[1,2], Prakash Lingasamy[1], Sergio Vela Moreno[1,2], Andres Salumets[1,2,3] and Vijayachitra Modhukur[1,2]*

[1]Competence Centre on Health Technologies, Tartu, Estonia, [2]Department of Obstetrics and Gynecology, Institute of Clinical Medicine, University of Tartu, Tartu, Estonia, [3]Division of Obstetrics and Gynecology, Department of Clinical Science, Intervention and Technology, Karolinska Institute and Karolinska University Hospital, Stockholm, Sweden

Childhood medulloblastoma is a malignant form of brain tumor that is widely classified into four subgroups based on molecular and genetic characteristics. Accurate classification of these subgroups is crucial for appropriate treatment, monitoring plans, and targeted therapies. However, misclassification between groups 3 and 4 is common. To address this issue, an AI-based R package called MBMethPred was developed based on DNA methylation and gene expression profiles of 763 medulloblastoma samples to classify subgroups using machine learning and neural network models. The developed prediction models achieved a classification accuracy of over 96% for subgroup classification by using 399 CpGs as prediction biomarkers. We also assessed the prognostic relevance of prediction biomarkers using survival analysis. Furthermore, we identified subgroup-specific drivers of medulloblastoma using functional enrichment analysis, Shapley values, and gene network analysis. In particular, the genes involved in the nervous system development process have the potential to separate medulloblastoma subgroups with 99% accuracy. Notably, our analysis identified 16 genes that were specifically significant for subgroup classification, including *EP300*, *CXCR4*, *WNT4*, *ZIC4*, *MEIS1*, *SLC8A1*, *NFASC*, *ASCL2*, *KIF5C*, *SYNGAP1*, *SEMA4F*, *ROR1*, *DPYSL4*, *ARTN*, *RTN4RL1*, and *TLX2*. Our findings contribute to enhanced survival outcomes for patients with medulloblastoma. Continued research and validation efforts are needed to further refine and expand the utility of our approach in other cancer types, advancing personalized medicine in pediatric oncology.

KEYWORDS

childhood medulloblastoma, subgroup classification, DNA methylation, machine learning, gene expression, deep learning, Wnt, sonic hedgehog

1 Introduction

Medulloblastoma (MB) is the most prevalent malignant form of brain tumor among children, accounting for approximately 20% of all central nervous system (CNS) malignancies. The pathological features of MB are heterogeneous, and its emergence in the cerebellum is attributed to genetic and epigenetic alterations that disrupt critical pathways in cerebellar development (Northcott and Dubuc, 2012). According to the World Health Organization (WHO) classification of CNS tumors, the following four major subgroups have been identified based on molecular and genetic characteristics: wingless (WNT)-activated, sonic hedgehog (SHH)-activated, and numerically designated non-WNT/non-SHH, representing Groups 3 and 4 (Louis et al., 2016; Northcott et al., 2019; Louis et al., 2021). Accurate classification of childhood MB and its subclasses is critical for selecting appropriate treatment, monitoring plans, preventing tumor progression, and reducing mortality rates. In addition, the accurate classification of MB subgroups plays a vital role in developing targeted therapies for each specific subclass (Ramaswamy et al., 2016; Yan et al., 2020).

Advancements in multi-omics, including genomics, transcriptomics, epigenomics, and proteomics, have significantly contributed to the reporting of the biological and clinical relevance of subgroups in MB (Northcott and Dubuc, 2012; Northcott et al., 2017; Capper et al., 2018; Sharma et al., 2019). Transcriptomic analysis can identify medulloblastoma subgroups, but it has limitations in capturing the microenvironment and impact of modifications on gene expression, as well as dealing with technical variations, noisy data, and incomplete transcriptome coverage. DNA methylation profiling is more reliable in accurately classifying medulloblastoma subgroups (Korshunov et al., 2017; Gomez et al., 2018). Moreover, later studies use integrative clustering methods, such as similarity network fusion, to analyze multiple data types in conjunction for improved results. However, these methods may not account for intratumor heterogeneity, which can lead to misclassification of subgroups (Northcott and Shih, 2012; Cavalli et al., 2017; Northcott et al., 2017; Alharbi et al., 2020).

Recently, various other methods have been explored for the accurate classification of medulloblastoma subgroups, including an AI-based pipeline that uses histopathological and textural images (Attallah and Zaghlool, 2022), radiomics-based machine learning models (Karabacak et al., 2022), and one-class logistic regression machine learning that integrates gene expression and DNA methylation data (Lian et al., 2019). While featuring certain limitations, such as smaller sample sizes, limited diverse datasets, and the need for high-quality images, these methods hold great potential for improving the diagnosis and treatment of medulloblastoma. The current gold standard for accurate MB subgroup classification is genome-wide transcriptional and methylation arrays, with high accuracy for WNT and SHH subgroups (Ramaswamy et al., 2016). On the other hand, classification based on immunohistochemistry (IHC) and MRI has also been utilized for subgrouping. However, the challenges associated with standardization and lack of specificity in clinical settings have limited its effectiveness (Ramaswamy et al., 2016; Yan et al., 2020). The classification of Group 3 and Group 4 tumors is particularly challenging due to their overlapping molecular features, low incidence of recurring mutations, and recurrent chromosomal alterations (Cavalli et al., 2017). To overcome this issue, integration of multi-omics data (including DNA methylation, gene expression, and clinical features) and application of machine learning algorithms for the development of accurate classification models are required (Hovestadt et al., 2020). Therefore, our study aims to develop an artificial intelligence (AI)-based framework to classify MB subgroups using publicly available DNA methylation data. Furthermore, our framework integrates DNA methylation and gene expression data. The relevance of our prediction biomarkers was further examined using Gene Ontology analysis, survival analysis, Shapley values, and network analysis.

2 Materials and methods

2.1 Data collection

We collected DNA methylation profiles of pediatric medulloblastoma patients from multiple Gene Expression Omnibus (GEO) datasets, including GSE85212 (N = 763), GSE130051 (N = 1390), GSE90496 (N = 390), GSE54880 (N = 276), GSE109379 (N = 128), and GSE75153 (N = 91) (Table 1). All the above-mentioned methylation data were profiled using the Illumina Infinium HumanMethylation450 platform. In addition, we also included gene expression data that matched the DNA methylation data from the GEO series GSE85217 (N = 763) profiled using Affymetrix Human Gene 1.1 ST Array.

2.2 Methylation data preprocessing

We downloaded raw data files in "idat" format for all the aforementioned GEO datasets and assessed their quality using the minfi Bioconductor package (Aryee et al., 2014a). Subsequently, we conducted the following preprocessing procedure:

a) We assessed the signal quality using the detectionP function from the Bioconductor minfi package. We then calculated the p-values for each CpG probe across all samples. Probes with a *p*-value >0.05 in over 5% of samples were removed from subsequent analysis.
b) As all samples used in the current study were from the cerebellum, we used the preprocessQuantile function from the minfi package to normalize the data. We excluded CpG probes related to sex chromosomes and probes associated with single nucleotide polymorphisms (SNPs). On average, the total number of remaining probes was 420,000.
c) The methylation beta values ranging between 0 and 1 were calculated using the getBeta function from the Bioconductor minfi package. Briefly, such values were obtained based on the methylated and unmethylated probe intensities using formula M/(M + U + 100) (Bibikova et al., 2011); M and U stand for fully methylated and fully unmethylated intensities, respectively.
d) To deduce missing demographic information, including age and sex, we employed the methyAge algorithm and the predictedSex function from the Enmix (Xu et al., 2021) and minfi (Aryee et al., 2014) packages, respectively. This allowed us to create a

TABLE 1 Overview of datasets used in the current study from GEO Series: Testing, training, validation, and integration Dataset. Age and sex were predicted for datasets with missing metadata information.

Dataset	GEO accession	Total samples	Age (years) (mean ± SD)	Gender (% male)	Country	References
Training/ Testing	GSE85212[a]	763	10.43 ± 9.43	65.65	Canada	Cavalli et al. (2017)
Integration	GSE85217, GSE85212	763	10.43 ± 9.43	65.65	Canada	Cavalli et al. (2017)
Validation	GSE130051	1390	5.78 ± 10.53	66.14	Europe, North America and Asia-Pacific	Sharma et al. (2019)
Validation	GSE90496	390	36.15 ± 6.27	60.26	Germany	Capper et al. (2018)
Validation	GSE54880[a]	276	8.27 ± 4.75	63.04	Germany	Hovestadt et al. (2013)
Validation	GSE109379	128	36.75 ± 6.84	60.47	Germany	Capper et al. (2018)
Validation	GSE75153	91	11.5 ± 18.39	59.78	Canada	-

[a]Series with original metadata.

summarized demographic view of the data types used in the current study.

2.3 Integration of DNA methylation and gene expression data using similarity network fusion (SNF)

In our study, we utilized the similarity network fusion (SNF) technique (Wang et al., 2014) proposed by Wang *et al.* to integrate the DNA methylation dataset with gene expression data and to further generate new labels. SNF allows for the identification of similarity networks, enabling the creation of the most appropriate labels for the methylation dataset using spectral clustering. To this end, we combined 763 samples from the methylation dataset (GSE85212) with the same number of samples from the gene expression dataset (GSE85217). The data integration was performed using the following parameters: 51 nearest neighbors, sigma = 0.85, and 120 iterations. As our study focused on medulloblastoma, which is characterized by the four subgroups, we set the cluster number to four and used the result of spectral clustering as the ground truth labels. We converted the cluster numbers into subgroups by comparing the sample number from the fused dataset and actual labels. Next, we evaluated the performance of the fused network by calculating the normalized mutual information (NMI) score, ranging from 0 to 1. An NMI score of 1 indicates that the fused network leads to the same labels as the actual labels, while a score of 0 indicates the opposite.

2.4 Feature selection

Feature selection is a critical step in machine learning, as it allows for the identification of the most relevant features, resulting in decreased prediction model error rates and computational time. In this study, we utilized a random forest model (RF) to train the top 5,000 most variable CpG probes obtained from Median Absolute Deviation (MAD) through the mad function in the stats package. To this end, we grew 300 trees using the RF model and determined the importance of each probe across all subgroups using the varImp function from the caret package.

2.5 Survival analysis

To evaluate the prognostic potential of prediction biomarkers, we conducted an overall survival analysis by adapting the MethSurv webtool pipeline (Modhukur et al., 2018; Modhukur, 2019). We utilized a multivariate Cox proportional hazards model to associate the methylation levels of each biomarker with patient survival using age, sex and MB subgroups as covariates. Patients were divided into high and low methylation groups based on a cut-off point such as the mean, median, or upper and lower quantiles. The specific cut-off values were determined based on models with high hazard ratios (HRs), maximizing the difference in survival outcomes between the groups. Next, we evaluated the goodness of fit of the Cox model using both the likelihood-ratio (LR) test and the Wald test.

2.6 Class imbalance correction

To overcome the challenge posed by imbalanced sample sizes for each MB subgroup in the methylome data, we implemented a technique called synthetic minority oversampling (SMOTE) (Chawla et al., 2002) using the DMwR package (Torgo, 2016). SMOTE generates synthetic samples by interpolating between existing minority class samples.

2.7 Data clustering

We utilized t-distributed stochastic neighbor embedding (t-SNE), a non-linear dimensionality reduction technique using the Rtsne package (Van Der Maaten and Hinton, 2008), to reduce the high-dimensional space to the most informative variables. The resulting cluster labels from the previous spectral

clustering step were applied to identify four subgroups in our dataset, which were visualized in a three-dimensional (3D) plot using the rgl package (Adler et al., 2003). To explore the distribution of beta values, we used the ComplexHeatmap R package (Gu et al., 2016) to generate heatmaps.

2.8 AI-based models to classify MB subgroups

Our aim was to address the multiclassification challenge of accurately classifying medulloblastoma (MB) subgroups by leveraging the DNA methylation levels as a key feature. To do this, we used a diverse set of machine learning algorithms. The six algorithms employed were random forest (RF), naive Bayes (NB), K-nearest neighbor (KNN), support vector machine (SVM), extreme gradient boosting (XGB), and linear discriminant analysis (LDA). Furthermore, to capture the intricate nonlinear relationships, we incorporated an artificial neural network (ANN) model. Since the ensemble-based algorithms RF and XGB combine the predictions of multiple weak models to improve overall performance, we included those models in our study. On the other hand, NB operates as a probabilistic model, employing Bayes' theorem to calculate the likelihood of class membership based on the independent features. KNN is classified as a nonparametric supervised learning algorithm, meaning that it does not make explicit assumptions about the underlying data distribution and defers computations until prediction. SVM can function either as a linear or as a nonlinear model, using a hyperplane or kernel trick to separate classes in the feature space. LDA is a linear model that projects data onto a lower-dimensional space to maximize class separation, aiding classification (Ray, 2019). The utilization of diverse machine learning algorithms in this classification conundrum enables a comprehensive evaluation of their efficacies, fostering heightened precision and resilience of the classification model. Additionally, ensemble methods (RF and XGB) can reduce variance and bias, while linear models (SVM and LDA) provide interpretability of the results (Sheth et al., 2022). Moreover, the ANN model is well known for its capability to learn complex nonlinear relationships between features. Unlike linear models, ANNs consist of interconnected nodes or neurons organized in layers, enabling them to capture intricate patterns and interactions in the data (Grossi and Buscema, 2007).

To train the abovementioned machine learning prediction models, we split the data into the training and test sets with a ratio of 0.8 for machine learning models using the sample. split function from the caTools package. Furthermore, we performed cross-validation in ten random folds (k = 10) using the createFolds function from the caret package (Kuhn, 2008).

The RF model was trained using the Random Forest package (Liaw and Wiener, 2002) with 300 trees and six as the maximum number of nodes. The SVM and NB models were trained using the e1071 package (Meyer, 2014), and a threshold of 0.8 was defined for NB to convert probabilities into subgroups. The KNN model was trained using the class package (Venables and Ripley, 2013) with three nearest neighbors, and the LDA model was trained using the lda function from the MASS package.

We implemented ANN models using the Keras package in R with TensorFlow 2.10 (Abadi et al., 2016). The data were split into training, testing, and validation sets with ratios of 0.6, 0.2, and 0.2, respectively. The ANN model had four layers: input, two hidden layers, and output, with neuron counts of 40, 30, 10, and 4. 'Leaky ReLU' activation was used for the first three layers, and softmax was used for the output layer.

To prevent overfitting, we applied regularization techniques, including dropout (50%, 40%, and 10% rates), L2 regularization on the second layer (regularizer_l2 = 0.009), and early stopping after five patients. The model was optimized using the categorical cross-entropy loss, stochastic gradient descent (SGD) optimizer, 200 epochs, batch size of 16, learning rate of 0.03, decay of 0.00006, momentum of 0.05, and Nesterov momentum.

To optimize the computational training time, we utilized the mclapply function from the parallel package to run the machine learning models in parallel on available CPUs. The training was performed on an Ubuntu machine equipped with an Intel Core i5-6200U processor and 16 GB RAM.

2.9 Performance evaluation

In our study, we evaluated the performance of each classification model using standard metrics, which included accuracy, sensitivity, specificity, precision, F1-score, and area under the curve (AUC) as described by similar studies (Le et al., 2017; Le et al., 2022). Briefly, the performance metrics were computed as follows:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$
$$Sensitivity = TP/(TP + FN)$$
$$Precision = TP/(TP + FP)$$
$$F1score = (2 \times TP)/(2 \times TP + FP + FN)$$

Here, true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) indicate whether the model predicted correctly or incorrectly. We also computed the AUC from the pROC package (Robin et al., 2011). The AUC score, presents the degree of separability between the classes.

2.10 Model visualization

To plot the training and testing results of a classifier, we designed a custom R script. Initially, the dataset was partitioned into a training set and a test set. Subsequently, principal component analysis (PCA) was conducted on the training and testing sets separately using the preProcess function from the caret package, enabling the extraction of two primary components that captured the most significant variability in the data. Following this, the training and test sets were transformed using the derived PCA outcomes. A grid structure was then constructed, encompassing values pertaining to the two principal components. Utilizing the trained classifier, labels are predicted for the grid set. Moreover, a color mapping scheme was employed to associate colors with the predicted and actual subgroups, enhancing the interpretability of the resulting plot.

2.11 Gene set enrichment analysis

To investigate the molecular function of the predicted CpG biomarkers and their relevance to the MB subgroups, we performed gene enrichment analysis. To annotate the CpGs with the genes, we utilized the minfi and IlluminaHumanMethylation450kanno.ilmn12.hg19 packages. The resulting genes were used as the input for the gprofiler2 package (Kolberg et al., 2020) to identify their gene ontology (GO) terms in the biological process (BP), KEGG, and Reactome pathways. To determine statistical significance, we used the false discovery rate (FDR) with a threshold of p-value <0.05.

2.12 Explaining the effect of each feature on the model output

To interpret the contribution of each identified biomarker to the MB subgroup prediction, we used the Shapley value, which is a local interpretation method in IML (Interpretable Machine Learning). Since the machine learning models employed in this study cannot directly elucidate the relationship between CpG probes and their target class, we employed the Shapley value to provide human-understandable explanations of the models' results. The Shapley value is computed as the average marginal contribution of a CpG probe or gene beta value across all possible coalitions. For a single prediction of each MB subgroup, it randomly changed the value of each beta value from zero to the actual value of the sample and calculated the prediction for all patterns of changes due to the addition of each CpG. We used the iml package (Molnar, 2018) to calculate the Shapley values. To perform the Shapley analysis, we first trained an ANN model with all converted gene symbols from the functional enrichment step and the respective parameters as described in the iml package. Following the prediction on the training set, we used the prediction variable as input to the Shapley function to explain four samples of the training set belonging to each subgroup.

2.13 Network analysis

In this study, we utilized the igraph package (Csardi and Nepusz, 2006) to perform gene network analysis and investigate the relationship between the predicted genes. To identify clusters of genes that are highly correlated, we computed the Pearson correlation coefficient between each pair of genes and generated an adjacency matrix. We filtered out any edges that formed loops or had multiple connections, as well as edges with a Pearson correlation value less than or equal to 0.6 or genes with fewer than two adjacent edges. Additionally, we scaled the size of each gene according to its methylation values by a factor of 10 to enhance the readability of the network. We then utilized Prim's algorithm to convert the graph adjacency object into a minimum spanning tree. Finally, we identified highly correlated gene clusters using a function called cluster_edge_betweenness.

3 Results

In this study, we used a combination of data integration and AI-based techniques to effectively classify subgroups of medulloblastoma. The methodology used in this study is presented in Figure 1 and involves the following six main steps:

(i) Collection of data from Gene Expression Omnibus (GEO), followed by pre-processing and processing steps;
(ii) Implementation of similarity network fusion (SNF) to establish new class labels by integrating DNA methylation and gene expression data;
(iii) Median Absolute Deviation (MAD) analysis was applied to select informative prediction biomarkers, followed by random forest (RF) analysis for feature selection. Furthermore, survival analysis was performed based on the prediction biomarkers.
(iv) Construction of AI-based prediction models following Synthetic Minority Oversampling Technique (SMOTE) application;
(v) Evaluation of the models using multiple parameters, including accuracy, sensitivity, precision, AUC, and F1-score;
(vi) Gene Ontology (GO) analysis was used to functionally annotate the selected genes.

We further conducted gene network analysis and interpreted the classifier decision by utilizing Shapley values. The subsequent sections provide detailed results from each of the steps mentioned above.

3.1 Integration of gene expression and methylation data through similarity network fusion

In this study, using similarity network fusion (SNF), we identified four distinct clusters in both the gene expression and methylation datasets (Supplementary Figures S1A, B). We then fused the resulting networks to obtain a comprehensive view of the data (Supplementary Figures S1C). The spectral clustering results on the fused network revealed two clusters (belonging to groups 3 and 4; Supplementary Figures S1C) with slightly different samples from the actual clusters (GSE85212) with a high NMI score of 0.926. Using the class labels obtained from SNF and implementing SMOTE, we addressed class imbalance, particularly in the minority subgroup (WNT = 70), by increasing the number of WNT samples to 210, resulting in a total of 910 samples (Supplementary Figures S2). Additionally, for the selection of the top 399 probes as features for prediction, we employed the random forest feature selection method among the 5,000 most variable probes identified using the median absolute deviation (MAD) method. This two-step process allowed us to first identify the 5,000 most variable probes based on MAD and then further reduce them to the top 399 probes using random forest feature selection (Supplementary Data S1).

The t-SNE visualization revealed (Figure 2A) only a minor overlap between groups 3 and 4; additionally, only one sample from the WNT cluster appeared in the SHH subgroup. Furthermore, we generated a heatmap of the CpG biomarkers to examine the distribution of methylation beta values across all subgroups (Figure 2B), in which a distinct methylation pattern among subgroups is notable.

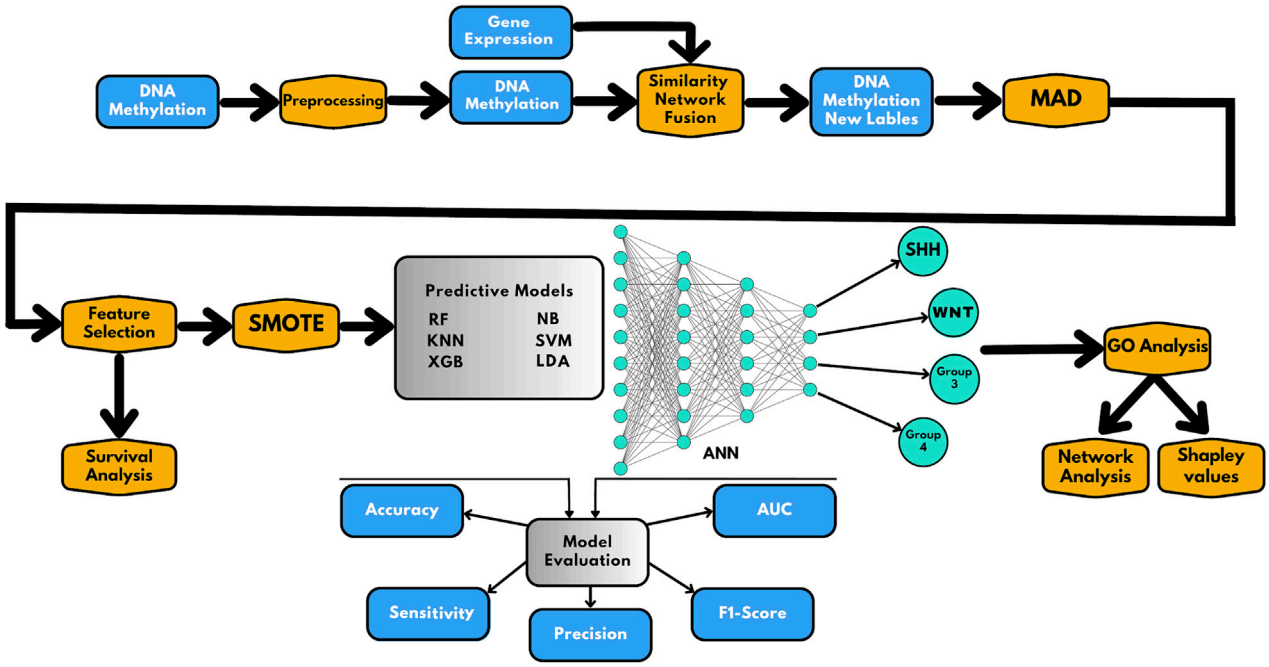


FIGURE 1
Schematic representation of the workflow presented in this study. The workflow includes the following steps: (1) preprocessing and integration of methylome and gene expression data; (2) similarity network fusion (SNF) to establish the new class labels; (3) MAD analysis for selecting informative prediction biomarkers and feature selection and survival analysis; (4) applying SMOTE for correction of class imbalance and construction of AI-based prediction models; (5) evaluation of the prediction models; and (6) Gene Ontology (GO) analysis for functional annotation of the prediction genes, where gene network analysis and Shapley values are used to understand the classifier results.

3.2 Performance evaluation of the prediction models for medulloblastoma subgroup classification based on DNA methylation profiles

In our study, we employed six robust machine-learning algorithms, namely, SVM, KNN, NB, RF, XGB, and LDA, along with an artificial neural network, to predict medulloblastoma subgroups based on DNA methylation samples using 399 predictive biomarkers. As a result of the fusion process, a subset of samples (n = 16) had their labels switched (Supplementary Table S1). These new labels predominantly belonged to the Group 3 and Group 4 subgroups, accounting for 14 out of the 16 samples. These switched labels were utilized specifically for training the model. However, during the validation process, the confusion matrices were constructed based on the original labels from validation sets and predicted labels. For testing and training, we utilized the dataset from GSE85212, while multiple datasets were used for validation. Detailed information regarding the testing/training and validation datasets can be found in Table 1.

The overall performance of the classifiers based on the validation set (GSE90496) is presented in Table 2. Briefly, the ANN model achieved the highest accuracy of 99.25%, followed by SVM with 99.50% accuracy. However, the KNN, NB, RF, XGB, and LDA models also achieved high accuracy ranging from 97.80% to 99.35%.

Since the focus of our study was the classification of MB subgroups, we evaluated the performance of each model, considering the different MB subgroups, across multiple validation datasets. Notably, all tested classifiers exhibited exceptional performance on the GSE90496 validation set, exceeding 0.92 in accuracy, precision, sensitivity, F1-Score, specificity, and AUC (Table 3; Figure 3A; Supplementary Figure S3). We specifically monitored the performance of the prediction models on the challenging Group 3 and Group 4 MB subgroups. The SVM, RF, and ANN models achieved excellent performance, with accuracy, precision, sensitivity, F1-Score, specificity, and AUC exceeding 0.96 (Table 3; Supplementary Table S2). Other models, including KNN, NB, LDA, and XGB, also demonstrated comparable performance, with accuracy, precision, sensitivity, F1-Score, specificity, and AUC ranging from 0.88 to 0.99 (Table 3; Supplementary Table S2; Supplementary Figure S3).

Furthermore, we visualized the ability of the classifiers based on the training and test sets, as shown in Figures 3B,C, using Principal Component Analysis (PCA) based on XGB as the reference model. The PCA plot revealed a clear separation between MB subgroups. Thus, the classifiers successfully captured the underlying variability and discriminating features among the different MB subgroups.

Across the different validation sets, our models consistently displayed higher performance. For example, on the GSE130051 dataset, the NB model emerged as a top-performing classifier with accuracy exceeding 0.96, while other models achieved accuracy ranging from 0.91 to 0.95 (Supplementary Tables S3, S4). The ANN model demonstrated robust performance on the GSE54880 dataset, achieving an accuracy of 0.97 with minimal misclassifications (Supplementary Tables S5, S6). On the GSE109379 dataset, the ANN and RF models performed

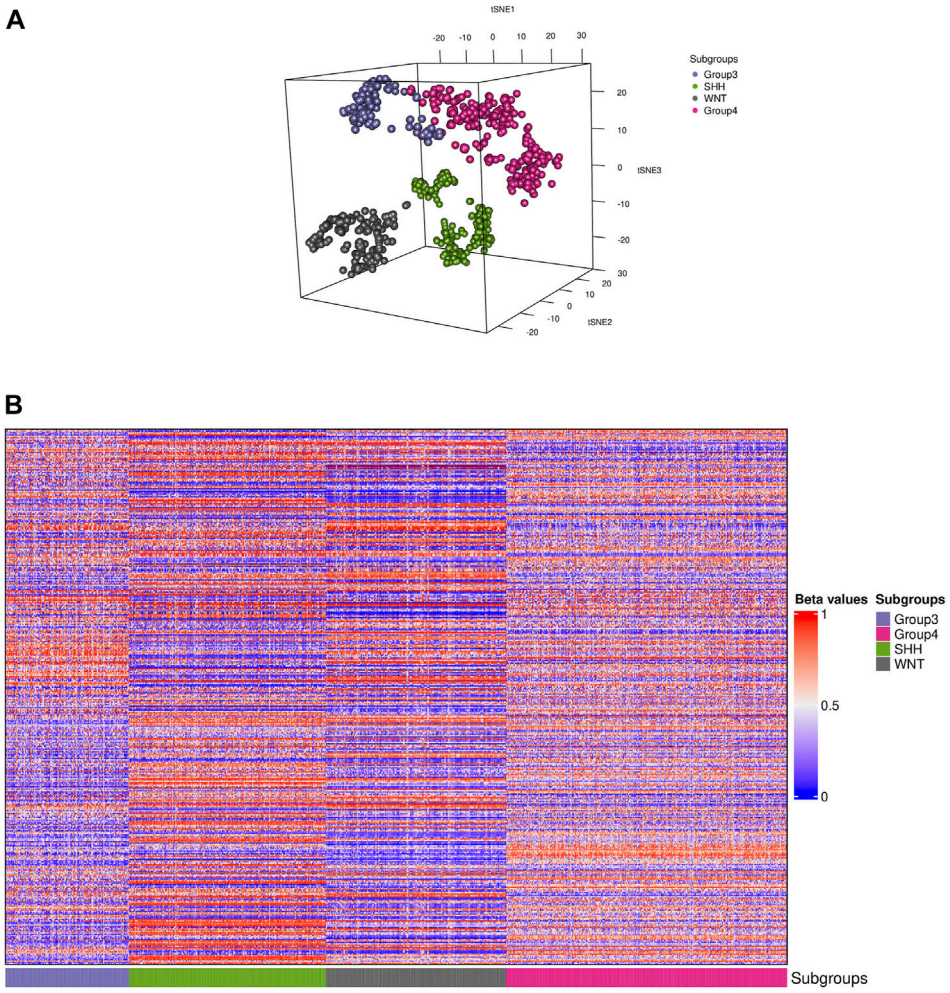


FIGURE 2
Visualization of the training data. **(A)** Distributed stochastic neighbor embedding (t-SNE) plot shows the presence of four distinct subgroups (colored dots) of medulloblastoma in the dataset. **(B)** A heatmap representation of the 910 samples depicting each subgroup is shown. The colors in the heatmap represent the levels of DNA methylation, with red indicating higher methylation levels and blue indicating lower methylation levels. The CpG biomarkers revealed a unique methylation pattern in groups 3 and 4, while the WNT and SHH subgroups displayed a distinct pattern.

TABLE 2 Overall performance metrics for each model using GSE90496 as a validation set.

Model	Accuracy	Precision	Sensitivity	F1.Score	Specificity	AUC
RF	0.9935 ± 0.005	0.98675 ± 0.013	0.988 ± 0.01	0.9875 ± 0.011	0.9955 ± 0.004	0.98 ± 0
SVM	0.995 ± 0.004	0.98875 ± 0.013	0.99125 ± 0.008	0.98975 ± 0.009	0.9965 ± 0.003	0.986 ± 0
XGB	0.9895 ± 0.005	0.979 ± 0.02	0.96875 ± 0.023	0.97325 ± 0.014	0.993 ± 0.005	0.973 ± 0
NB	0.9935 ± 0.005	0.98575 ± 0.017	0.9895 ± 0.01	0.9875 ± 0.011	0.99575 ± 0.004	0.983 ± 0
LDA	0.978 ± 0.017	0.95875 ± 0.032	0.95625 ± 0.04	0.9575 ± 0.036	0.9845 ± 0.012	0.928 ± 0
KNN	0.9885 ± 0.009	0.97775 ± 0.019	0.9765 ± 0.022	0.97725 ± 0.02	0.99175 ± 0.007	0.961 ± 0
ANN	0.9925 ± 0.078	0.98475 ± 0.17	0.98475 ± 0.17	0.98475 ± 0.17	0.9945 ± 0.058	0.995 ± 0

exceptionally well, achieving accuracy above 0.97, while the SVM, XGBoost, and KNN models also exhibited favorable performance, albeit with slightly lower precision and sensitivity for Groups 3 and 4 (Supplementary Tables S7, S8). Finally, for the GSE75153 dataset, all models performed comparably well, with accuracy above 0.97 (Supplementary Tables S9, S10).

TABLE 3 Performance metrics of each model for MB subgroup classification using GSE90496 as a validation set.

Subgroup	Accuracy	Precision	Sensitivity	F1-score	Specificity	AUC	Model
Group3	0.987	0.962	0.974	0.968	0.99	0.98	RF
Group4	0.987	0.985	0.978	0.982	0.992	0.98	
SHH	1	1	1	1	1	0.98	
WNT	1	1	1	1	1	0.98	
Group3	0.99	0.962	0.987	0.974	0.99	0.986	SVM
Group4	0.99	0.993	0.978	0.985	0.996	0.986	
SHH	1	1	1	1	1	0.986	
WNT	1	1	1	1	1	0.986	
Group3	0.982	0.938	0.974	0.955	0.984	0.973	XGB
Group4	0.987	0.985	0.978	0.982	0.992	0.973	
SHH	0.997	0.993	1	0.996	0.996	0.973	
WNT	0.992	1	0.923	0.96	1	0.973	
Group3	0.987	0.95	0.987	0.968	0.987	0.983	NB
Group4	0.987	0.993	0.971	0.982	0.996	0.983	
SHH	1	1	1	1	1	0.983	
WNT	1	1	1	1	1	0.983	
Group3	0.959	0.907	0.883	0.895	0.978	0.928	LDA
Group4	0.956	0.935	0.942	0.939	0.964	0.928	
SHH	0.997	0.993	1	0.996	0.996	0.928	
WNT	1	1	1	1	1	0.928	
Group3	0.977	0.947	0.935	0.941	0.987	0.961	KNN
Group4	0.977	0.964	0.971	0.968	0.98	0.961	
SHH	1	1	1	1	1	0.961	
WNT	1	1	1	1	1	0.961	
Group3	0.985	0.961	0.961	0.961	0.99	0.995	ANN
Group4	0.985	0.978	0.978	0.978	0.988	0.995	
SHH	1	1	1	1	1	0.995	
WNT	1	1	1	1	1	0.995	

In summary, our analysis revealed slight variability in the performance of different prediction models across a diverse range of validation sets, with an average accuracy exceeding 0.96.

3.3 Biological and clinical significance of the prediction biomarkers

We performed an overall survival analysis on 399 prediction biomarkers after adjusting for the covariates age, sex and sugroups using the methodology adapted from MethSurv (Modhukur et al., 2018; Modhukur, 2019). We found that all 399 prediction biomarkers showed a significant association with patient survival (log rank test p-value <0.05). The top biomarkers with the lowest p values included *CBFA2T3*, *PRDM16*, *TRIM65*, *KIAA0182*, *SEMA4F*, *OR6N1*, *RPTOR*, *KIAA0415*, *SAG*, and *TTC15* (Figure 4; Supplementary Figure S4, and Supplementary Data S2).

To further gain biological insights into the prediction biomarkers, we performed functional enrichment analysis. We annotated each probe with its gene symbol and excluded CpGs without gene annotations. For CpGs with duplicated gene names, we calculated the median value. The latter resulted in a total of 239 unique gene symbols, which were used as input for gprofiler2 (Peterson et al., 2020). Our analysis identified the 20 most significant biological processes (adjusted p-value <0.05) in which the selected genes were enriched (Supplementary Figure S5). Some of these biological processes included nervous system development, neurogenesis, neuron projection development, and

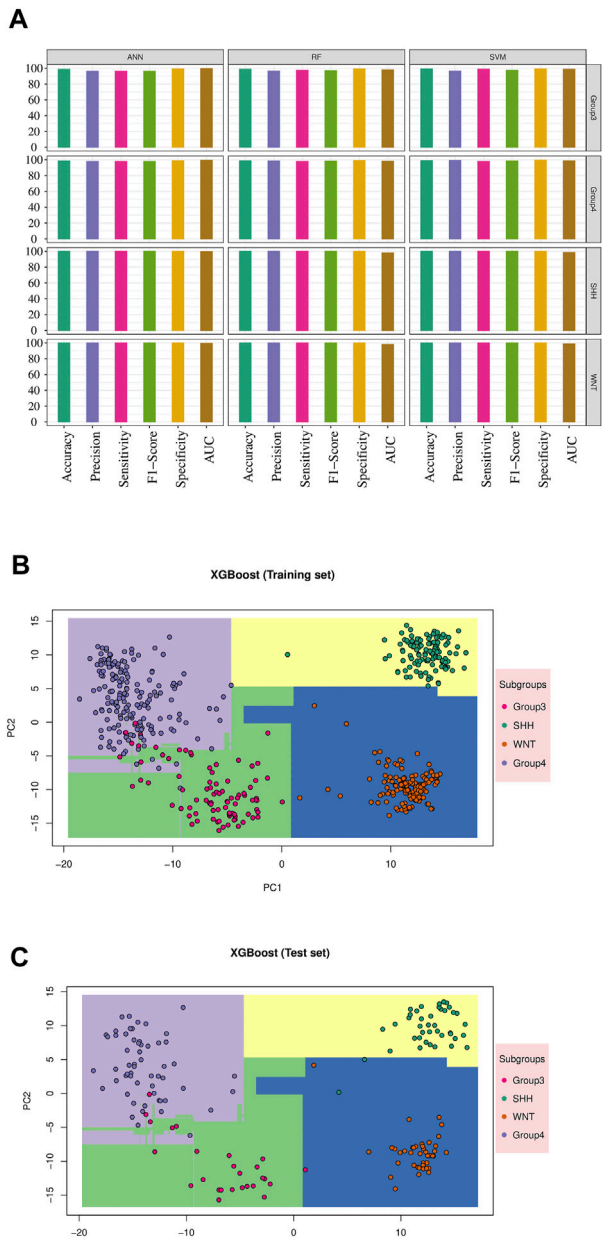


FIGURE 3
Machine learning model's performance in predicting medulloblastoma subgroups: The result for validation cohort GSE90496 and XGBoost model results. **(A)** The performance of the top three models' on the validation cohort GSE90496 (n = 390) is displayed (see Supplementary Figure S3 for other models). The *X*-axis represents various metrics, including accuracy, precision, sensitivity, F1-Score, specificity, and AUC. Each metric is represented by a bar plot using different colors, indicating the corresponding percentage. These metrics were calculated separately for each subgroup, highlighting the accurate classification of WNT and SHH subgroups, as well as some misclassifications within Groups 3 and 4. **(B)** The training result for the XGBoost model is shown. The distinct decision boundaries for each subgroup are denoted by various colors. **(C)** The testing result for the XGBoost model performance is displayed. Similar to the training results, distinct decision boundaries are depicted for each subgroup using different colors.

differentiation. To evaluate the effectiveness of the enriched genes, we employed a neural network as our optimal model to analyze genes associated with the top 20 biological processes. The neural network consisted of five layers with 50, 30, 20, 10, and 4 neurons and a learning rate of 0.03. We trained each gene set ten times and computed the average performance results. Although all models produced similar outcomes with AUC scores above 0.9, the nervous system development process consisting of 49 genes had the highest mean AUC score of 0.995 (Supplementary Figure S6; Supplementary Table S11).

3.4 Explaining feature effects on model output through Shapley values

To investigate the individual impact of the prediction genes (N = 49) on the model performance, we computed the Shapley values for the trained neural network model. Each gene with its corresponding beta values and their contribution in terms of Shapley values on the ANN model across different subgroups are shown in Figure 5. Briefly, maroon color indicates a positive effect, and blue denotes an adverse effect.

For example, we found that *ZIC4*'s hypermethylation state (beta value = 0.882) has a highly positive impact on the model's ability to predict Group 3 but has a negative effect on the WNT subgroup. At the same time, *ZIC4* has a low negative impact on the model's ability to forecast SHH and Group 4 subgroups. Additionally, we identified other genes, such as *ARTN* and *SLC8A1*, which have a positive contribution to the model's ability to predict Group 3, with beta values equal to 0.847 and 0.327, respectively.

Furthermore, we observed that higher methylation levels of the *CXCR4* and *MEIS1* genes and lower methylation levels of *NFASC* had a positive impact on the ANN model's ability to predict the Group 4 subgroup. In the WNT subgroup, *ASCL2*, *SYNGAP1*, *RTNR4L*, and *NFASC* gene hypermethylation status, as well as *KIDINS220* and *S100A10* gene hypomethylation, had a highly positive impact on prediction. For the SHH subgroup, we found that higher methylation levels of *SLC8A1* and lower methylation levels of *ROR1*, *CXCR4*, and *RTN4RL1* had a high contribution to the prediction.

3.5 Network analysis

We conducted network analysis using the methylation beta values of 49 genes enriched in the nervous system development process identified based on the functional enrichment analysis (Supplementary Figure S5; Supplementary Table S11). The resulting network revealed 41 genes with a Pearson correlation coefficient greater than 0.6, distributed among six distinct clusters (Figure 6A). To evaluate the classification ability of each cluster's genes, we trained artificial neural network (ANN) models for each cluster. However, upon assessing the performance of the individual models on the test data (Figure 7A), we observed that some models

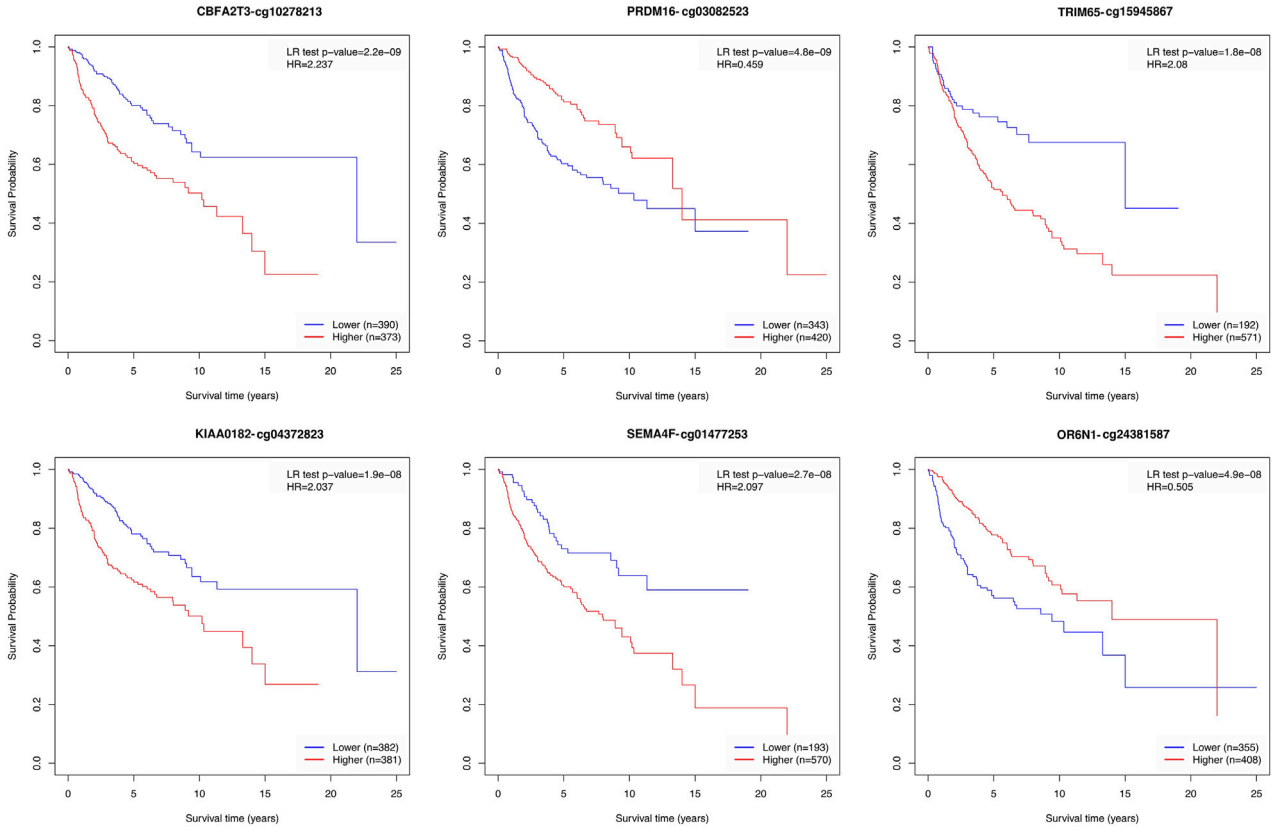


FIGURE 4
Kaplan–Meier plots depicting the effect of the top six prediction biomarkers (log-rank test <0.05). Methylation groups are dichotomized by higher and lower methylation groups based on a cut-off point such as the mean, median, or upper and lower quantiles. The *X*-axis denotes survival time in years, and the *Y*-axis denotes the probability of patient survival.

exhibited poor performance for certain subgroups. To address this limitation, we devised a unique strategy to enhance the model's performance. Specifically, we incorporated genes from other clusters into each model until we achieved improved performance (Figure 7B). This iterative process allowed us to leverage the collective predictive power of multiple gene clusters, ultimately leading to enhanced classification accuracy. The performance of each model on the test data is shown in Figure 7A, where all models except for cluster 3 exhibited poor performance. To improve the model's performance, we gradually added genes from other clusters to each model until the performance improved (Figure 7B). Accordingly, we confirmed the significance of *ARTN* and *WNT4* for Group 3 and WNT subgroups, respectively. These genes suggest possible associations with their respective subgroups, highlighting their importance in driving molecular characteristics and prognostic outcomes. Building upon these findings, we integrated *ARTN*, *WNT4*, *EP300* and *ROR1* into the gene list of cluster 4, resulting in improved performance for cluster 1.

Furthermore, by adapting a similar procedure, we intended to improve cluster 5, which initially exhibited the lowest performance. To achieve this, we incorporated the additional genes *RTN4RL1*, *TLX2*, *ARTN*, *WNT4*, *EP300*, and *ROR1* into the existing gene list from cluster 5. Additionally, cluster 6 was improved by using the same gene list as cluster 5. However, for cluster 3, we included *SEMA4F*, *SLC8A1*, *CXCR4*, *SYNGAP1*, *NFASC*, and *MEIS1* in the existing list of significant genes, thereby improving its predictive power.

Figure 6B displays the beta values associated with the predicted prognostic genes. Furthermore, Table 4 provides a comprehensive list of these significant genes, highlighting their functional annotations and their relevance to each molecular subgroup.

4 Discussion

Accurate classification of molecular subgroups in medulloblastoma (MB) is vital for initiating appropriate treatment plans. In our study, we utilized a comprehensive approach integrating data and AI-based methods and utilized synthetic sample generation using SMOTE to address limited data and maintain class balance. Our developed prediction framework, MBMethPred, was designed explicitly for medulloblastoma subgroup classification using DNA methylation data. MBMethpred incorporates multiple AI models to enhance accuracy, processing speed, ease of use, and user-friendliness.

Compared to the molecular-based MB subgroup classification methods (Schwalbe et al., 2013; 2017; Korshunov et al., 2017; Capper

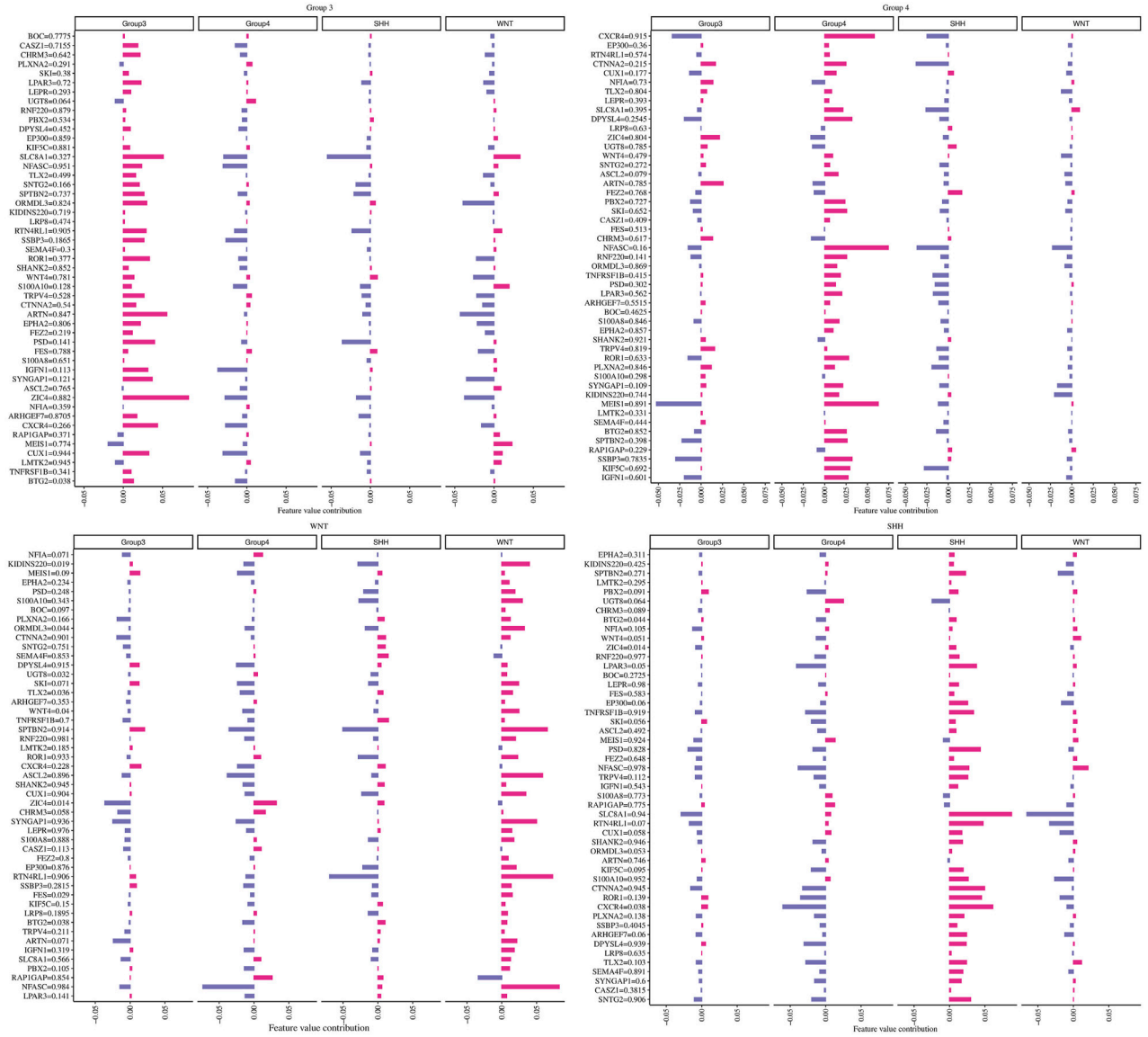


FIGURE 5
The contribution of each gene in predicting every MB subgroup. Each plot title corresponds to an associated group where we computed Shapley values. The *Y*-axis represents genes with their beta values, while the *X*-axis demonstrates the genes' contribution to the ANN model based on the phi value.

et al., 2018; Gomez et al., 2018; Korshunov et al., 2019; Sharma et al., 2019; Rathi et al., 2020) (Supplementary Table S12), MBMethPred demonstrates several distinctive characteristics and advantages. Previous studies (Schwalbe et al., 2013; Korshunov et al., 2017; Schwalbe et al., 2017; Capper et al., 2018; Gomez et al., 2018; Korshunov et al., 2019; Sharma et al., 2019) employed a single classifier, in contrast to MBMethPred, which applies multiple classifiers. While MBMethPred achieves an AUC score above 0.99, the primary focus of Capper *et al.*'s (2018) study was the classification of central nervous system tumors, rather than focusing on medulloblastoma. Furthermore, it lacked an accuracy score specifically for medulloblastoma. Similarly, Sharma *et al.* exclusively concentrated on the classification of Groups 3 and 4 subgroups. Additionally, both Korshunov et al., 2017 and Korshunov et al. (2019) utilized smaller sample sizes (N = 239 and N = 78, respectively) compared to MBMethPred's sample size of 910 samples. Likewise, Korshunov et al. (2019) solely focused on classifying the WNT subgroup. Moreover, Rathi et al. (2020) and Gomez et al. (2018) reported accuracies ranging between 85% and 100% using a single classifier, which is lower than the accuracy achieved by MBMethPred with multiple classifiers. In contrast, Attallah and Zaghlool (2022) utilized histopathology images and achieved 100% accuracy (Attallah and Zaghlool, 2022). However, there is limited availability of histopathological images and a lack of precision (Kim et al., 2022). This approach may restrict its widespread applicability. In this context, MBMethPred remains an accessible and valuable alternative for medulloblastoma subgroup classification

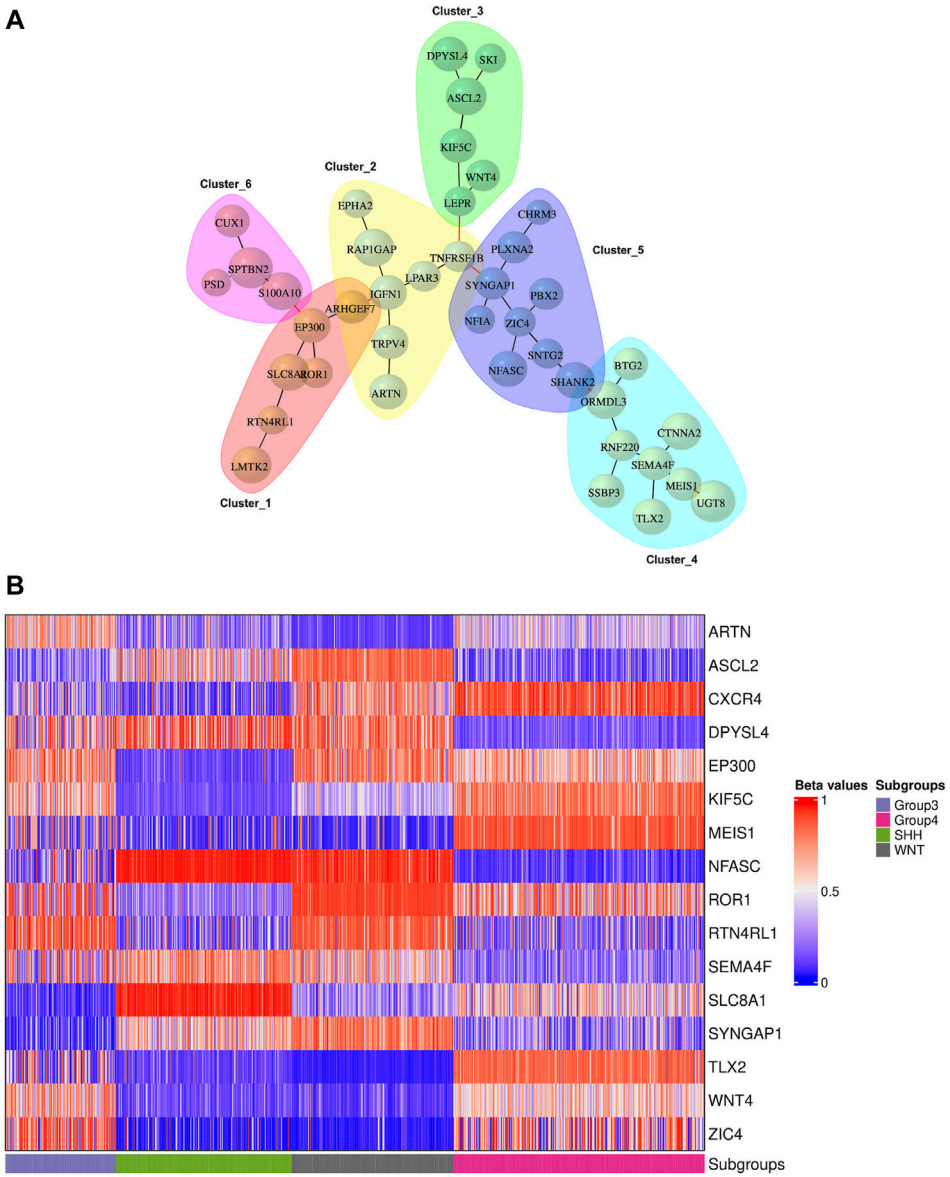


FIGURE 6
Gene network analysis and heatmap plot of the beta values associated with the prognostic genes. **(A)** Gene network representation of 41 out of 49 genes enriched in the nervous system development process, showing the correlation between genes belonging to six distinct clusters. The size of the vertices represents the beta values of the genes with a Pearson correlation coefficient above 0.6. **(B)** Methylation statuses (hypo in blue, hyper in red) of significant genes for the precise prediction of MB subgroups.

complemented by its robust performance and comprehensive evaluation in comparison to the existing methods.

Our study comprehensively evaluated the models' effectiveness in classifying MB subgroups using multiple validation datasets. Although slight variations were observed in the performance of prediction models across different datasets, the overall high performance observed in our study strengthens the reliability and generalizability of the models. Thus, incorporating multiple validation sets and prediction models is essential for robust evaluation of model reliability.

Gene-specific effects on model prediction were identified using Shapley values, offering insights into the contributions of specific genes to subgroup classification. Additionally, survival analysis identified significant associations between the identified biomarkers and survival outcomes in MB patients. Moreover, the biomarkers with significant survival outcomes correlated with previously reported oncogenes. For example, the CBFA complex, which includes *CBFA2T3* (Hendrikse et al., 2022; Gorini et al., 2023), is suggested to play a critical role in tumor development through its interactions with epigenetic modifiers, contributing to the pathogenesis of medulloblastoma. Similarly, the study by Menyhárt et al. (2019) demonstrated epigenetic changes in the *RPTOR* gene, along with other identified biomarkers, in classifying non-WNT/non-SHH medulloblastomas. These

FIGURE 7
Performance evaluation of ANN models for predicting MB subgroups. **(A)** Prediction outcomes of MB subgroups using genes within each cluster derived from the network analysis. **(B)** Performance improvement of the ANN model by including additional genes in the existing gene list within each cluster, resulting in the creation of a new cluster designated by the prime symbol.

findings suggest that the identified biomarkers hold the potential for predicting patient prognosis and guiding treatment decisions.

Our functional enrichment analysis highlighted the association between the model performance and biological relevance. For instance, *EP300* encodes a histone acetyltransferase protein that activates the expression of genes critical for the development and progression of medulloblastoma (Northcott et al., 2017). *CXCR4* has been suggested to be the oncogenic driver of MB (Amarante et al., 2018). In addition, *SYNGAP1* is a GTPase-activating protein that is known to cause cognitive deficits by inducing alterations in glutamatergic neurotransmission (Berryer et al., 2016). Finally, *WNT4* is a member of the Wnt signaling pathway and has been associated with the pathogenesis of WNT and SHH subgroups (Taylor et al., 2012). Thus, the functional insights gained from our study may contribute to identifying potential therapeutic targets for each medulloblastoma subgroup.

Finally, network analysis considered correlations among genes enriched in nervous system development and identified distinct

TABLE 4 Predicted key prognostic genes associated with molecular subgroups of medulloblastoma.

Gene name	SHH	WNT	Group 3	Group 4	Function
EP300	✓				Histone acetyltransferase; regulates cell proliferation and differentiation
CXCR4	✓		✓	✓	Chemokine receptor with high expression in breast cancer cells
WNT4		✓			Involved in oncogenesis and developmental processes, such as embryogenesis
ZIC4			✓		Transcription factor; involved in cerebellum development
MEIS1				✓	Plays a crucial role in normal development
SLC8A1	✓		✓	✓	Sodium-calcium exchanger
ASCL2		✓			Transcription factor; involved in the determination of the neuronal precursors in the peripheral nervous system and the central nervous system (CNS)
NFASC		✓		✓	Cell adhesion
KIF5C	✓			✓	Transport of cargo in CNS
SYNGAP1		✓	✓	✓	Ras GTPase; regulates synaptic plasticity and neuronal homeostasis
SEMA4F	✓				Neural development
ROR1	✓	✓	✓	✓	Neurite growth in CNS
DPYSL4	✓			✓	Development of the enteric nervous system (in mouse)
ARTN			✓		Supports the survival of several peripheral neuron populations and at least one population of dopaminergic CNS neurons
RTN4RL1	✓	✓	✓		Negative regulation of axon regeneration
TLX2			✓	✓	Transcription factor; involved in development of the enteric nervous system

clusters with potential relevance to medulloblastoma. Moreover, training a separate artificial neural network model for each cluster improved the classification accuracy by gradually incorporating genes from different clusters. Thus, our integrative approach enhances the understanding of the complex molecular heterogeneity underlying medulloblastoma and provides a basis for further research.

It is important to acknowledge some limitations of our study. Although we utilized gene expression profiles for data integration and further implemented SNF to define the new labels, our prediction models exclusively rely on the DNA methylation datasets. However, it is worth highlighting that the availability and accessibility of additional datasets, especially those including diverse patient populations, are currently limited, potentially impacting the generalizability of our findings. Therefore, further research in this direction is highly warranted to explore the clinical applicability of our study.

In conclusion, we developed a robust classifier for medulloblastoma subgroup classification. Moreover, our functional enrichment analysis offers valuable insights into the molecular pathogenesis of medulloblastoma. Survival analysis enables the evaluation of prognostic relevance for individual biomarkers. By identifying key genes in medulloblastoma subgroup classification and their functional relevance, our study provides insights into disease stratification. While our approach has the potential to be adapted for subgroup prediction in other cancer types, it requires careful validation and adaptation to specific datasets to ensure its reliability. Despite the underlying limitations, our findings contribute to the advancement of medulloblastoma research, with the potential to improve patient outcomes.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/. The package developed in this study is available from https://cran.r-project.org/web/packages/MBMethPred/index.html.

Ethics statement

This article does not contain any studies performed by the authors that involve human participants or animals.

Author contributions

VM and AS identified the problem statement and conceptualized and supervised the study. ER and AL identified the datasets. AL downloaded and processed the methylome samples. ER was responsible for the data visualization. SV tested the package and provided inputs for the improvement. PL provided a comprehensive literature search and provided input for drafting the manuscript. ER developed the R package and is responsible for its maintenance.All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the Estonian Research Council (grant PRG1076), Horizon 2020 innovation grant (ERIN, grant no. EU952516), and Enterprise Estonia (grant no EU48695).

Acknowledgments

We would like to sincerely thank Associate Professor Darja Lavogina for her valuable insights and suggestions, which greatly improved the clarity and coherence of our manuscript. We are also grateful to Berrin Buyueren, Specialist in Pathology at Hacettepe University, Turkey, and Dauren Sarsenov, Specialist in General Surgery at Mater Dei Hospital, Malta, for their helpful review and clinical guidance. Additionally, we extend our gratitude to the researchers who generously shared their data through the Gene Expression Omnibus (GEO). Their contributions were essential in developing our computational framework and advancing our research in this area.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1233657/full#supplementary-material

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "TensorFlow: A system for large-scale machine learning," in Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2–4, 2016, 265–283. doi:10.48550/arxiv.1605.08695

Adler, D., Nenadic, O., and Zucchini, W. (2003). "RGL: A R-library for 3D visualization with OpenGL," in Proceedings of the 35th Symposium of the Interface: Computing Science and Statistics, Bend OR USA, 29 October 2022- 2 November 2022.

Alharbi, M., Mobark, N., Bashawri, Y., Abu Safieh, L., Alowayn, A., Aljelaify, R., et al. (2020). Methylation profiling of medulloblastoma in a clinical setting permits sub-classification and reveals new outcome predictions. *Front. Neurology* 11, 167. doi:10.3389/fneur.2020.00167

Amarante, M. K., Vitiello, G. A. F., Rosa, M. H., Mancilla, I. A., and Watanabe, M. A. E. (2018). Potential use of CXCL12/CXCR4 and sonic hedgehog pathways as therapeutic targets in medulloblastoma. *Acta Oncol.* 57 (9), 1134–1142. doi:10.1080/0284186X.2018.1473635

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014a). Minfi: A flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30 (10), 1363–1369. doi:10.1093/BIOINFORMATICS/BTU049

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014b). Minfi: A flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30 (10), 1363–1369. doi:10.1093/bioinformatics/btu049

Attallah, O., and Zaghlool, S. (2022). AI-based pipeline for classifying pediatric medulloblastoma using histopathological and textural images. *Life* 12 (2), 232. doi:10.3390/LIFE12020232

Berryer, M. H., Chattopadhyaya, B., Xing, P., Riebe, I., Bosoi, C., Sanon, N., et al. (2016). Decrease of SYNGAP1 in GABAergic cells impairs inhibitory synapse connectivity, synaptic inhibition and cognitive function. *Nat. Commun.* 7, 13340. doi:10.1038/ncomms13340

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., et al. (2011). High density DNA methylation array with single CpG site resolution. *Genomics* 98 (4), 288–295. doi:10.1016/j.ygeno.2011.07.007

Capper, D., Jones, D. T. W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., et al. (2018). DNA methylation-based classification of central nervous system tumours. *Nature* 555 (7697), 469–474. doi:10.1038/nature26000

Cavalli, F. M. G., Remke, M., Rampasek, L., Peacock, J., Shih, D. J. H., Luu, B., et al. (2017). Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer Cell* 31 (6), 737–754. doi:10.1016/J.CCELL.2017.05.005

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi:10.1613/jair.953

Csardi, G., and Nepusz, T. (2006). *'The igraph software package for complex network research', InterJournal, Complex Sy(1–9)*, 1695. Available at: https://cran.r-project.org/web/packages/igraph/citation.html (Accessed March 10, 2023).

Gomez, S., Garrido-Garcia, A., Garcia-Gerique, L., Lemos, I., Suñol, M., de Torres, C., et al. (2018). A novel method for rapid molecular subgrouping of medulloblastoma. *Clin. Cancer Res.* 24 (6), 1355–1363. doi:10.1158/1078-0432.CCR-17-2243

Gorini, F., Miceli, M., de Antonellis, P., Amente, S., Zollo, M., and Ferrucci, V. (2023). Epigenetics and immune cells in medulloblastoma. *Front. Genet.* 14, 1135404. doi:10.3389/FGENE.2023.1135404

Grossi, E., and Buscema, M. (2007). Introduction to artificial neural networks. *Eur. J. Gastroenterol. Hepatol.* 19, 1046–1054. doi:10.1097/MEG.0b013e3282f198a0

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32 (18), 2847–2849. doi:10.1093/bioinformatics/btw313

Hendrikse, L. D., Haldipur, P., Saulnier, O., Millman, J., Sjoboen, A. H., Erickson, A. W., et al. (2022). Failure of human rhombic lip differentiation underlies medulloblastoma formation. *Nature* 609 (7929), 1021–1028. doi:10.1038/s41586-022-05215-w

Hovestadt, V., Remke, M., Kool, M., Pietsch, T., Northcott, P. A., and Fischer, R. (2013). Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. *Acta. Neuropathologica.* 125, 913–916. doi:10.1007/s00401-013-1126-5

Hovestadt, V., Ayrault, O., Swartling, F. J., Robinson, G. W., Pfister, S. M., and Northcott, P. A. (2020). Medulloblastomics revisited: biological and clinical insights from thousands of patients. *Nat. Rev. Cancer* 20 (1), 42–56. doi:10.1038/S41568-019-0223-8

Karabacak, M., Ozkara, B. B., Ozturk, A., Kaya, B., Cirak, Z., Orak, E., et al. (2022). Radiomics-based machine learning models for prediction of medulloblastoma subgroups: A systematic review and meta-analysis of the diagnostic test performance. *Acta Radiol.* 64, 1994–2003. doi:10.1177/02841851221143496

Kim, J. W., Park, S. H., Choi, S. A., Kim, S. K., Koh, E. J., Won, J. K., et al. (2022). Molecular subgrouping of medulloblastoma in pediatric population using the NanoString assay and comparison with immunohistochemistry methods. *BMC Cancer* 22 (1), 1221–1311. doi:10.1186/s12885-022-10328-6

Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., and Peterson, H. (2020). gprofiler2 - an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *Profiler* F1000h (9), 709. doi:10.12688/f1000research.24956.2

Korshunov, A., Chavez, L., Northcott, P. A., Sharma, T., Ryzhova, M., Jones, D. T. W., et al. (2017). DNA-methylation profiling discloses significant advantages over NanoString method for molecular classification of medulloblastoma. *Acta Neuropathol. Acta Neuropathol.* 134, 965–967. doi:10.1007/s00401-017-1776-9

Korshunov, A., Sahm, F., Zheludkova, O., Golanov, A., Stichel, D., Schrimpf, D., et al. (2019). DNA methylation profiling is a method of choice for molecular verification of pediatric WNT-activated medulloblastomas. *Neuro-Oncology* 21 (2), 214–221. doi:10.1093/neuonc/noy155

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28 (5), 1–26. doi:10.18637/jss.v028.i05

Le, N. Q. K., Ho, Q. T., Nguyen, V. N., and Chang, J. S. (2022). BERT-promoter: an improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection. *Comput. Biol. Chem.* 99, 107732. doi:10.1016/j.compbiolchem.2022.107732

Le, N. Q. K., Nguyen, T. T. D., and Ou, Y. Y. (2017). Identifying the molecular functions of electron transport proteins using radial basis function networks and biochemical properties. *J. Mol. Graph. Model.* 73, 166–178. doi:10.1016/j.jmgm.2017.01.003

Lian, H., Han, Y. P., Zhang, Y. C., Zhao, Y., Yan, S., Li, Q. F., et al. (2019). Integrative analysis of gene expression and DNA methylation through one-class logistic regression machine learning identifies stemness features in medulloblastoma. *Mol. Oncol.* 13 (10), 2227–2245. doi:10.1002/1878-0261.12557

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R. News* 2 (3), 18–22.

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 World Health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol. Acta Neuropathol.* 131, 803–820. doi:10.1007/s00401-016-1545-1

Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., et al. (2021). The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro-Oncology* 23 (8), 1231–1251. doi:10.1093/neuonc/noab106

Menyhárt, O., Giangaspero, F., and Gyorffy, B. (2019). Molecular markers and potential therapeutic targets in non-WNT/non-SHH (group 3 and group 4) medulloblastomas. *J. Hematol. Oncol.* 12 (1), 29–17. doi:10.1186/s13045-019-0712-y

Meyer, D. (2014). *Package 'e1071'. Misc Functions of the Department of Statistics (e1071)*. Available at: https://cran.r-project.org/web/packages/e1071/index.html (Accessed March 10, 2023).

Modhukur, V., Iljasenko, T., Metsalu, T., Lokk, K., Laisk-Podar, T., and Vilo, J. (2018). MethSurv: A web tool to perform multivariable survival analysis using DNA methylation data. *Epigenomics* 10 (3), 277–288. doi:10.2217/epi-2017-0118

Modhukur, V. (2019). *Profiling of DNA methylation patterns as biomarkers of human disease*. Tartu, Estonia: Tartu University.

Molnar, C. (2018). iml: an R package for interpretable machine learning. *J. Open Source Softw.* 3 (26), 786. doi:10.21105/joss.00786

Northcott, P. A., Shih, D. J. H., Remke, M., Cho, Y. J., Kool, M., Hawkins, C., et al. (2012). Rapid, reliable, and reproducible molecular sub-grouping of clinical medulloblastoma samples. *Acta Neuropathol.* 123 (4), 615–626. doi:10.1007/s00401-011-0899-7

Northcott, P. A., Buchhalter, I., Morrissy, A. S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., et al. (2017). The whole-genome landscape of medulloblastoma subtypes. *Nature* 547 (7663), 311–317. doi:10.1038/nature22973

Northcott, P. A., Dubuc, A. M., Pfister, S., and Taylor, M. D. (2012). Molecular subgroups of medulloblastoma. *Expert Rev. Neurother.* 12 (7), 871–884. doi:10.1586/ern.12.66

Northcott, P. A., Robinson, G. W., Kratz, C. P., Mabbott, D. J., Pomeroy, S. L., Clifford, S. C., et al. (2019). Medulloblastoma. *Nat. Rev. Dis. Prim.* 5, 11–20. doi:10.1038/s41572-019-0063-6

Peterson, H., Tank, A., Geller, D. S., Yang, R., Gorlick, R., Hoang, B. H., et al. (2020). Characterization of bony anatomic regions in pediatric and adult healthy volunteers using diffuse optical spectroscopic imaging. *Profiler*' F1000 (9), 1–17. doi:10.1117/1.JBO.25.8.086002

Ramaswamy, V., Remke, M., Bouffet, E., Bailey, S., Clifford, S. C., Doz, F., et al. (2016). Risk stratification of childhood medulloblastoma in the molecular era: the current consensus. *Acta Neuropathol.* 131 (6), 821–831. doi:10.1007/s00401-016-1569-6

Rathi, K. S., Arif, S., Koptyra, M., Naqvi, A. S., Taylor, D. M., Storm, P. B., et al. (2020). A transcriptome-based classifier to determine molecular subtypes in medulloblastoma. *PLoS Comput. Biol.* 16 (10), 10082633–e1008315. doi:10.1371/journal.pcbi.1008263

Ray, S. (2019). "A quick review of machine learning algorithms," in Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, Piscataway, New Jersey, 14th-16th February, 2019, 35–39. doi:10.1109/COMITCON.2019.8862451

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinforma.* 12, 77. doi:10.1186/1471-2105-12-77

Schwalbe, E. C., Hicks, D., Rafiee, G., Bashton, M., Gohlke, H., Enshaei, A., et al. (2017). Minimal methylation classifier (mimic): A novel method for derivation and rapid diagnostic detection of disease-associated DNA methylation signatures. *Sci. Rep.* 7 (1), 13421–13428. doi:10.1038/s41598-017-13644-1

Schwalbe, E. C., Williamson, D., Lindsey, J. C., Hamilton, D., Ryan, S. L., Megahed, H., et al. (2013). DNA methylation profiling of medulloblastoma allows robust subclassification and improved outcome prediction using formalin-fixed biopsies. *Acta Neuropathol.* 125 (3), 359–371. doi:10.1007/s00401-012-1077-2

Sharma, T., Schwalbe, E. C., Williamson, D., Sill, M., Hovestadt, V., Mynarek, M., et al. (2019). Second-generation molecular subgrouping of medulloblastoma: an international meta-analysis of group 3 and group 4 subtypes. *Acta Neuropathol.* 138 (2), 309–326. doi:10.1007/s00401-019-02020-0

Sheth, V., Tripathi, U., and Sharma, A. (2022). A comparative analysis of machine learning algorithms for classification purpose. *Procedia Comput. Sci.* 215, 422–431. doi:10.1016/j.procs.2022.12.044

Taylor, M. D., Northcott, P. A., Korshunov, A., Remke, M., Cho, Y. J., Clifford, S. C., et al. (2012). Molecular subgroups of medulloblastoma: the current consensus. *Acta Neuropathol.* 123 (4), 465–472. doi:10.1007/s00401-011-0922-z

Torgo, L. (2016). *Data mining with R: Learning with case studies*. United States: Chapman and Hall/CRC. doi:10.1201/9781315399102

Van Der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2625.

Venables, W. N., and Ripley, B. D. (2013). Functions for classification - modern applied statistics with S (MASS). *R. News* 2013. doi:10.1007/978-0-387-21706-2

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11 (3), 333–337. doi:10.1038/nmeth.2810

Xu, Z., Niu, L., and Taylor, J. A. (2021). The ENmix DNA methylation analysis pipeline for Illumina BeadChip and comparisons with seven other preprocessing pipelines. *Clin. Epigenetics* 13 (1), 216–218. doi:10.1186/s13148-021-01207-1

Yan, J., Liu, L., Wang, W., Zhao, Y., Li, K. K. W., Li, K., et al. (2020). Radiomic features from multi-parameter MRI combined with clinical parameters predict molecular subgroups in patients with medulloblastoma. *Front. Oncol.* 10, 558162. doi:10.3389/fonc.2020.558162

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 06 October 2023
DOI 10.3389/fgene.2023.1190863

OPEN ACCESS

EDITED BY
Angelo Facchiano,
National Research Council (CNR), Italy

REVIEWED BY
Sanga Mitra,
Indian Institute of Technology Madras, India
Carole Sousa,
International Iberian Nanotechnology Laboratory (INL), Portugal

*CORRESPONDENCE
Tianxiao Zhang,
joshuaz@mail.xjtu.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 02 May 2023
ACCEPTED 26 September 2023
PUBLISHED 06 October 2023

CITATION
Wang S, Fang X, Wen X, Yang C, Yang Y and Zhang T (2023), Prioritization of risk genes for Alzheimer's disease: an analysis framework using spatial and temporal gene expression data in the human brain based on support vector machine. *Front. Genet.* 14:1190863. doi: 10.3389/fgene.2023.1190863

Prioritization of risk genes for Alzheimer's disease: an analysis framework using spatial and temporal gene expression data in the human brain based on support vector machine

Shiyu Wang[1†], Xixian Fang[1†], Xiang Wen[2], Congying Yang[1], Ying Yang[1] and Tianxiao Zhang[1,3*]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, China, [2]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Beijing, China, [3]National Anti-Drug Laboratory Shaanxi Regional Center, Xi'an, China

Background: Alzheimer's disease (AD) is a complex disorder, and its risk is influenced by multiple genetic and environmental factors. In this study, an AD risk gene prediction framework based on spatial and temporal features of gene expression data (STGE) was proposed.

Methods: We proposed an AD risk gene prediction framework based on spatial and temporal features of gene expression data. The gene expression data of providers of different tissues and ages were used as model features. Human genes were classified as AD risk or non-risk sets based on information extracted from relevant databases. Support vector machine (SVM) models were constructed to capture the expression patterns of genes believed to contribute to the risk of AD.

Results: The recursive feature elimination (RFE) method was utilized for feature selection. Data for 64 tissue-age features were obtained before feature selection, and this number was reduced to 19 after RFE was performed. The SVM models were built and evaluated using 19 selected and full features. The area under curve (AUC) values for the SVM model based on 19 selected features (0.740 [0.690–0.790]) and full feature sets (0.730 [0.678–0.769]) were very similar. Fifteen genes predicted to be risk genes for AD with a probability greater than 90% were obtained.

Conclusion: The newly proposed framework performed comparably to previous prediction methods based on protein-protein interaction (PPI) network properties. A list of 15 candidate genes for AD risk was also generated to provide data support for further studies on the genetic etiology of AD.

KEYWORDS

Alzheimer's disease, risk gene prioritization, gene expression patterns, machine learning, genome-wide association analyses

1 Introduction

Alzheimer's disease (AD) is a chronic neurodegenerative disorder that is characterized by cognitive impairment and memory loss. It affected approximately 50 million people worldwide in 2020, which is expected to increase to 150 million by 2050 (Breijyeh and Karaman, 2020). Advanced age is the most important risk factor for AD (Knopman et al., 2021). A significant increase in the incidence rate of AD was observed in senior citizens after the age of 65 years (Knopman et al., 2021). Equal incidence rates of AD were identified for males and females after adjusting for age, indicating that sex might not be associated with the risk of AD (Knopman et al., 2021). The pathological features of AD include senile plaques formed by the accumulation of β-amyloid protein and neurofibrillary tangles composed of highly phosphorylated τ proteins. Several hypotheses have been proposed to explain the pathogenesis of AD, including oxidative stress (Yang et al., 2022), inflammation (Yang et al., 2022), and DNA damage (Tanaka et al., 2021). However, no consensus has yet been reached.

Previous studies have indicated that AD is a complex disorder, and its risk is attributed to multiple genetic and environmental factors (Carmona et al., 2018; Bertram and Tanzi, 2019). In the last decade, genome-wide association (GWA) analyses have significantly contributed to the genetic etiology of AD (Bertram and Tanzi, 2019). Jansen et al. confirmed 29 risk loci and several relevant pathways related to AD through a GWA meta-analysis (Jansen et al., 2019). In addition, Celeste et al. reviewed the relationship between several AD risk genes, including ABCA7, BIN1, CASS4, and CD33, and the cellular and neuropathological characteristics of AD (Karch et al., 2014). Nevertheless, a recent study indicated that approximately half of the heritability of AD remains unaccounted (Raybould and Sims, 2021). It is probable that a large number of susceptibility loci for AD have not yet been discovered. However, recent studies have indicated that larger-scale GWA studies in the future are less cost effective due to the intrinsic deficiency rooted in the study design of GWA studies; therefore, it might not be a preferable choice for unraveling these hidden genomic regions that contribute to the risk of AD (Escott-Price and Hardy, 2022). In this sense, prioritizing AD risk genes based on evidence gained from different perspectives and then validating these candidate risk genes in subsequent candidate gene-based association studies might be an effective strategy for discovering more relevant genes for AD risk. In a recent study, Cogill et al. applied machine-learning-based methods using brain developmental gene expression data to prioritize high-confidence candidate genes for autism spectrum disorder (Cogill and Wang, 2016). This study established a feasible analysis pipeline for prioritizing candidate risk genes for complex disorders, using spatial and temporal gene expression data.

Multiple lines of evidence have indicated that the expression of AD risk genes has specific spatial and temporal features (Moradifard et al., 2018; Grubman et al., 2019). Extracting and properly synthesizing information from these gene expression features might be an effective way to prioritize the risk genes for AD. In this study, we aimed to construct and evaluate a machine-learning-based model to identify high-confidence risk genes for AD using spatial and temporal gene expression data extracted from a publicly available database.

2 Materials and methods

The statistical analysis pipeline is shown in Figure 1. In this study, we propose an AD risk gene prediction framework based on spatial and temporal features of gene expression data (STGE). In this analysis framework, the gene expression data of providers of different tissues and ages were utilized as model features. Human genes were classified as AD risk or non-risk sets and randomly split into training and validation sets. Support vector machine (SVM) models were constructed to capture the expression patterns of genes that were believed to contribute to the risk of AD in the training set, which were then applied to the validation set to evaluate model performance. The STGE model was then applied to a gene set with an unknown status for AD risk, and a confidence score was assigned to each gene.

2.1 Data extraction

The data used in the present study were extracted from three publicly available databases: the GTEx database (https://gtexportal.org/home/) (GTEx Consortium, 2020), AlzData database (http://www.alzdata.org/) (Xu et al., 2018), and GWAS catalog (https://www.ebi.ac.uk/gwas/) (Buniello et al., 2019).

Spatial and temporal expression data for each gene were obtained from the GTEx database. Gene expression data related to tissues of the human brain (including the cerebellum, cortex, anterior cingulate cortex, hippocampus, substantia nigra, caudate, cerebellar hemisphere, frontal cortex, hypothalamus, nucleus accumbens, putamen, spinal cord, and amygdala) were extracted. Data from tissue sample providers under 20 or over 70 years of age were not included, and all these providers were healthy. In addition, we also removed tissue providers who scored 0 or 4 points on the death classification provided by GTEx database basing on the 4-point Hardy Scale (Hardy et al., 1985), because those scores represent the death of the provider is associated with chronic disease. Specifically, the score of 0 added by GTEx database stands for ventilator case (all cases on a ventilator immediately before death), and the score of 4 stands for slow death case (death after a long illness, with a terminal phase longer than 1 day; deaths that are not unexpected). Finally, gene expression data in 13 types of brain-related tissues for 14,697 genes were extracted from 317 tissue sample providers of various ages and genders (Supplementary Table S1 and Supplementary Figure S1).

AlzData is a database for scoring correlations between human genes and the risk of AD, based on evidence from high-throughput omics data. The CFG scores ranged from 0 to 5, with a higher score indicating a stronger correlation between the gene and AD. Genes with scores of 4–5 were extracted to form the AD risk gene set ("the right answer"). For genes with scores of 0–3, we supplemented the "DISEASE/TRAIT" (we always call it "trait" for short) from the GWAS catalog and excluded genes related to AD to obtain AD non-risk genes. Finally, 3,899 genes comprising 340 AD risk genes and 3,559 non-AD risk genes were identified, and these genes' CFG scores and GWAS traits are shown in Supplementary Table S2.

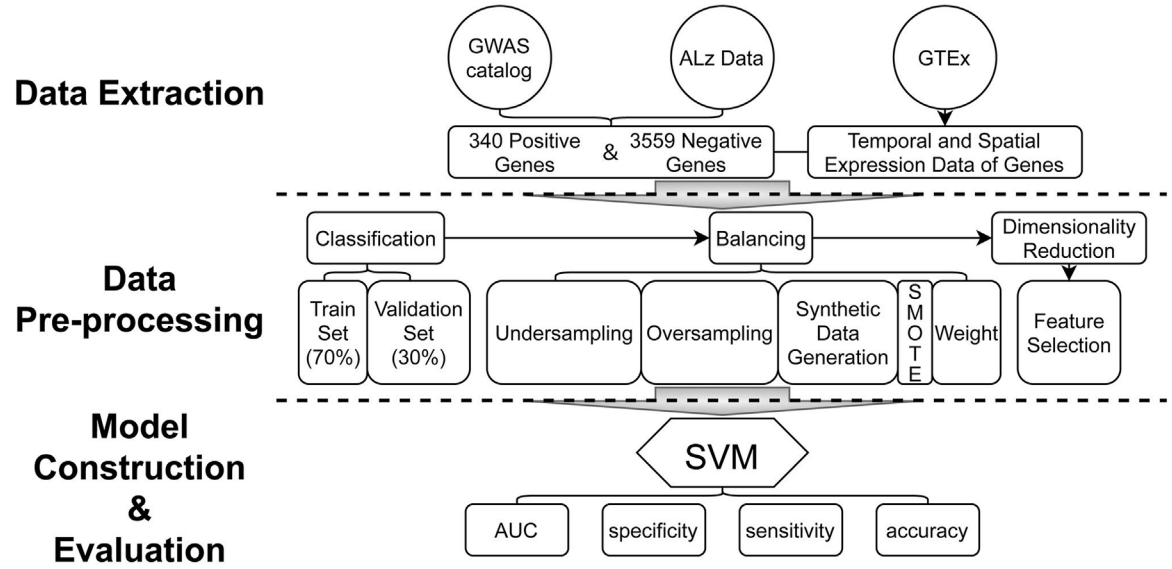


FIGURE 1
Analysis pipeline of the model construction and evaluations.

2.2 Model construction and evaluation

The SVM models were constructed based on spatial and temporal gene expression data extracted from relevant databases using the e1071 package of the R project, and both spatial and temporal aspects of the data are contained in the data features, which is going to be used for feature selection. Gene expression data were first grouped by the tissue type and age of the tissue providers. The median expression level of each gene in the tissue type age group was calculated and used as features in the SVM models. A total of 64 brain tissue-related features were obtained for model construction (Supplementary Table S3). The dataset was randomly divided into training and validation sets in a ratio of 7: 3. There were 238 AD risk genes and 2,491 AD non-risk genes in the training set. The SMOTE function in the DMwR package was used to balance gene numbers. Feature selection was conducted using the caret package, and 19 spatial and temporal features were selected based on recursive feature elimination (RFE). Accuracy and Kappa statistics were chosen as the evaluation indicators to estimate the performance of the selected features, and we chose the feature set with both the greatest value and least variance to build the SVM model. Parameter optimization was performed using a grid search strategy. Parameters including model accuracy, specificity, sensitivity, and area under the curve (AUC) were utilized to evaluate the performance of the SVM model. The R packages pROC and ROCR were used to draw the ROC curve and calculate the AUC, respectively. The R package ggplot2 was used for data visualization.

2.3 Results validation

After the genes with high confidence were predicted by SVM model, the normalized expression in AlzData database (http://www.alzdata.org/Normalized_differential1.php) will be used for providing these genes' differential expression data. Besides, KOBAS platform (http://kobas.cbi.pku.edu.cn/) will be used to do gene ontology (GO) and KEGG pathway enrichment analysis in all available databases (including OMIM, KEGG Disease and NHGRI GWAS Catalog).

3 Results

3.1 Feature selection based on recursive feature elimination

RFE was used for feature selection. Data for 64 tissue-age features were obtained before feature selection, and this number was reduced to 19 after RFE was performed. SVM models based on each of these 19 features (the gene expression levels were obtained by median values of samples) were built and evaluated for accuracy, specificity, sensitivity, and AUC (Table 1). The feature with the highest AUC was the human tissue of the brain cerebellum at the age of 40–49 (AUC = 0.688).

3.2 Comparison and validation of SVM models

SVM models were built and evaluated using 19 selected and full features (Table 2 and Figure 2). The AUC values for the SVM model based on 19 selected features (0.74 [0.690–0.790]) and full feature sets (0.730 [0.678–0.769]) were very similar. To evaluate model robustness, we also constructed these models based on the mean expression level of each gene in the tissue type age group. In addition, to examine the potential effects of sex, SVM models were constructed based on the expression data from male and female samples. The results are

TABLE 1 The mean accuracy, sensitivity, specificity and AUC of each model built by each selected feature from the RFE method.

Tissue	Age	Accuracy	Sensitivity	Specificity	AUC
Brain-Cerebellum	40–49	0.604	0.657	0.599	0.688
Brain-Amygdala	50–59	0.768	0.441	0.799	0.684
Brain-Frontal Cortex BA9	50–59	0.657	0.598	0.663	0.683
Brain-Anterior cingulate cortex BA24	30–39	0.67	0.559	0.681	0.683
Brain-Putamen basal ganglia	30–39	0.645	0.569	0.653	0.682
Brain-Anterior cingulate cortex BA24	40–49	0.701	0.529	0.717	0.678
Brain-Cerebellum	60–69	0.551	0.676	0.539	0.674
Brain-Cerebellar Hemisphere	60–69	0.689	0.549	0.702	0.667
Brain-Frontal Cortex BA9	60–69	0.644	0.520	0.656	0.665
Brain-Substantia nigra	40–49	0.715	0.500	0.736	0.664
Brain-Putamen basal ganglia	60–69	0.691	0.520	0.707	0.655
Brain-Caudate basal ganglia	30–39	0.695	0.559	0.708	0.640
Brain-Anterior cingulate cortex BA24	50–59	0.733	0.461	0.759	0.636
Brain-Cerebellum	30–39	0.770	0.461	0.800	0.633
Brain-Substantia nigra	60–69	0.774	0.461	0.803	0.628
Brain-Amygdala	60–69	0.736	0.480	0.760	0.626
Brain-Nucleus accumbens basal ganglia	40–49	0.564	0.598	0.561	0.621
Brain-Nucleus accumbens basal ganglia	60–69	0.561	0.539	0.563	0.606
Brain-Hypothalamus	30–39	0.555	0.52	0.558	0.604

TABLE 2 The average accuracy, sensitivity, specificity and AUC of the two models based on ten-fold cross validation.

	Selected feature set	Full feature set
Accuracy	0.756 ± 0.016	0.754 ± 0.023
Sensitivity	0.588 ± 0.069	0.500 ± 0.054
Specificity	0.772 ± 0.016	0.778 ± 0.021
AUC(95%CI)	0.740 (0.690–0.790)	0.730 (0.678–0.769)

summarized in Supplementary Table S4. There are no significant differences when mean values were utilized compared to median values. The model performance based on males or females was also very similar to that of models constructed using all samples. Finally, we chose the selected feature and median values to construct the SVM model because of its highest AUC. Besides, some known AD risk genes (such as APOE, PICALM and BIN1) were recovered with the final SVM model, and the probabilities of them being classified as AD risk genes are ranged from 0.723–0.783 and shown in Supplementary Table S5.

3.3 Risk genes of AD predicted by the SVM model

Based on the SVM models constructed using tissue-age-specific gene expression data, the risk contributions to AD onset and development were evaluated for 10,798 genes that were not included in the model construction and evaluations (the external gene set). 15 genes predicted to be risk genes for AD with a probability greater than 90% were obtained (Table 3). Among these genes, GUCY1B3 had the highest confidence score as a risk gene for AD (0.93). To further investigate this gene set, we examined the gene expression patterns of these 15 genes in the human brain and made a heatmap showing in Supplementary Figure S2. In addition, 191 risk genes for AD with a probability greater than 80% are shown in Supplementary Table S6.

3.4 Differential gene expression analysis and pathway/ontology analysis

After the normalized differential gene expression analysis, there exist 8 genes among 15 candidate genes expressing differentially in AD. The differential expression data of these 8 genes are shown in Table 4. The GO and KEGG pathway enrichment analyses find out 15 pathways that are statistically correlated with candidate genes, which are shown in Supplementary Figures S3, S4.

4 Discussion

In the present study, we propose a novel machine-learning-based analysis pipeline using data extracted from the GTEx database

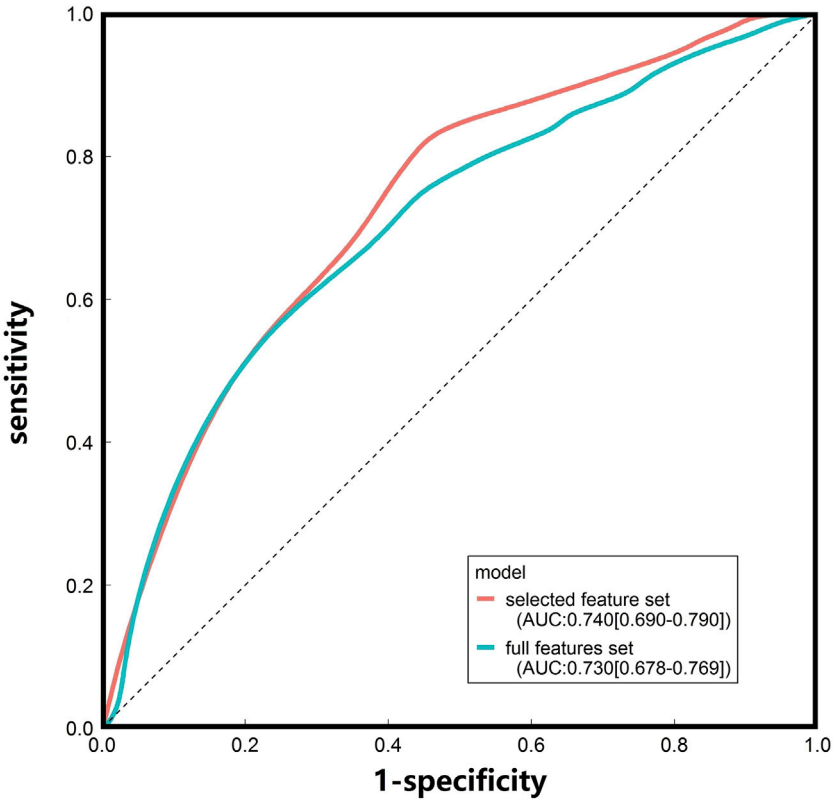


FIGURE 2
ROC curves of the SVM models constructed based on the median gene expression levels in different tissue-age groups.

to prioritize candidate AD risk genes. The performance measured by the AUC of the SVM models was promising, and a list of 15 candidate AD risk genes was presented according to the prediction model. In the last decade, several studies have been published to identify candidate AD risk genes, and most of these studies were based on protein–protein interaction (PPI) networks to identify hub genes using GWA data. The model performance measured by the AUC of these previous studies ranged from 0.63 to 0.84 depending on different settings (Luo et al., 2019; Lagisetty et al., 2022; Wang et al., 2022; Pei et al., 2023). The methods used in these comparative studies and their AUC are shown in the Supplementary Table S7. Unlike these previous studies, the STGE framework was used to predict AD candidate genes based on the spatial and temporal features of AD risk gene expression. The performance of our model (AUC = 0.74) was comparable to that of previous studies. In this sense, the present study proposed and validated an alternative framework for prioritizing risk genes for AD. In the future, an analysis framework integrating information from gene expression features and PPI network properties might be a promising method to further promote the accuracy and effectiveness of prediction models for prioritizing candidate AD risk genes.

Although most patients with AD experience the first symptom in their mid-60s, previous studies have indicated that changes in the molecular levels occur at a much earlier stage (Egan et al., 2019; Vermunt et al., 2019). A previously published family-based longitudinal study has shown that familial AD may have a long prodromal phase of several years (Chiotis et al., 2018). A recent cohort study also indicated that plasma phospho-tau181 levels were much higher from 16 years prior to the onset of AD symptoms in AD patients with specific DNA mutations (Wang et al., 2021; Karikari et al., 2022). The results of the current study offer new evidence at the gene expression level for prodromal changes in AD patients. Although AD is a late-onset disorder, more than half of the selected features were obtained from sample providers before the age of 60 years. Five of the 19 features, including tissues of the anterior cingulate cortex, putamen basal ganglia, caudate basal ganglia, cerebellum, and hypothalamus, were obtained from providers who are 30–39 years old. In accordance with multiple lines of previous evidence, these findings indicate that molecular-level changes might be identified several years before early symptoms appear in patients with AD. Nevertheless, since a couple of the AD risk genes used in this study were extracted from studies focusing on early-onset AD, we need to be cautious in interpreting these results. Future research using longitudinal data might provide more clues for identifying prodromal biomarkers for AD and, in turn, shed light on early screening and prevention of this complex neurodegenerative disorder.

Among the 15 candidate genes identified through STGE, a few are of particular interest. Sine oculis homeobox homolog 3 (SIX3) encodes a type of transcription factor belonging to the sine oculis homeobox transcription factor family (Steinmetz et al., 2010). Multiple lines of evidence based on animal models have linked this locus to brain development (Steinmetz et al., 2010; Schacht et al.,

TABLE 3 Genes predicted by the SVM model with their confidence score, location, length (bp) and biotype.

Genes	Confidence	Location	Length (bp)	Type
SIX3	0.911	2p21	4,370	protein coding
EFEMP1	0.904	2p16.1	58,197	protein coding
GUCY1B3	0.939	4p32.1	48,820	protein coding
MTPN	0.921	7q33	50,600	protein coding
ACTR3B	0.904	7q36.2	311,231	protein coding
BAG3	0.932	10q26.11	26,440	protein coding
INPP5A	0.924	10q26.3	245,697	protein coding
LRRC10B	0.917	11q12.2	2,270	protein coding
ELMOD1	0.914	11q22.3	75,771	protein coding
DRD2	0.900	11q23.2	66,087	protein coding
GABRA5	0.922	15q12	82,490	protein coding
PITPNM3	0.916	17p13.2-p13.1	105,293	protein coding
SEZ6	0.909	17q11.2	51,540	protein coding
ICAM5	0.922	19p13.2	7,428	protein coding
CSDC2	0.916	22q13.2	16,732	protein coding

TABLE 4 Differential expression results of candidate genes with FDRs < 0.05.

Genes	Brain region	log2 FoldChange	*p*-value	FDR
EFEMP1	Hippocampus	0.52	0.001	0.020
GUCY1B3	Hippocampus	−0.24	0.004	0.048
	Temporal Cortex	−0.73	0.000	0.000
ACTR3B	Temporal Cortex	−0.8	0.001	0.006
BAG3	Temporal Cortex	0.81	0.000	0.001
	Frontal Cortex	0.42	0.001	0.012
INPP5A	Temporal Cortex	−0.47	0.000	0.001
GABRA5	Hippocampus	−0.57	0.003	0.036
	Temporal Cortex	−1.37	0.000	0.000
SEZ6	Temporal Cortex	−1.37	0.000	0.000
ICAM5	Entorhinal Cortex	−0.61	0.001	0.015
	Frontal Cortex	−0.41	0.000	0.000

2020). A recent GWA study associated genetic polymorphisms of SIX3 with math ability, and its weakening was considered a sign of the progression of AD patients (Lee et al., 2018). Actin-related protein 3B (ACTR3B) encodes a member of the actin-related protein (ARP) family, which might regulate and induce cell shape changes and motility (Hu et al., 2018). Several previous studies have linked ACTR3B to brain aging progression, although no direct GWA study has validated the connection between genetic polymorphisms of these loci and AD (Hu et al., 2018; Seefelder and Kochanek, 2021). In addition, multiple animal models and population-based evidence have been published for dopamine receptor D2 (DRD2) and gamma-aminobutyric acid type A receptor subunit alpha 5 (GABRA5) being associated with brain-related disorders and traits, including schizophrenia, bipolar disorder, Parkinson's disorder, and neurotransmission (Prisciandaro et al., 2017; Escamilla et al., 2018; Mundorf et al., 2021; Zhang et al., 2021). In a recent study, Blum et al. concluded that the DRD2 Taq1A A1 allele might increase the risk of Alzheimer's aging in African Americans by integrating and reviewing previously published data (Blum et al., 2018). Additionally, the genes BAG Cochaperone 3 (BAG3), inositol polyphosphate-5-phosphatase A (INPP5A), seizure related 6 homolog (SEZ6), and intercellular adhesion molecule 5 (ICAM5) are involved in the progression of AD has been proposed in several functional studies using animal models (Hoarau et al., 2011; Paetau et al., 2017; Zhu et al., 2018; Zhou et al., 2020; Zhu et al., 2021). Within these genes, through proteomic study, BAG3 may affect AD by influencing the interpretation of Aβ and tau protein, and patients with AD have much lower levels of SEZ6 in their cerebrospinal fluid than those without dementia (Khoonsari et al., 2016; Gonzalez-Rodriguez et al., 2021). Further *in vivo* and *in vitro* studies are needed to validate the functional connections between the risk of AD and the genes on the predicted list.

Three of the 15 pathways identified by GO and KEGG pathway enrichment analyses are worthy of attention, including regulation of synapse structural plasticity, branching morphogenesis of a nerve and forced vital capacity. According to a review, synapse structural plasticity is related to the number of spines, and post-mortem reports of Alzheimer's brains showed reduced spine number in the hippocampus and cortex (Chidambaram et al., 2019). One research studying novel compounds' effect on neuronal branching morphogenesis of PC12 cells indicates that branching morphogenesis is one of the entry points for research to promote recovery of nerve regeneration following neurodegenerative diseases, like AD (Katebi et al., 2019). A prospective cohort study of 431,834 individuals shows that per unit decrease in lung function measure was each associated with increased risk for all-cause dementia (including AD). As for forced vital capacity, its hazard ratio (HR) is 1.16 and *p*-value is 2.04×10^{-5} (Ma et al., 2023).

The current study has several limitations. First, there is still much space for the promotion of STGE, although the performance of STGE is comparable to that of previous models based on PPI network properties. In addition, as bioinformatics data mining is based on publicly available databases, the completeness of the current work might be limited owing to data availability. The gene expression data in the brain substantia nigra in the age group of 30–39 years were unavailable from the database; therefore, this feature was not included in the model

construction and evaluation. Besides, although the data for training the model contains non-coding RNA, which have been shown to play an important role in the pathogenesis of complex disorders (Goyal et al., 2018), all candidate AD risk genes are protein-coding genes in the current study. Furthermore, the data we used in our research can only correlate to tissues, so we were unable to associate these genes with specific brain cell types.

In summary, in the present study, an efficient analysis framework based on spatial and temporal features of gene expression was proposed to prioritize AD risk genes. The newly proposed framework performed comparably to previous prediction methods based on PPI network properties. A list of 15 candidate genes for AD risk was also generated to provide data support for further studies on the genetic etiology of AD.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.gtexportal.org/home/datasets; https://www.ebi.ac.uk/gwas/docs/file-downloads; http://www.alzdata.org/CFG_rank1.php.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

TZ, SW, and XF designed the study. TZ, SW, and XF wrote the main manuscript text. SW, XF and XW conducted the statistical analysis. SW, XF, CY, and YY prepared all the tables, figures and Supplementary Materials for this manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (NSFC) Young Scientists Fund (31900407).

Acknowledgments

We would thank Yingying Wei who has provided insightful suggestions and significantly promoted the manuscript. A preprint version of this manuscript could be found on medRxic (link: https://www.medrxiv.org/content/10.1101/2023.02.06.23285522v1).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1190863/full#supplementary-material

References

Bertram, L., and Tanzi, R. E. (2019). Alzheimer disease risk genes: 29 and counting. *Nat. Rev. Neurol.* 15 (4), 191–192. doi:10.1038/s41582-019-0158-4

Blum, K., Badgaiyan, R. D., Dunston, G. M., Baron, D., Modestino, E. J., McLaughlin, T., et al. (2018). The DRD2 Taq1A A1 allele may magnify the risk of Alzheimer's in aging african-Americans. *Mol. Neurobiol.* 55 (7), 5526–5536. doi:10.1007/s12035-017-0758-1

Breijyeh, Z., and Karaman, R. (2020). Comprehensive review on Alzheimer's disease: causes and treatment. *Molecules* 25 (24), 5789. doi:10.3390/molecules25245789

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47 (D1), D1005–D1012. doi:10.1093/nar/gky1120

Carmona, S., Hardy, J., and Guerreiro, R. (2018). The genetic landscape of Alzheimer disease. *Handb. Clin. Neurol.* 148, 395–408. doi:10.1016/B978-0-444-64076-5.00026-0

Chidambaram, S. B., Rathipriya, A. G., Bolla, S. R., Bhat, A., Ray, B., Mahalakshmi, A. M., et al. (2019). Dendritic spines: revisiting the physiological role. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 92, 161–193. doi:10.1016/j.pnpbp.2019.01.005

Chiotis, K., Saint-Aubert, L., Rodriguez-Vieitez, E., Leuzy, A., Almkvist, O., Savitcheva, I., et al. (2018). Longitudinal changes of tau PET imaging in relation to hypometabolism in prodromal and Alzheimer's disease dementia. *Mol. Psychiatry* 23 (7), 1666–1673. doi:10.1038/mp.2017.108

Cogill, S., and Wang, L. (2016). Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates. *Bioinformatics* 32 (23), 3611–3618. doi:10.1093/bioinformatics/btw498

Egan, M. F., Kost, J., Voss, T., Mukai, Y., Aisen, P. S., Cummings, J. L., et al. (2019). Randomized trial of verubecestat for prodromal Alzheimer's disease. *N. Engl. J. Med.* 380 (15), 1408–1420. doi:10.1056/NEJMoa1812840

Escamilla, R., Camarena, B., Saracco-Alvarez, R., Fresán, A., Hernández, S., and Aguilar-García, A. (2018). Association study between COMT, DRD2, and DRD3 gene variants and antipsychotic treatment response in Mexican patients with schizophrenia. *Neuropsychiatr. Dis. Treat.* 14, 2981–2987. doi:10.2147/NDT.S176455

Escott-Price, V., and Hardy, J. (2022). Genome-wide association studies for Alzheimer's disease: bigger is not always better. *Brain Commun.* 4 (3), fcac125. doi:10.1093/braincomms/fcac125

Gonzalez-Rodriguez, M., Villar-Conde, S., Astillero-Lopez, V., Villanueva-Anguita, P., Ubeda-Banon, I., Flores-Cuadrado, A., et al. (2021). Neurodegeneration and

astrogliosis in the human CA1 hippocampal subfield are related to hsp90ab1 and bag3 in Alzheimer's disease. *Int. J. Mol. Sci.* 23 (1), 165. doi:10.3390/ijms23010165

Goyal, N., Kesharwani, D., and Datta, M. (2018). Lnc-ing non-coding RNAs with metabolism and diabetes: roles of lncRNAs. *Cell Mol. Life Sci.* 75 (10), 1827–1837. doi:10.1007/s00018-018-2760-9

Grubman, A., Chew, G., Ouyang, J. F., Sun, G., Choo, X. Y., McLean, C., et al. (2019). A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* 22 (12), 2087–2097. doi:10.1038/s41593-019-0539-4

GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369 (6509), 1318–1330. doi:10.1126/science.aaz1776

Hardy, J. A., Wester, P., Winblad, B., Gezelius, C., Bring, G., and Eriksson, A. (1985). The patients dying after long terminal phase have acidotic brains; implications for biochemical measurements on autopsy tissue. *J. Neural Transm.* 61 (3-4), 253–264. doi:10.1007/BF01251916

Hoarau, J. J., Krejbich-Trotot, P., Jaffar-Bandjee, M. C., Das, T., Thon-Hon, G. V., Kumar, S., et al. (2011). Activation and control of CNS innate immune responses in health and diseases: a balancing act finely tuned by neuroimmune regulators (NIReg). *CNS Neurol. Disord. Drug Targets* 10 (1), 25–43. doi:10.2174/187152711794488601

Hu, Y., Pan, J., Xin, Y., Mi, X., Wang, J., Gao, Q., et al. (2018). Gene expression analysis reveals novel gene signatures between Young and old adults in human prefrontal cortex. *Front. Aging Neurosci.* 10, 259. doi:10.3389/fnagi.2018.00259

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* 51 (3), 404–413. doi:10.1038/s41588-018-0311-9

Karch, C. M., Cruchaga, C., and Goate, A. M. (2014). Alzheimer's disease genetics: from the bench to the clinic. *Neuron* 83 (1), 11–26. doi:10.1016/j.neuron.2014.05.041

Karikari, T. K., Ashton, N. J., Brinkmalm, G., Brum, W. S., Benedet, A. L., Montoliu-Gaya, L., et al. (2022). Blood phospho-tau in alzheimer disease: analysis, interpretation, and clinical utility. *Nat. Rev. Neurol.* 18 (7), 400–418. Epub ahead of print. doi:10.1038/s41582-022-00665-2

Katebi, S., Esmaeili, A., Ghaedi, K., and Zarrabi, A. (2019). Superparamagnetic iron oxide nanoparticles combined with NGF and quercetin promote neuronal branching morphogenesis of PC12 cells. *Int. J. Nanomedicine* 14, 2157–2169. doi:10.2147/IJN.S191878

Khoonsari, P. E., Häggmark, A., Lönnberg, M., Mikus, M., Kilander, L., Lannfelt, L., et al. (2016). Analysis of the cerebrospinal fluid proteome in Alzheimer's disease. *PLoS One* 11 (3), e0150672. doi:10.1371/journal.pone.0150672

Knopman, D. S., Amieva, H., Petersen, R. C., Chételat, G., Holtzman, D. M., Hyman, B. T., et al. (2021). Alzheimer disease. *Nat. Rev. Dis. Prim.* 7 (1), 33. doi:10.1038/s41572-021-00269-y

Lagisetty, Y., Bourquard, T., Al-Ramahi, I., Mangleburg, C. G., Mota, S., Soleimani, S., et al. (2022). Identification of risk genes for Alzheimer's disease by gene embedding. *Cell Genom* 2 (9), 100162. doi:10.1016/j.xgen.2022.100162

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* 50 (8), 1112–1121. doi:10.1038/s41588-018-0147-3

Luo, P., Tian, L. P., Ruan, J., and Wu, F. X. (2019). Disease gene prediction by integrating PPI networks, clinical RNA-seq data and OMIM data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16 (1), 222–232. doi:10.1109/TCBB.2017.2770120

Ma, Y. H., Shen, L. X., Li, Y. Z., Leng, Y., Yang, L., Chen, S. D., et al. (2023). Lung function and risk of incident dementia: A prospective cohort study of 431,834 individuals. *Brain Behav. Immun.* 109, 321–330. doi:10.1016/j.bbi.2023.02.009

Moradifard, S., Hoseinbeyki, M., Ganji, S. M., and Minuchehr, Z. (2018). Analysis of microRNA and gene expression profiles in Alzheimer's disease: A meta-analysis approach. *Sci. Rep.* 8 (1), 4767. doi:10.1038/s41598-018-20959-0

Mundorf, A., Kubitza, N., Hünten, K., Matsui, H., Juckel, G., Ocklenburg, S., et al. (2021). Maternal immune activation leads to atypical turning asymmetry and reduced DRD2 mRNA expression in a rat model of schizophrenia. *Behav. Brain Res.* 414, 113504. doi:10.1016/j.bbr.2021.113504

Paetau, S., Rolova, T., Ning, L., and Gahmberg, C. G. (2017). Neuronal ICAM-5 inhibits microglia adhesion and phagocytosis and promotes an anti-inflammatory response in LPS stimulated microglia. *Front. Mol. Neurosci.* 10, 431. doi:10.3389/fnmol.2017.00431

Pei, Y., Chen, S., Zhou, F., Xie, T., and Cao, H. (2023). Construction and evaluation of Alzheimer's disease diagnostic prediction model based on genes involved in mitophagy. *Front. Aging Neurosci.* 15, 1146660. doi:10.3389/fnagi.2023.1146660

Prisciandaro, J. J., Tolliver, B. K., Prescot, A. P., Brenner, H. M., Renshaw, P. F., Brown, T. R., et al. (2017). Unique prefrontal GABA and glutamate disturbances in co-occurring bipolar disorder and alcohol dependence. *Transl. Psychiatry* 7 (7), e1163. doi:10.1038/tp.2017.141

Raybould, R., and Sims, R. (2021). Searching the dark genome for Alzheimer's disease risk variants. *Brain Sci.* 11 (3), 332. doi:10.3390/brainsci11030332

Schacht, M. I., Schomburg, C., and Bucher, G. (2020). six3 acts upstream of foxQ2 in labrum and neural development in the spider Parasteatoda tepidariorum. *Dev. Genes Evol.* 230 (2), 95–104. doi:10.1007/s00427-020-00654-9

Seefelder, M., and Kochanek, S. (2021). A meta-analysis of transcriptomic profiles of Huntington's disease patients. *PLoS One* 16 (6), e0253037. doi:10.1371/journal.pone.0253037

Steinmetz, P. R., Urbach, R., Posnien, N., Eriksson, J., Kostyuchenko, R. P., Brena, C., et al. (2010). Six3 demarcates the anterior-most developing brain region in bilaterian animals. *Evodevo* 1 (1), 14. doi:10.1186/2041-9139-1-14

Tanaka, H., Kondo, K., Fujita, K., Homma, H., Tagawa, K., Jin, X., et al. (2021). HMGB1 signaling phosphorylates Ku70 and impairs DNA damage repair in Alzheimer's disease pathology. *Commun. Biol.* 4 (1), 1175. doi:10.1038/s42003-021-02671-4

Vermunt, L., Sikkes, S. A. M., van den Hout, A., Handels, R., Bos, I., van der Flier, W. M., et al. (2019). Duration of preclinical, prodromal, and dementia stages of Alzheimer's disease in relation to age, sex, and APOE genotype. *Alzheimers Dement.* 15 (7), 888–898. doi:10.1016/j.jalz.2019.04.001

Wang, Y., Chen, G., and Shao, W. (2022). Identification of ferroptosis-related genes in Alzheimer's disease based on bioinformatic analysis. *Front. Neurosci.* 16, 823741. doi:10.3389/fnins.2022.823741

Wang, Y. L., Chen, J., Du, Z. L., Weng, H., Zhang, Y., Li, R., et al. (2021). Plasma p-tau181 level predicts neurodegeneration and progression to Alzheimer's dementia: A longitudinal study. *Front. Neurol.* 12, 695696. doi:10.3389/fneur.2021.695696

Xu, M., Zhang, D. F., Luo, R., Wu, Y., Zhou, H., Kong, L. L., et al. (2018). A systematic integrated analysis of brain expression profiles reveals YAP1 and other prioritized hub genes as important upstream regulators in Alzheimer's disease. *Alzheimers Dement.* 14 (2), 215–229. doi:10.1016/j.jalz.2017.08.012

Yang, Y., Wang, L., Zhang, C., Guo, Y., Li, J., Wu, C., et al. (2022). Ginsenoside Rg1 improves Alzheimer's disease by regulating oxidative stress, apoptosis, and neuroinflammation through Wnt/GSK-3β/β-catenin signaling pathway. *Chem. Biol. Drug Des.* 99 (6), 884–896. doi:10.1111/cbdd.14041

Zhang, W., Xiong, B. R., Zhang, L. Q., Huang, X., Yuan, X., Tian, Y. K., et al. (2021). The role of the GABAergic system in diseases of the central nervous system. *Neuroscience* 470, 88–99. doi:10.1016/j.neuroscience.2021.06.037

Zhou, J., Chow, H. M., Liu, Y., Wu, D., Shi, M., Li, J., et al. (2020). Cyclin-dependent kinase 5-dependent BAG3 degradation modulates synaptic protein turnover. *Biol. Psychiatry* 87 (8), 756–769. doi:10.1016/j.biopsych.2019.11.013

Zhu, J. W., Jia, W. Q., Zhou, H., Li, Y. F., Zou, M. M., Wang, Z. T., et al. (2021). Deficiency of TRIM32 impairs motor function and purkinje cells in mid-aged mice. *Front. Aging Neurosci.* 13, 697494. doi:10.3389/fnagi.2021.697494

Zhu, K., Xiang, X., Filser, S., Marinković, P., Dorostkar, M. M., Crux, S., et al. (2018). Beta-site amyloid precursor protein cleaving enzyme 1 inhibition impairs synaptic plasticity via seizure protein 6. *Biol. Psychiatry* 83 (5), 428–437. doi:10.1016/j.biopsych.2016.12.023

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 12 October 2023
DOI 10.3389/fgene.2023.1215232

OPEN ACCESS

EDITED BY
Angelo Facchiano,
National Research Council (CNR), Italy

REVIEWED BY
Amit Dubey,
Independent Researcher, Khushinagar, India
Ankush Sharma,
University of Oslo, Norway
Maurady Amal,
Faculty of Science and Technology of Tangier, Abdelmalek Essaadi University, Morocco

*CORRESPONDENCE
Boutaina Ettetuani,
b.ettetuani@uae.ac.ma

RECEIVED 01 May 2023
ACCEPTED 28 August 2023
PUBLISHED 12 October 2023

CITATION
Ettetuani B, Chahboune R and Moussa A (2023), Adjustment of *p*-value expression to ontology using machine learning for genetic prediction, prioritization, interaction, and its validation in glomerular disease.
Front. Genet. 14:1215232.
doi: 10.3389/fgene.2023.1215232

Adjustment of *p*-value expression to ontology using machine learning for genetic prediction, prioritization, interaction, and its validation in glomerular disease

Boutaina Ettetuani[1]*, Rajaa Chahboune[2] and Ahmed Moussa[1]

[1]Systems and Data Engineering Team, National School of Applied Sciences, Abdelmalek Essaadi University, Tétouan, Morocco, [2]Life and Health Sciences Team, Faculty of Medicine and Pharmacy, Abdelmalek Essaadi University, Tétouan, Morocco

The results of gene expression analysis based on *p*-value can be extracted and sorted by their absolute statistical significance and then applied to multiple similarity scores of their gene ontology (GO) terms to promote the combination and adjustment of these scores as essential predictive tasks for understanding biological/clinical pathways. The latter allows the possibility to assess whether certain aspects of gene function may be associated with other varieties of genes, to evaluate regulation, and to link them into networks that prioritize candidate genes for classification by applying machine learning techniques. We then detect significant genetic interactions based on our algorithm to validate the results. Finally, based on specifically selected tissues according to their normalized gene expression and frequencies of occurrence from their different biological and clinical inputs, a reported classification of genes under the subject category has validated the abstract (glomerular diseases) as a case study.

KEYWORDS

glomerular diseases, gene expression, gene ontology, machine learning, ETL

1 Introduction

C3 glomerulopathies (C3G) are a group of related conditions that cause kidney dysfunction (Riedl et al., 2017), characterized by the presence of glomerular deposits composed of C3 (Cook and Pickering, 2015). Many conditions in glomerular diseases (GD) have a variety of genetic/environmental causes (Coelho et al., 2019). C3G is associated with changes (mutations) in many genes. Most of these genes provide instructions for making proteins that help regulate a part of the body's immune response known as the complement system (Iatropoulos et al., 2016). This system works together as a group of proteins to destroy foreign invaders/triggers/inflammation. The complement system must be regulated, targeting all unwanted materials without damaging the body's healthy cells. A specific mutation in the complement system-related genes, like C3, ADAM19, ADAMTS13, C3AR1, C8A, CD46, CFB, CFD, CFI, CFHR (1-5), in addition to other complement system-related genes (Tsai et al., 2000; Xiao et al., 2014), risk haplotypes of CFH and CD46 have been identified that modify disease penetration and severity (Legendre et al., 2013; de Cordoba et al., 2014). In most cases, the cause of the C3G is unknown.

Many kinds of research are still devoted to discovering genes involved in specific phenotypes and diseases. Multiple gene selection techniques are defined in the literature. Whichever confidence in using a single criterion for selecting genes is not always adopted which specific used one should be diffident.

This question inspired us to consider the ranking of all criteria in the evaluation of the gene and propose a new selection of genes for transcriptomic data, focusing on the gene expression adjustment to the similarity score. Thus, the genes for each criterion would be systematically computed and validated by our algorithm. Our solution can be considered the most informative and stable method for gene prediction/selection and classification steps. In the meta-analysis process, the input of gene expression results consisted of normalized gene expression measurements. From this, a linear model fit for all genes of our transcriptomics data can be computed as an appropriate contrast function to test hypotheses of interest and to find genes with significant differential expression (DE) between different conditions (Klaus and Reisenauer, 2016) from understudied raw data (as a set of binary files in CEL format), accessible via the public repository of microarray data, the NCBI Gene Expression Omnibus (GEO) (Clough and Barrett, 2016). The raw data were chosen to be used as extracted from the source rather than processed data, although their analysis is very similar, as mentioned in (Figure 1) representing a literature review. The first step in pre-processing is data quality control. The latter is an essential step in any analytical process and a relative concept that depends on the nature of the biological sample, experimental settings, and other factors. Hence, poor-quality data can directly lead to the absence of some positive results. Moreover, we evaluated the measure of precision to reduce deficiencies over time and under varying operating conditions (Bolstad et al., 2003; Kauffmann et al., 2009; Carvalho and Irizarry, 2010). Different normalization methods have been developed in the context of gene expression analysis. A specific normalization method in microarray data analysis is crucial to ensuring accurate and reliable results. RMA (Robust Multi-array Average) was chosen over other methods such as MASS, GCRMA, PLIER, PUMA, etc.

Microarray data and RNA-Seq data are generated through different technologies and have distinct characteristics. Microarrays measure the relative abundance of pre-selected probes for known genes, while RNA-Seq directly sequences and quantifies the transcriptome, including known and novel transcripts. Initially developed for microarray data, the RMA algorithm can be applied to RNA-Seq data with different disease modalities and normalization methods, offering new insights into gene expression analysis for different biological contexts. While some concepts and principles from microarray data analysis may be relevant to RNA-Seq analysis, it is crucial to use appropriate RNA-Seq-specific methods to accurately handle the data and obtain reliable results.

The RMA algorithm was performed on our data to background-correct, normalize, and summarize the process (Okoniewski and Miller, 2006; McCall and Irizarry, 2011), offering several advantages, such as reducing the impact of extreme values. The choice of RMA over other methods depends on the specific characteristics of the dataset and the specific research questions of the analysis. RMA is often favored due to its robustness and simplicity compared to other methods like GCRMA, PLIER, or PUMA that might be more suitable.

The organization of this article is presented in two main sections, according to the following principles. The first one (methods) includes gene signature identification in connection with the search for a therapeutic target involved in the detection of differentially expressed (DE) genes, followed by a subsequent step leading to the construction of the workflow and its structure to prioritize and interact with a gene on each cluster based on Expression-Similarity-Frequency of occurrence. A second section (results and discussion) covers the interpretation of the biological/clinical results as a form of evaluation and validation of our hypothesis.

2 Materials and methods

The main focus of this follow-up study (Figure 2) is to propose and validate a novel matrix-expression-similarity-frequency consisting of a new scoring scheme based on a given combined/adjusted linear DE measurement selection in diverse experimental conditions for individual samples. Machine learning tasks then combine a mathematical algorithm with our prediction results

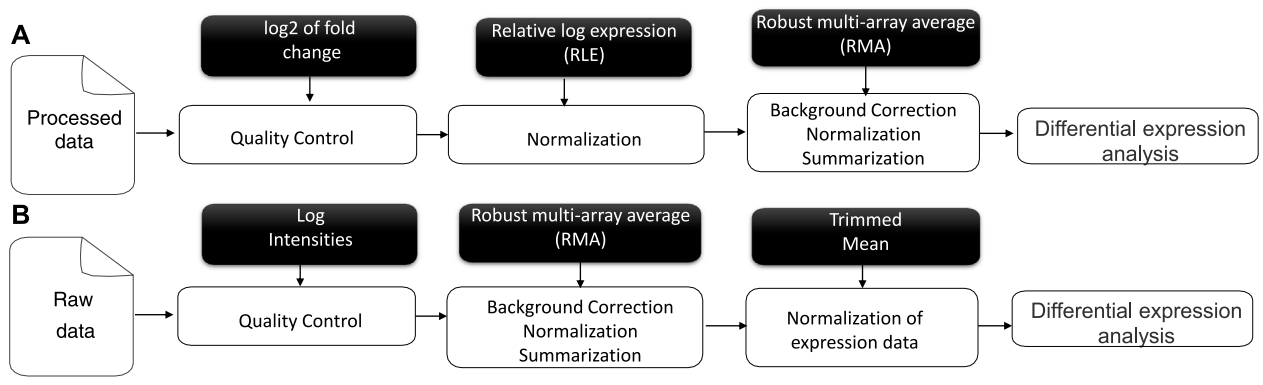


FIGURE 1
Literature review of Affymetrix microarray data pre-processing for processed **(A)** and raw data **(B)**.

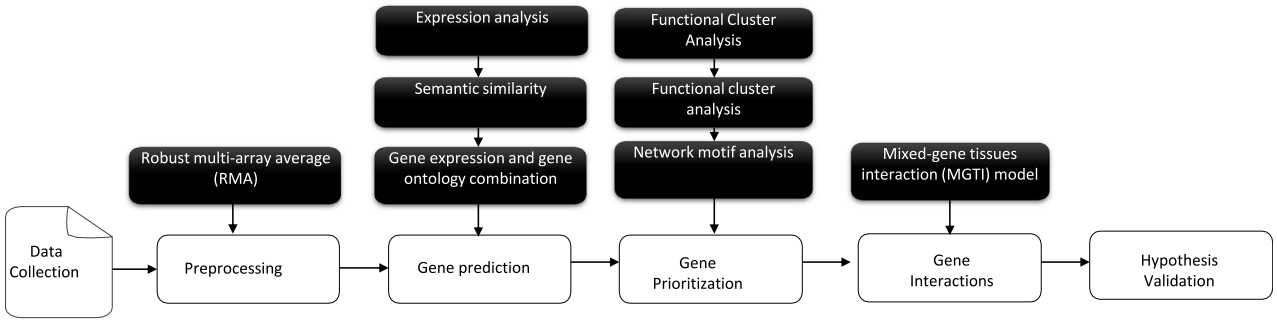


FIGURE 2
An overview of the schematic workflow of the developed approach, consisting of five components: first, transcriptomics data collection and pre-processing steps; then, gene prediction and prioritization. Finally, a human genome-scale genetic interaction model is followed by hypothesis evaluation and validation.

analysis (Tarca et al., 2007; Xu and Jackson, 2019; Sabir et al., 2021). The gene cluster lists suggested using unsupervised learning based on their normalized matrix-expression-similarity-frequency-based scores of occurrences first to find out the structure as groups of a similar category. In this case, the data contains only inputs and no desired output labels. Then, classification is the second step in which the algorithm keeps in check both the inputs and the desired outputs (a limited set of outputs) to construct co-regulation and link them into networks that prioritize candidate genes as a logistic regression tool. In addition, significant genetic interactions for a specific tissue type, genetic background, experimental stimulus, or clinical variable are detected and validated in the results.

2.1 A computational algorithm for gene correction

The statistical methods used to detect DE genes were calculated as moderated t-statistics for the microarray data based on a linear model fit by fixing three different *p*-values (p1 = 0.01, p2 = 0.001, p3 = 0.0001), which are estimated as the prior probability that a gene is DE (Jeffery et al., 2006; Sartor et al., 2006). It should be noted that when we fixed a different *p*-value distribution to our dataset, we demonstrated that the expression of these candidate genes between these *p*-values is expected to change the methods employed in GD diagnosis and prognosis. The semantic similarity computation was assumed based on the Wang method (Wang et al., 2007), and using the GOSemSim (Yu et al., 2010) package between gene products based on the information content (IC) and a directed acyclic graph (DAG) (Mazandu and Mulder, 2013). The IC-based measures depend on the frequencies of two GO terms involved in their closest common ancestor term in a specific corpus of GO annotations. The GO terms were classified into three different aspects: molecular function (MF), biological process (BP), and cellular component (CC). The molecular function is a process extended by two actions described as biochemical binding activities, referring to a protein that functions as a receptor. The second aspect is that the biological process represents an organism's specific and significant objective (Gaudet et al., 2017; Thomas, 2017). Finally, the cellular component, in terms of cellular structures and location, provides information about where a molecular process may occur. The best-match average (BMA) method calculates the average of all maximum similarities on each row and column. Furthermore, a new score for the gene similarity measures was calculated for each pair of genes as demonstrated here (Eq. 1), through which the matrix evaluates whether the mean and normally distributed score within each independent pair of genes of samples evaluates an important significance or not by introducing the matrix-similarity-based.

$$M_{simsc}(g_1, g_2) = \sum_{n=1}^{Lengthx} \sum_{n=n+1}^{Lengthx} \left(\sqrt{\frac{1}{2} * \left(\frac{\delta_1^2}{\delta_2^2} + \frac{\delta_2^2}{\delta_1^2} \right)} \right) * \left(\frac{\mu_1 + \mu_2}{2} \right)^2 \quad (1)$$

- M_{simsc} represents the similarity of the gene measurement matrix.
- Only paired groups of genes can perform the paired test.
- Based on the first two DE genes, $\sum_{n=1} \sum_{n=n+1}$ up to the full set of genes as mentioned in *Lengthx* were selected for M_{simsc}.
- The sample means are denoted as μ_1 and μ_2 for each similarity score.
- Each score is sampled independently and randomly.
- The sample standard deviations δ_1 and δ_2 are normally distributed within each of the two rows.

Gene prediction model (based on their gene expression and gene-GO similarity) represented as matrix-expression-similarity-frequency-based consisting of a new adjusted scoring scheme of the score and frequencies (as results) for a given linear DE measurement selection results mixed with their scores and frequencies of occurrence of matrix-similarity-based, which yield the final association score. The genes with the highest scores were first selected and improved to serve as inputs for the machine learning steps.

$$M_{CombSc} = n_{expr} * M_{expr} + n_{sel} * M_{simsc} \quad (2)$$

- A number of expressed genes n_{expr} provided with fixed *p*-values M_{expr} were combined into the expression matrix.
- A number of correlated genes n_{sel} were combined into the semantic similarity matrix M_{simsc}.

2.2 Prioritization of candidate genes based on machine learning

Machine learning methods (Ganggayah et al., 2019; Zhang, 2020; Ke et al., 2021) were first performed as clustering algorithms (Chou et al., 2007; Maulik et al., 2009), building a mathematical model from the normalized final scores of our matrix-expression-similarity-frequency-based (Eq. 2) to segment them into k clusters to understand biological processes in addition to molecular functions. Each cluster represents a group of similar observations performed using the Ward method and Euclidean distance for a given value of k as a possible solution (i.e., high intra-class similarity). Although objects from different clusters are as dissimilar as possible (i.e., low inter-class similarity), to improve the initial partition obtained from hierarchical clustering. The algorithm can stop when the assignment of genes to clusters no longer changes or when the specified maximum number of iterations has been reached (Gry et al., 2009; de Winter et al., 2016). Justifying the choice of distance metric and comparing the results with several known distances are essential steps that can significantly impact the results and ensure the robustness and reliability of the machine learning model, including candidate gene classification. The appropriateness of the Euclidean distance depends on the nature of the data and the problem at hand for candidate gene classification:

The classical methods for distance measures are Euclidean, Manhattan, and correlation-based distances, used for gene expression, such as Pearson correlation distance and Spearman correlation distance. The correlation-based distance considers two similar objects if their features are highly correlated. The convergence between the Manhattan distance and Euclidean distance for gene expression depends on some specific characteristics and distributions of gene expression data. The lasts are widely used to measure the similarity or dissimilarity between samples based on their gene expression profiles. In summary, the choice between Manhattan and Euclidean distances for gene expression data should consider the vector space (dimensionality) of the data and the number of genes being analyzed. When the number of genes is relatively small due to the pre-processing and DE analysis, both Manhattan and Euclidean distances may behave similarly, especially if the genes are highly correlated or there is little variability in the data. Additionally, it was essential to experiment with different metrics and compare their performance using appropriate evaluation techniques such as cross-validation to select the best distance metric, which is why we tried all the discussed methods, and the results are accessible in our two previous published papers (Ettetuani et al., 2019; Ettetuani et al., 2020). Finally, the Euclidean distance assumes that all features have the same scale and are equally important. Euclidean distance treats each feature independently, without considering correlations between them.

Further, each gene cluster list was exposed to the (hgu133a,hgu133plus2) database of *Homo sapiens* as a direct mapping of a gene symbol to a vector containing the corresponding Entrez gene identifier (Smedley et al., 2009; Yu et al., 2012), then implemented in a hypergeometric model to assess whether the number of selected genes is linked/associated with the pathogenesis of the diseases (Yu, 2012; Fabregat et al., 2018). All enriched terms were associated with their enrichment scores (p-values) as a first step of supervised learning, allowing the possibility to cross from high-level concepts to detailed pathway diagrams showing bio-molecular events using the groupGO(), and enrichGO() functions (Sidiropoulos et al., 2017; Wu et al., 2021). Genetic variation was also studied through our first step of supervised learning, which is the genome-wide association study (GWAS); based on the GWAS catalog used to tag variation across the genome and enable investigations to identify causal variants (Johnson et al., 2010; Scharf et al., 2013; Butler et al., 2017; Garfield, 2020) and variant-trait associations mapped to their chromosomal positions in the human genome. Many computational approaches have been developed to support the identification of the most promising candidates (Zolotareva and Kleine, 2019). oPOSSUM (Ho Sui et al., 2007) web applications containing a great variety of the conserved non-coding regions of the promoters/enhancers were used to select the interaction between our candidate's genes, whose interactions between genes and transcription factors (TFs) were a major to understand gene regulation and the origin of complex protein components (Suravajhala and Benso, 2017).

To facilitate the prioritization of causative genes as the second step of supervised learning and based on algorithmic tools, we illustrate a new gene prioritization model for candidate genes based on the logistic regression method (Lee et al., 2018; Zhang et al., 2018; Christodoulou et al., 2019; Nusinovici et al., 2020). Gene prioritization schemes boost the power to identify the most promising phenotype-associated among those clusters under normal conditions in different tissues [where each gene can have a normalized expression score in the tissue expression database (Palasca et al., 2018)].

$$Score_{prioritization}(Gene_i \mid Tissue_t) = \mu_0 + \mu_1 X_1 + \mu_2 X_2 + \mu_3 X_3 + \cdots + \mu_n X_n \tag{3}$$

- μ_0 is the tissue-specific means of expression for a given gene with fixed tissues.
- $\mu_1, \mu_2, \mu_3, \ldots \mu_n$ are the means of (frequencies of occurrence) for target genes for each process.
- X1, X2, X3, , Xn represent the normalized expression values of biological processes, GWAS, TFs, etc.
- The parameters in logistic regression cannot simply be replaced by average values, especially when the output is a probability-related value. The parameters in logistic regression represent the relationship between the input variables (in this case, gene expression means) and the probability of normalized expression values based on biological processes.
- Replacing the parameters with average values would be highly unusual and would be approved later by the methodology results.

Following the hypothesis that genes underlying similar tissues will share functional and phenotypic characteristics, we incorporated logistic regression for any training genes that need to be prioritized (Eq. 3). When applying logistic regression, it is generally recommended to split the available data into three separate sets: a training set to estimate the model parameters, a validation set to tune the model hyperparameters and assess its performance, and a

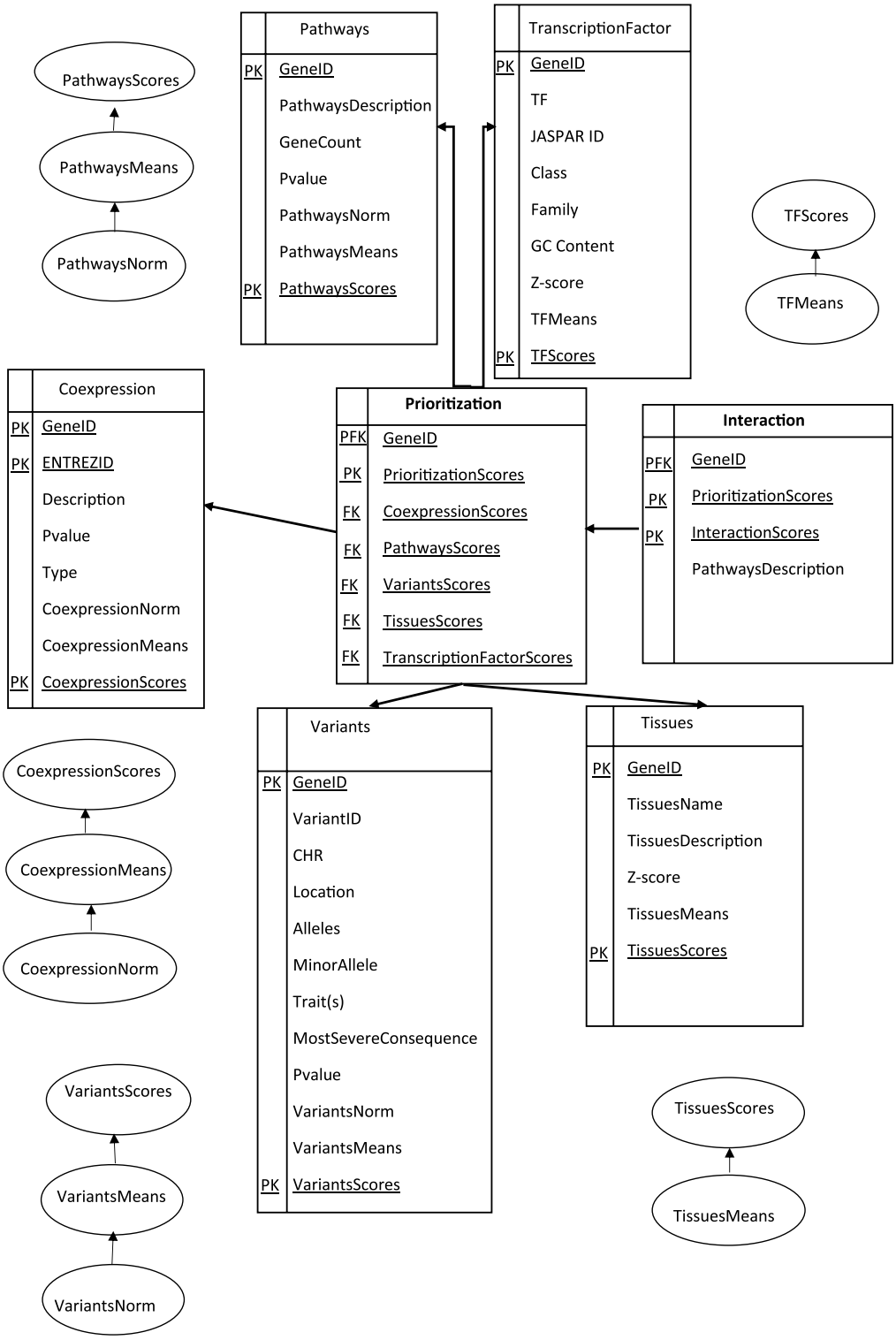


FIGURE 3
Data Warehouse Star Schema represented in SQL: A fact table representing the prioritization and interaction model for different dimension tables. Data were collected based on Eq. 3, stored in SQL databases and structured in tables with predefined schemas.

test set. The formulated model arranged the candidate genes (i) in the order of their tissues (t) first to be sure they were associated with the pathology. The algorithm took two inputs: a collection of evidence sources defining a phenotype/trait of interest, and enhancer/promoter interaction information, extracted for a given gene, linked to each other with a normalized score reflecting the "likelihood" for each gene to be responsible for the phenotype. Then a second factor; frequencies of occurrence of each entire represents a

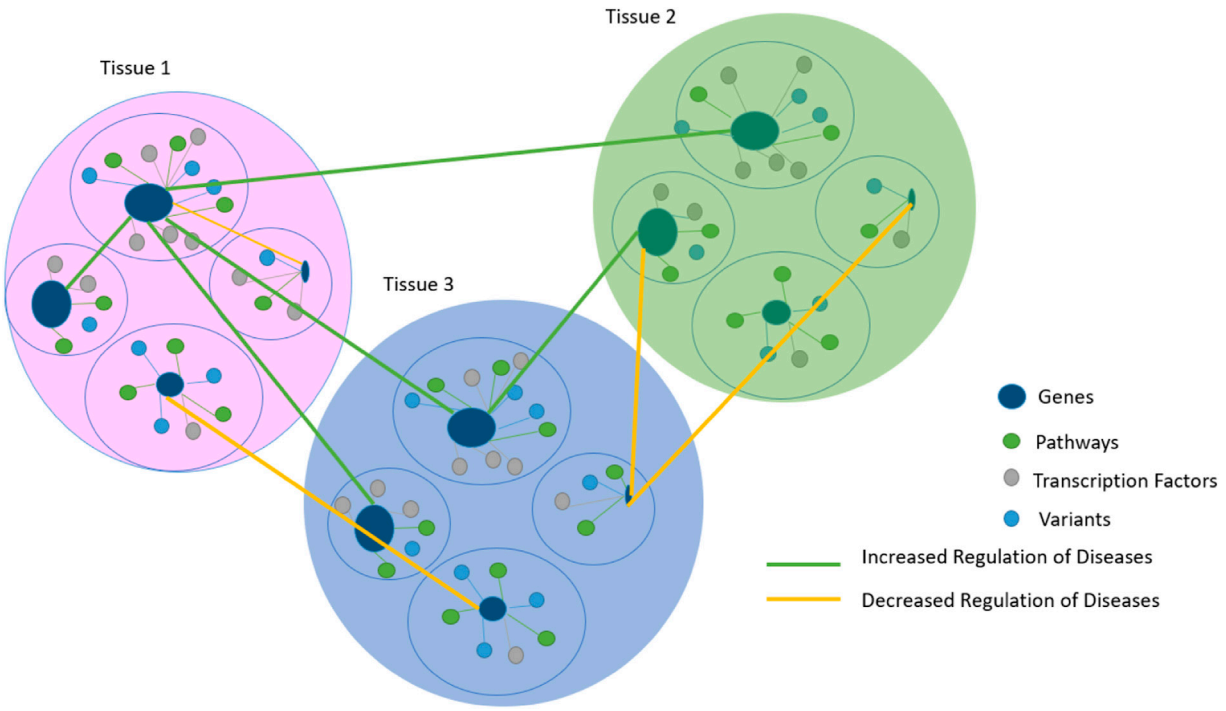


FIGURE 4
Gene prioritization results are based on tissues that are related to each other by regulating disease links. The prioritization scheme in Figure 4 is used to facilitate the comprehension of different Supplementary Figures and serves as a starting point for the sequence in Figure 3.

mean of the prioritization factor. The output of the algorithm resulted in a list of candidate genes arranged according to the calculated scores for each tissue.

Overall, the process of extracting, transforming, and loading ETL data from homogeneous or heterogeneous sources was established to access our data (Biswas et al., 2020; Biswas et al., 2018). In short, it was an essential component in cleansing, customizing, reformatting, integrating, and inserting the prerequired data (Greiff et al., 2015; Tecnico, 2015). In this paper, we tried to navigate through our adjusted genes to conceptualize the ETL processes, as shown in (Figure 3), first modeling a prioritization tool, then modeling genetic interaction constructs into proper storage (format/structure) for querying and analysis.

The Primary Key (PK) uniquely identified each row in an interactive table. In the star schema, the dimension tables were typically designed with surrogate keys that were independent of the source data. These surrogate keys were used to establish relationships between the schema dimensions and the fact table(s). A proper star schema design often includes a central fact table that contains the Primary Foreign Key (PFK) of the dimension tables, along with the numerical values (facts) associated with those dimensions as a combination of a primary key and a foreign key in a database. It was used to establish a relationship between different tables, allowing and creating a link between a table to reference another table's primary key. The dimension tables have descriptive attributes that provide additional information about the dimensions. A collection of candidate geneIDs arranged according to a specific tissue (kidney and urine and Immune system and blood and Embryonic dev) with the highest information of the gene related to specific biological pathway terms, the gene-variant-trait associations among the GWAS catalog, and the genes and transcription factors (TFs) interaction are the most efficient approaches to representing genomic data. A Foreign Key (FK) column in the prioritization table refers to the primary key in another table to create our study a connection between each calculated score-related data, allowing for data retrieval and enforcing referential integrity.

One way to understand these terms and enforce data integrity involves defining the structure and relationships based on primary and foreign keys (PK and FK). Primary keys uniquely identify each record in a table, while foreign keys establish relationships between tables to design, build, and manage efficient and reliable databases, preventing invalid or inconsistent data from being inserted or updated.

2.3 Gene interaction

Genetic interactions of omics data refer to a combination of pairs of genes in different tissues of fixed clusters, as shown in (Figure 4), whose contribution to a phenotype between specific variants in complex traits and tissues remains a challenge (Gomez-Cabrero et al., 2014; Vasaikar et al., 2018; Subramanian et al., 2020).

A novel algorithm-based model called Mixed-Gene Tissue Interaction (MGTI) was developed based on previous data. Gene prioritization scores were first calculated, and then significantly interacted mixed genes between the selected tissues were

identified using (Eq. 4), reflecting the strengths of regulatory interactions to understand the etiology of our glomerular disease as a use case.

$$Score_{inter}(g_1, g_2) = \frac{\log_{10}\left(S_p(g_1)\right) + \log_{10}\left(S_p(g_2)\right)}{2 + \log_{10}\left(\max\left(S_p(g_1, g_2)\right)\right)} \quad (4)$$

- $Score_{inter}(g_1, g_2)$ represents an interaction score for each pair of genes.
- Each gene is represented by the prioritization score S_p, which is based on multiple calculated scores as mentioned in Eq. 3.
- $\max(S_p(g_1, g_2))$ represents the maximum gene prioritization score between the two selected genes.
- The interaction scores are between (0, 1).
- When the interaction scores are ≅ (0.31, 1), the MGTI model reflects an increased regulation of disease.
- When the interaction scores are ≅ (0.1, 0.3), the MGTI model reflects a reduced regulation of disease, where we should interact more with other possible genes.
- There is no gene score between ≅ (0, 0.1), because each gene is selected following a selection score and then validated to have n minimum pathway information in the prioritization section.

Moreover, as the number of interacting genes increases, traditional statistical methods are limited in their ability to identify interacting genes in high-dimensional data (Yang et al., 2011; Gordon et al., 2020; Sun et al., 2020).

3 Results and discussion

The detection of significant genetic interactions was focused on large-scale studies based on a selection of gene expression mixed with multiple similarity scores of their gene ontology (GO) terms. As well as their sources of biological became widely adopted. Our algorithm adjusted them to evaluate the regulation and link them into networks that prioritize candidate genes as classification by applying machine learning techniques related to glomerular diseases (GD).

The spectrum of glomerular diseases is defined by the abnormal control of complement cascade activation, whose actions are considered part of the innate immune system, procuring an immune complex deposition of fragments of C3 in glomeruli (Pickering et al., 2013). GD often results in kidney damage, the cause of which is unknown.

The first thing to do to measure the level of transcriptional genes was to validate the information stored in the raw data measurements (Dalman et al., 2012; Kharchenko et al., 2014), which were analyzed as described in our paper (Ettetuani et al., 2019), published by the ACM organization, Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society, as a validation process of the information stored in expression sets. In this meta-analysis approach, each experiment was first analyzed separately, and all the results were then combined based on their primary statistics (p-values) (Walsh et al., 2015). Here, we fit the linear model for all genes and defined appropriate contrast functions to test hypotheses of interest to find genes with significant DE within each condition.

3.1 Experimental data

Based on the PubMed database, for Data retrieval, summarizing and comparing topics according to their frequencies of occurrence (Rani and Ramachandran, 2015; Gusenbauer and Haddaway, 2020), to broaden or/and narrow a search Kovalchik (2015), and to exclude unwanted search terms/concepts from a specific speech as “glomerulopathies” “diabetic kidney disease” “tumor Nephrectomy” “diabetic nephropathy” “focal segmental glomerulosclerosis” “rapidly progressive glomerulonephritis” “minimal change disease” and “membranous glomerulonephritis.”

In addition, five datasets (Table 1) were extracted consisting of human kidney biopsies of patients are used in our analysis, providing a comparison of the glomerular transcriptome for multiple profiles as the adult-onset steroid-sensitive focal segmental glomerulosclerosis and minimal change disease (Tong et al., 2015), transcriptomic and proteomic profiling reveals insights of mesangial cell function in patients with IgA nephropathy (Liu et al., 2017), glomerular transcriptome from subjects in the NEPTUNE cohort (Mitrofanova et al., 2018), and shared molecular targets in the glomerular transcriptome from patients with nephrotic syndrome and ANCA-associated vasculitis, and glomerular transcriptome from European renal cDNA bank subjects and living donors (Grayson et al., 2018).

3.2 Genetic contributions and their statistical impact on the estimation of predictive models of gene

The results of each number of DE genes were extracted as mentioned in (Figure 1), combined, and estimated in a uniform distribution for the p values corresponding to five different datasets (Table 1). Our study was performed based on P3 (p-value) for the simple reason that the significance of other selected genes (P2 and P1) is correlated with P3, as visualized in the corresponding figure (Supplementary Figure S1). The results of the ontology analysis are represented as the distance of (dis)/similarity between our list of genes in the interval of (0, 1). When $Sim(g_i, g_j) = 1$, it means that $(i = j)$, and when $Sim(g_i, g_j) \cong [0.6, 1]$, it means that the precision of semantic similarities over genes in $GO(g_i, g_j)$ is more significant and related to common pathologies. Then, when $Sim(g_i, g_j) \cong [0, 0.5]$; is referred to the precision of semantic similarities over genes in $GO(i, j)$, which may be significant or related to other common pathologies. However, many genes had a high score on the expression in parallel to a low score of GO (or the reverse). Their classical functional analysis in the literature is based only on the selection of genes from experiments while searching for their dis/similarity score is the input of the classification/regulation analysis in most cases. This problem inspired us to combine/adjust both scores (expression and similarity-based GO annotation) into a single formula that gives us a better chance to predict genes related to our pathologies based on their occurrence frequencies. Consequently, we propose a novel gene selection method by introducing a novel matrix-expression-similarity-frequency-based. Different threshold values give different levels of sensitivity and specificity. Whether the low threshold represented with red color refers to a false positive and the highest with blue color refers to a true positive, as fixed for the validation of our study. This makes it more likely to be specific with more high positives

TABLE 1 Description of the five datasets, in which all types of data were transcribed by array, with their fixed *p*-values when annotated to the human databases, and selected matrix-expression-similarity-frequency-based.

Id	Status	Number of samples	Number of genes	P1	P2	P3
E-GEOD-69814	4-Jan-17	11	32,321	10	22	71
E-GEOD-93798	3-Jul-17	42	54,675	11	22	34
GSE108113/E-GEOD-104066	26-Jun-19	76	1,416,100	3	11	53
GSE108113/E-GEOD-108109	17-Jul-18	111	1,416,100	6	23	68
E-GEOD-104948	24-Jan-18	196	76,958	22	51	183
Total genes				52	129	415
Not duplicated				52	129	412
Annotated				33	78	209
Selected				18	45	100

against more sensitive with more low positives, as shown in the corresponding figures (Supplementary Figure S2; Supplementary Tables S1, S2) with a given fixed threshold of observations.

3.3 Evaluation and prioritization of tissue using a regression model

An estimating matrix-expression-similarity-frequency-based score for each gene was the input for computing the clustering algorithms (unsupervised learning) to understand biological processes along their molecular functions. Four clusters were found to represent a group of similar observations. To search for shared functions as the first step of supervised learning, all selected genes (in terms of the fixed threshold provided by our prediction analysis) were linked to different databases as enrichment, traits (non-coding variants), tissue databases, and the over-represented conserved transcription factor binding sites based on their score of %GC content. GC content is a measure of the proportion of nucleotides in a DNA sequence that are either guanine (G) or cytosine (C). It is often expressed as a percentage. "GC" is one of the factors considered when identifying over-represented transcription factor binding sites (TFBS) in co-expressed genes Gao et al. (2022). Such analyses generate a mixture of data that requires a biological interpretation. The majority of these genes fall under (kidney, blood, urine, immune system, and embryonic) tissues.

In the hypothesis interpretation and validation section for the GD, we were based on a specific entire to generate a list of the highest information of the gene related to particular biological pathway terms such as the regulation of inflammatory response, regulation of acute inflammatory response, regulation of protein processing/maturation, positive regulation of glomerulus development, of glomerular mesangial cell proliferation, of the adaptive immune response, and complement activation, etc., as shown in the Circos plot (Supplementary Figure S3) as one of the most efficient approaches to visualize genomic data; it allowed us to easily represent all this information on a single plot.

Before evaluating the prioritization model as the second step of supervised learning for our gene clusters, we reanalyzed the highest (more conservative) and lowest threshold (more sensitive) to be sure and to validate the score of a prediction selection, while also justifying that genes selected with the lowest threshold score are not sufficiently expressed in the tissues and/or traits and/or biological processes of interest. A heatmap-like functional classification plot (Supplementary Figure S5) was used to visualize the most significant terms (with the terms expected from the literature), according to some scores, while simultaneously visualizing the sub-ontologies of causative genes as (EGR1, IL33, BMP2, SLAMF8, etc.), in which we filtered/selected the most dominant terms according to our pathology (GD), as well as their GO annotations include (kidney vasculature development, regulation of cell activation, inflammatory, immune effector, adaptive immune, glomerulus, and glomerular mesangial cell proliferation development, etc.).

A bar plot (Supplementary Figure S4) was used to visualize the gene-variant-trait associations among the GWAS catalog used for the most dominant terms such as (complement C3, C4, C7 measurement, and serum IgE/IgA measurement, c-reactive protein measurement, nephrotic syndrome, immune system disease, tuberculosis, glomerular filtration rate, chronic kidney disease, urinary metabolite measurement, C-reactive protein measurement, glomerular filtration rate, etc.). In addition, many genes such as (TNXA, FCER1A, NME3, FMOD, BTG2, PTGER4, AXL, CYP1A2, CYTL1, BHLHE40, IFI16, SPON1, ETNPPL, COL14A1, ITGAV, MYOZ2, CAMK2A, SORT1, RANBP1, etc.) showed some trait association.

Finally, a PieChart plot (Supplementary Figure S6) was used to represent the mean expression of genes in the selected tissues (kidney, renal cancer cell, immune system, urine, blood, blood vessel, blood plasma, hematopoietic stem cell, parenchyma, uroepithelium, HEK 293 EBNA cell, HEK 293ET cell, HK 2 cell).

The gene prioritization model could be formulated as follows (two parts): arrange candidate genes in the order of their normalized scores and frequencies of occurrence from matrix-expression-similarity-frequency-based, then to their normalized scores and frequencies of occurrence for categorical tissues, biological processes, TFs, and variants. The logistic regression model (parametric regression) was performed on categorical data to prioritize the dependent variable using a given set of independent variables to solve classification problems. In logistic regression, linear connections between the dependent and independent variables are not needed. However, there should be no collinearity between the independent variables. As discussed above, logistic regression was used to classify the elements of each cluster under different tissues by calculating the mean of the normalized expression of

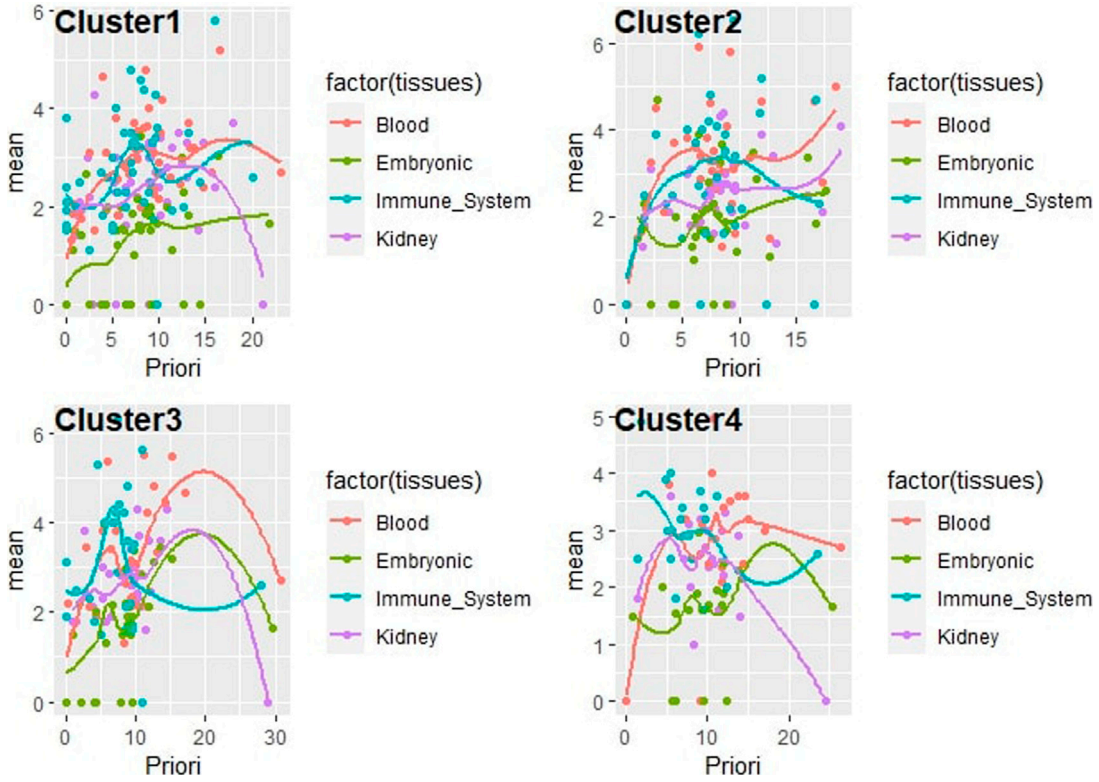


FIGURE 5
Correlation plot line of the prioritization model as logistic regression for gene clusters consisting of a new scoring including expression, similarity, and frequency from mean expression of various tissues of interest, under each element of the set (phenotype/trait, and enhancer/promoter).

each set. A confidence regression line provides a representation of the uncertainty in the sense that, among a cluster of genes, we can prioritize the most promising ones based on their prioritization scores on each selected tissue, as visualized in (Figure 5). Many redundant genes were more promising by applying the priority model to fixed tissues, as noted in the table (Table 2). This has only one interpretation, which has great value in the expression in the different tissues analyzed. The process of arranging all possible prioritizing disease-causing genes based on their logistic regression has shown that consistent genes may reside in distinct pathways and affect the promoter/regulatory region of location-related organisms. Based on PubMed resources, 972 abstracts were extracted and approved that included 11 genes (COL4A5, EGR1, GDF15, CPE, CASK, NT5E, JUN, AXL, CCL3, IL33, ITGAV) from our candidate genes that have already been reported in previous studies on GD.

3.4 Genetic interactions bridging transcription factors and pathways in genome-wide association studies

Traditional statistical methods consider gene-gene interactions and estimate interactions among only a fixed or small number of phenotypes information with significant main effects. However, our MGTI algorithm-based model can be applied when the data are highly dimensional (many attributes or independent variables) or when interactions between more than two tissues may play a role in human disease etiology and regulation data mining analysis. To perform ETL (extract-transform-load) operations, a vector of SQL commands was used to select data based on a specific entity, focusing on choosing the best score of gene-gene interaction and supporting and validating the hypothesis validation. Representatives Table 3 show some random gene-gene interaction algorithm tools based MGTI model.

Based on the interaction between IL33 and RANBP1 genes, which are expressed in the immune system and kidney tissue, respectively, they contribute to indirect immune-mediated renal disease, often many acute forms of renal disease, and play a central role in the progression of chronic kidney disease. When the interaction between these genes is high, the network interaction leads to new diagnostic methods and treatment solutions to inhibit further progression and promote appropriate tissue repair.

A glomerular filtration rate, expressed between urine and kidney tissue, checks how well the kidneys are working. The kidneys are two organs on either side of the spine, near the waist. They have tiny filters called glomeruli. These filters remove waste and extra water from the blood and get rid of them through urine. When the kidneys are damaged by kidney disease, they cannot filter blood as fast as they should. The interaction between the MYOZ2 and SORT1 genes can be used to check for kidney disease by measuring how much blood is filtered in the kidneys and how much C-reactive protein is increased.

Finally, the gene interaction results encode a superfamily of proteins that plays a role in transcriptional expression, whether

TABLE 2 The most promising genetic head data prioritization-models for the four clusters.

Cluster N 1				
Kidney	Urine	Immune system	Blood	Embryonic dev
IL33	SPON1	SPON1	SPON1	SPON1
RANBP1	IL33	IL33	IL33	IL33
ANK1	RANBP1	RANBP1	RANBP1	RANBP1
MYOZ2	ANK1	ANK1	MYOZ2	ANK1
MEOX1	MYOZ2	MYOZ2	PALB2	PALB2
CACNA1G	MEOX1	MEOX1	CACNA1G	MYOZ2
PALB2	CACNA1G	CACNA1G	ANK1	MEOX1
Cluster N 2				
Kidney	Urine	Immune system	Blood	Embryonic dev
BMP2	BMP2	COL15A1	BMP2	BMP2
COL15A1	COL15A1	ITGAV	COL15A1	COL15A1
ITGAV	ITGAV	BMP2	ITGAV	ITGAV
ATP13A3	ATP13A3	ATP13A3	ATP13A3	ATP13A3
TNFSF10	TNFSF10	TNFSF10	TNFSF10	TNFSF10
FGF7	FGF7	FGF7	FGF7	FGF7
Cluster N 3				
Kidney	Urine	Immune system	Blood	Embryonic dev
SPON1	SPON1	SPON1	SPON1	SPON1
RET	RET	RET	RET	RET
FOS	FOS	FOS	FOS	FOS
SFRP4	SFRP4	SFRP4	AXL	AXL
AXL	NME3	P2RY14	SFRP4	SFRP4
NME3	CA10	NME3	GDF15	GDF15
CA10	BTG2	CA10	NME3	CCL4
Cluster N 4				
Kidney	Urine	Immune system	Blood	Embryonic dev
SPON1	SPON1	SPON1	SPON1	SPON1
MYOZ2	MYOZ2	MYOZ2	TDG	TDG
PALB2	PALB2	PALB2	PALB2	PALB2
SORT1	RAP1B	RAP1B	SPRY1	MYOZ2
RAP1B	NME3	P2RY14	MYOZ2	SORT1
TDG	SORT1	NME3	RAP1B	SPRY1

ligands of this family bind various enzyme binding and receptors/initiators leading to recruitment and activation of family transcription and signaling factors regulating the level and stability of gene expression. The encoded proteins possess different motifs composed of intracellular and extracellular domains. Several cellular functions may be involved in multiple cell types and various tissues. While alternative splicing results in multiple transcript variants of these candidate genes. While alternative splicing results in multiple transcript variants of these candidate genes.

Based on a suite of queries such as network analysis, functional enrichment analysis, and cross-validation with network analysis that represent different types of analyses performed on the data warehouse comparing the results to previously validated ones.

TABLE 3 Global summary of the mixed-gene tissue interaction (MGTI) model based on gene prioritization results and randomly selected genes from the results.

Cluster N 1		
	IL 33: Interleukine 33	Genes 2: RANBP1: RAN binding protein 1
Mean-coExp (adjusted)	0.17	0.87
Mean-tissues	5.8	2.4
Description-tissues	"Immune system"	"Kidney"
Mean-BP	0.59	-
Description-BP	"Regulation of: inflammatory response." of immune effector process."response to external stimulus." cytokine production. "inflammatory response." adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains." proteolysis." adaptive immune response." neuroinflammatory response." of cell activation."	-
Mean-TFs	0.77	1.54
Description-TFs	"Myc", "CEBPA", "USF1", "TBP"	"Zfx", "MYC::MAX"
		"Mycn", "RXRA::VDR"
	"FOXO3", "Arnt::Ahr", "Sox5"	"ZNF354C″, "Myc", "Egr1"
Mean-GWAS	4.75	0.4
Description-GWAS	"Acute kidney injury"	"Urinary metabolite
		Measurement"
		"Chronic kidney disease"
		"Blood protein measurement"
		"Platelet count"
		"Serum IgG glycosylation
		Measurement"
Prioritization score	16.53	16
Interaction score: 0.75		
Cluster N 2		
	Genes 1: ITGAV: Integrin alpha V	Genes 2: FCER1A: Fc fragment of IgE
Mean-coExp	1.05	1.69
Mean-tissues	3.35	1.9
Description-tissues	"Hematopoietic stem cell" "Parenchyma"	"Kidney"
Mean-BP	0	0
Description-BP		
Mean-TFs	1.18	0.96
Description-TFs	"IRF2", "Sox17", "Esrrb", "ARID3A″	"NFYA", "NHLH1", "Gata1"
	"RORA1", "SOX9"	"Myf", "FEV"
Mean-GWAS	2.38	0.003
Description-GWAS	"Urinary albumin to creatinine ratio"	"C-reactive protein measurement"
	"Microalbuminuria"	"Leukocyte count"
		"Serum IgE measurement"
Prioritization score	16.01	6.6

(Continued on following page)

TABLE 3 (*Continued*) Global summary of the mixed-gene tissue interaction (MGTI) model based on gene prioritization results and randomly selected genes from the results.

Cluster N 1		
	IL 33: Interleukine 33	Genes 2: RANBP1: RAN binding protein 1
Interaction score: 0.63		
Cluster N 3		
	Genes 1: PHLDA2: Pleckstrin Homology Like Domain Family A Member 2	Genes 2: BTG2: BTG Anti-Proliferation Factor 2
Mean-coExp	0.96	0.23
Mean-tissues	2.4	3.15
Description-tissues	"Kidney"	"Blood"
		"Blood vessel"
Mean-BP	-	-
Description-BP	"Regulation of binding"	-
	"Regulation of protein binding"	
Mean-TFs	1.19	1
Description-TFs	"GABPA", "NFE2L2"	"INSM1", "NFYA", "Tcfcp2l1"
	"IRF1", "ELK4"	"CREB1", "Zfx", "Klf4"
	"Egr1", "FEV"	
	"RELA"	
Mean-GWAS	-	0.49
Description-GWAS	-	"Mean corpuscular hemoglobin"
		"Red blood cell distribution width"
		"Immunoglobulin isotype switching
		Measurement"
		"Multiple sclerosis"
Prioritization score	10.17	9.26
Interaction score: 0.65		
Cluster N 4		
	Genes 1 : MYOZ2 : Myozenin 2	Genes 2: SORT1: Sortilin 1
Mean-coExp	1.8	3.15
Mean-tissues	1.5	1.7
Description-tissues	"Kidney"	"Urine"
Mean-BP	-	-
Description-BP	-	-
Mean-TFs	1.08	1.5
Description-TFs	"PLAG1", "MEF2A″	"RORA2", "MEF2A″, "RREB1"
	"ELF5", "Myb", "FOXD1"	"NFYA", "CREB1", "Gfi"
Mean-GWAS	2.27	0.06
Description-GWAS	"Body height" "glomerular filtration rate" "serum IgE measurement""renal transplant outcome measurement" "donor genotype effect measurement"	"Blood protein measurement" "C-reactive protein measurement" "glomerular filtration rate" "C-reactive protein measurement" "creatinine measurement" "body height" "chronic kidney disease"
Prioritization score	13.9	11.7
Interaction score:0.7		

The currently proposed search validation approach to gene prioritization and data warehouse results is based on selecting these improved best interaction scores to solve the sequencing analysis. The best-improved scores were considered biomarker modules to detect and rank novel forms of the activation of glomerular disease genes.

The early diagnosis and prognosis of any type of disease are correlated with the need for bioinformatics tools, as they can facilitate the subsequent clinical management of patients, in which the follow-up analysis can be applied to other types of data such as cancer data, etc. Finally, our global objective was to incorporate/develop a Shiny web-based application as an R framework designed to help and facilitate users' navigation under "Shiny apps" as to test our algorithms tools: gene prediction, prioritization, and interaction.

4 Conclusion

The long-term goal of this research is to improve our understanding of the molecular/biological mechanisms of activation and regulation of a set of novel/common genes implicated in our pathology in different target cells. To address the above issues, we proposed three contributions combining and adjusting multiple similarity scores of gene expression gene ontology terms based on similarity scores, which were explained by our algorithm as essential prediction tasks for evaluating the regulatory pathways. Then, machine learning techniques to prioritize candidate genes were demonstrated. Finally, some significant genetic interactions were detected as a validation of the results by applying our algorithm model.

The linked resources of biological/clinical data-based expression profiles (adjusted scores) are used to validate molecular biology research. Experimental validation of all associations facilitates the discovery of causative genes related to glomerular diseases (GD). Genes such as EGR1, IL33, BMP2, and SLAMF8 have their GO annotations such as kidney vasculature development, regulation of cell activation/inflammation/immune effectors/adaptive immune/glomerulus/glomerular mesangial cell proliferation] development, etc.

Other genes such as TNXA, FCER1A, NME3, FMOD, BTG2, PTGER4, AXL, CYP1A2, CYTL1, BHLHE40, IFI16, SPON1, ETNPPL, COL14A1, ITGAV, MYOZ2, CAMK2A, SORT1, RANBP1, in which their variants information include complement a set of C(3,4,7) protein measurement, serum IgE/IgA measurements, c-reactive protein measurement, nephrotic syndrome, immune system disease, tuberculosis, glomerular filtration rate, chronic kidney disease, etc.

The latter enables a rapid interpretation of complex gene expression studies and illustrates an overview of a computational model for gene prioritization and their genetic interactions.

Finally, the majority of our prioritized genes fall under transcription co-factor binding, regulation of glomerular mesangial cell proliferation, regulation of adaptive immune response, complement activation, etc.

As a future area of study, new deep neural network algorithms are proposed to summarize the clustering of genes based on their regulatory pathway results, not only on their expression. This can be challenging due to the many gene ontology (GO) terms connected as directed acyclic graphs. This new area of analysis can bring about changes in molecular, cellular, and biological processes. Finally, we demonstrated that genotype-phenotype associations can be adjusted and updated by using our feed and back-propagation algorithms, which minimize the loss function for gene ontology (GO) terms.

Data availability statement

The datasets and R Shiny application presented in this study can be found in online repositories: https://github.com/boutaina-ettetuani/Glomerular-Diseases-Analysis Accession: (GSE69814, GSE93798, GSE108113, GSE104948).

Author contributions

BE contributed to R-scripting for the development of GenePPI structural analyses, statistics, and the implementation of the method. RC contributed to the interpretation of the results. AM contributed to the interpretation and implementation of the method and structural analysis. AM and RC contributed equally to the literature search. All authors contributed to the article and approved the submitted version.

Acknowledgments

BE would like to thank AM and RC for their supervision and support in this work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1215232/full#supplementary-material

References

Biswas, N., Sarkar, A., and Mondal, K. C. (2020). Efficient incremental loading in etl processing for real-time data integration. *Innovations Syst. Softw. Eng.* 16, 53–61. doi:10.1007/s11334-019-00344-4

Biswas, N., Sarkar, A., and Mondal, K. C. (2018). "Empirical analysis of programmable etl tools," in International Conference on Computational Intelligence, Communications, and Business Analytics, Kalyani, India, July 27–28, 2018, 267–277.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193. doi:10.1093/bioinformatics/19.2.185

Butler, J. M., Hall, N., Narendran, N., Yang, Y. C., and Paraoan, L. (2017). Identification of candidate protective variants for common diseases and evaluation of their protective potential. *BMC genomics* 18, 575–611. doi:10.1186/s12864-017-3964-3

Carvalho, B. S., and Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26, 2363–2367. doi:10.1093/bioinformatics/btq431

Chou, J. W., Zhou, T., Kaufmann, W. K., Paules, R. S., and Bushel, P. R. (2007). Extracting gene expression patterns and identifying co-expressed genes from microarray data reveals biologically responsive processes. *BMC Bioinforma.* 8, 427–516. doi:10.1186/1471-2105-8-427

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* 110, 12–22. doi:10.1016/j.jclinepi.2019.02.004

Clough, E., and Barrett, T. (2016). The gene expression omnibus database. *Stat. Genomics Methods Protoc.* 1418, 93–110. doi:10.1007/978-1-4939-3578-9_5

Coelho, S. S., Fernandes, A. R., Soares, E., Valério, P., Matos, B., Romão, H., et al. (2019). A new complement factor b mutation associated with crescentic c3 glomerulopathy; a case report. *J. Nephropathol.* 8, 30. doi:10.15171/jnp.2019.30

Cook, H. T., and Pickering, M. C. (2015). Histopathology of mpgn and c3 glomerulopathies. *Nat. Rev. Nephrol.* 11, 14–22. doi:10.1038/nrneph.2014.217

Dalman, M. R., Deeter, A., Nimishakavi, G., and Duan, Z.-H. (2012). Fold change and p-value cutoffs significantly alter microarray interpretations. BMC Bioinforma. *Biomed. Cent.* 13, S11–S14. doi:10.1186/1471-2105-13-S2-S11

de Cordoba, S. R., Hidalgo, M. S., Pinto, S., and Tortajada, A. (2014). "Genetics of atypical hemolytic uremic syndrome (ahus)," in *Seminars in thrombosis and hemostasis* (New York, United States: Thieme Medical Publishers), 422–430.

de Winter, J. C., Gosling, S. D., and Potter, J. (2016). Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychol. methods* 21, 273–290. doi:10.1037/met0000079

Ettetuani, B., Chahboune, R., and Moussa, A. (2019). Meta-analysis for a therapeutic target involved in the activation of the genes associated with c3 glomerulopathy. *Proc. New Challenges Data Sci. Acts Second Conf. Moroc. Classif. Soc.* 1–6. doi:10.1145/3314074.3314095

Ettetuani, B., Moussa, A., and Chahboune, R. (2020). "Functional cluster analysis of glomerular disease," in *International conference on advanced intelligent systems for sustainable development* (Berlin, Germany: Springer), 1116–1123.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The reactome pathway knowledgebase. *Nucleic acids Res.* 46, D649–D655. doi:10.1093/nar/gkx1132

Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., and Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med. Inf. Decis. Mak.* 19, 48–17. doi:10.1186/s12911-019-0801-4

Gao, Y., Lu, Y., Song, Y., and Jing, L. (2022). Analysis of codon usage bias of wrky transcription factors in helianthus annuus. *BMC Genomic Data* 23, 46. doi:10.1186/s12863-022-01064-8

Garfield, V. (2020). Sleep duration: A review of genome-wide association studies (gwas) in adults from 2007 to 2020. *Sleep. Med. Rev.* 56, 101413. doi:10.1016/j.smrv.2020.101413

Gaudet, P., Škunca, N., Hu, J. C., and Dessimoz, C. (2017). Primer on the gene ontology. *Gene Ontology Handb.* 25–37.

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8, 1–10. doi:10.1186/1752-0509-8-S2-I1

Gordon, D. E., Watson, A., Roguev, A., Zheng, S., Jang, G. M., Kane, J., et al. (2020). A quantitative genetic interaction map of hiv infection. *Mol. Cell* 78, 197–209. doi:10.1016/j.molcel.2020.02.004

Grayson, P. C., Eddy, S., Taroni, J. N., Lightfoot, Y. L., Mariani, L., Parikh, H., et al. (2018). Metabolic pathways and immunometabolism in rare kidney diseases. *Ann. rheumatic Dis.* 77, 1226–1233. doi:10.1136/annrheumdis-2017-212935

Greiff, V., Miho, E., Menzel, U., and Reddy, S. T. (2015). Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.* 36, 738–749. doi:10.1016/j.it.2015.09.006

Gry, M., Rimini, R., Strömberg, S., Asplund, A., Pontén, F., Uhlén, M., et al. (2009). Correlations between rna and protein expression profiles in 23 human cell lines. *BMC genomics* 10, 365–414. doi:10.1186/1471-2164-10-365

Gusenbauer, M., and Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources. *Res. synthesis methods* 11, 181–217. doi:10.1002/jrsm.1378

Ho Sui, S. J., Fulton, D. L., Arenillas, D. J., Kwon, A. T., and Wasserman, W. W. (2007). opossum: integrated tools for analysis of regulatory motif over-representation. *Nucleic acids Res.* 35, W245–W252. doi:10.1093/nar/gkm427

Iatropoulos, P., Noris, M., Mele, C., Piras, R., Valoti, E., Bresin, E., et al. (2016). Complement gene variants determine the risk of immunoglobulin-associated mpgn and c3 glomerulopathy and predict long-term renal outcome. *Mol. Immunol.* 71, 131–142. doi:10.1016/j.molimm.2016.01.010

Jeffery, I. B., Higgins, D. G., and Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinforma.* 7, 359–416. doi:10.1186/1471-2105-7-359

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., et al. (2010). Accounting for multiple comparisons in a genome-wide association study (gwas). *BMC genomics* 11, 724–726. doi:10.1186/1471-2164-11-724

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayqualitymetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics* 25, 415–416. doi:10.1093/bioinformatics/btn647

Ke, P.-f., Xiong, D.-s., Li, J.-h., Pan, J., Zhou, J., et al. (2021). An integrated machine learning framework for a discriminative analysis of schizophrenia using multi-biological data. *Sci. Rep.* 11, 14636–14711. doi:10.1038/s41598-021-94007-9

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. methods* 11, 740–742. doi:10.1038/nmeth.2967

Klaus, B., and Reisenauer, S. (2016). An end to end workflow for differential gene expression using affymetrix microarrays. *F1000Research* 5, 1384. doi:10.12688/f1000research.8967.1

Kovalchik, S. (2015). *Rismed: download content from ncbi databases*. R package version 2.

Lee, H.-C., Yoon, S. B., Yang, S.-M., Kim, W. H., Ryu, H.-G., Jung, C.-W., et al. (2018). Prediction of acute kidney injury after liver transplantation: machine learning approaches vs. logistic regression model. *J. Clin. Med.* 7, 428. doi:10.3390/jcm7110428

Legendre, C. M., Licht, C., Muus, P., Greenbaum, L., Babu, S., Bedrosian, C., et al. (2013). Terminal complement inhibitor eculizumab in atypical hemolytic–uremic syndrome. *N. Engl. J. Med.* 368, 2169–2181. doi:10.1056/NEJMoa1208981

Liu, P., Lassén, E., Nair, V., Berthier, C. C., Suguro, M., Sihlbom, C., et al. (2017). Transcriptomic and proteomic profiling provides insight into mesangial cell function in iga nephropathy. *J. Am. Soc. Nephrol.* 28, 2961–2972. doi:10.1681/ASN.2016101103

Maulik, U., Mukhopadhyay, A., and Bandyopadhyay, S. (2009). Combining pareto-optimal clusters using supervised learning for identifying co-expressed genes. *BMC Bioinforma.* 10, 27–16. doi:10.1186/1471-2105-10-27

Mazandu, G. K., and Mulder, N. J. (2013). Information content-based gene ontology semantic similarity approaches: toward a unified framework theory. *BioMed Res. Int.* 2013, 292063. doi:10.1155/2013/292063

McCall, M. N., and Irizarry, R. A. (2011). Thawing frozen robust multi-array analysis (frma). *BMC Bioinforma.* 12, 369–377. doi:10.1186/1471-2105-12-369

Mitrofanova, A., Molina, J., Santos, J. V., Guzman, J., Morales, X. A., Ducasa, G. M., et al. (2018). Hydroxypropyl-β-cyclodextrin protects from kidney disease in experimental alport syndrome and focal segmental glomerulosclerosis. *Kidney Int.* 94, 1151–1159. doi:10.1016/j.kint.2018.06.031

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., et al. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *J. Clin. Epidemiol.* 122, 56–69. doi:10.1016/j.jclinepi.2020.03.002

Okoniewski, M. J., and Miller, C. J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinforma.* 7, 276–314. doi:10.1186/1471-2105-7-276

Palasca, O., Santos, A., Stolte, C., Gorodkin, J., and Jensen, L. J. (2018). Tissues 2.0: an integrative web resource on mammalian tissue expression. *Database* 2018, bay003. doi:10.1093/database/bay003

Pickering, M. C., D'agati, V. D., Nester, C. M., Smith, R. J., Haas, M., Appel, G. B., et al. (2013). C3 glomerulopathy: consensus report. *Kidney Int.* 84, 1079–1089. doi:10.1038/ki.2013.377

Rani, J., and Ramachandran, S. (2015). pubmed. miner: an r package with text-mining algorithms to analyse pubmed abstracts. *J. Biosci.* 40, 671–682. doi:10.1007/s12038-015-9552-2

Riedl, M., Thorner, P., and Licht, C. (2017). C3 glomerulopathy. *Pediatr. Nephrol.* 32, 43–57. doi:10.1007/s00467-015-3310-4

Sabir, B., Ullah, F., Babar, M. A., and Gaire, R. (2021). Machine learning for detecting data exfiltration: A review. *ACM Comput. Surv. (CSUR)* 54, 1–47. doi:10.1145/3442181

Sartor, M. A., Tomlinson, C. R., Wesselkamper, S. C., Sivaganesan, S., Leikauf, G. D., and Medvedovic, M. (2006). Intensity-based hierarchical bayes method improves

testing for differentially expressed genes in microarray experiments. *BMC Bioinforma.* 7, 538–617. doi:10.1186/1471-2105-7-538

Scharf, J. M., Yu, D., Mathews, C. A., Neale, B. M., Stewart, S. E., Fagerness, J. A., et al. (2013). Genome-wide association study of tourette's syndrome. *Mol. psychiatry* 18, 721–728. doi:10.1038/mp.2012.69

Sidiropoulos, K., Viteri, G., Sevilla, C., Jupe, S., Webber, M., Orlic-Milacic, M., et al. (2017). Reactome enhanced pathway visualization. *Bioinformatics* 33, 3461–3467. doi:10.1093/bioinformatics/btx441

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., et al. (2009). Biomart–biological queries made easy. *BMC genomics* 10, 22–12. doi:10.1186/1471-2164-10-22

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinforma. Biol. insights* 14, 1177932219899051. doi:10.1177/1177932219899051

Sun, H., Guo, Y., Lan, X., Jia, J., Cai, X., Zhang, G., et al. (2020). Phenomodifier: a genetic modifier database for elucidating the genetic basis of human phenotypic variation. *Nucleic acids Res.* 48, D977–D982. doi:10.1093/nar/gkz930

Suravajhala, P., and Benso, A. (2017). Prioritizing single-nucleotide polymorphisms and variants associated with clinical mastitis. *Adv. Appl. Bioinforma. Chem. AABC* 10, 57–64. doi:10.2147/AABC.S123604

Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R., and Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Comput. Biol.* 3, e116. doi:10.1371/journal.pcbi.0030116

Tecnico, R. (2015). *Etls for importing ncbi entrez gene, mirbase, mircancer and microrna into a bioinformatics graph database*.

Thomas, P. D. (2017). The gene ontology and the meaning of biological function. *gene ontology Handb.* 15–24.

Tong, J., Xie, J., Ren, H., Liu, J., Zhang, W., Wei, C., et al. (2015). Comparison of glomerular transcriptome profiles of adult-onset steroid sensitive focal segmental glomerulosclerosis and minimal change disease. *PLoS One* 10, e0140453. doi:10.1371/journal.pone.0140453

Tsai, H.-M., Rice, L., Sarode, R., Chow, T. W., and Moake, J. L. (2000). Antibody inhibitors to von willebrand factor metalloproteinase and increased binding of von willebrand factor to platelets in ticlopidine-associated thrombotic thrombocytopenic purpura. *Ann. Intern. Med.* 132, 794–799. doi:10.7326/0003-4819-132-10-200005160-00005

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). Linkedomics: analyzing multi-omics data within and across 32 cancer types. *Nucleic acids Res.* 46, D956–D963. doi:10.1093/nar/gkx1090

Walsh, C. J., Hu, P., Batt, J., and Santos, C. C. D. (2015). Microarray meta-analysis and cross-platform normalization: integrative genomics for robust biomarker discovery. *Microarrays* 4, 389–406. doi:10.3390/microarrays4030389

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics* 23, 1274–1281. doi:10.1093/bioinformatics/btm087

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* 2, 100141. doi:10.1016/j.xinn.2021.100141

Xiao, X., Pickering, M. C., and Smith, R. J. (2014). "C3 glomerulopathy: the genetic and clinical findings in dense deposit disease and c3 glomerulonephritis," in *Seminars in thrombosis and hemostasis* (New York, United States: Thieme Medical Publishers), 465–471.

Xu, C., and Jackson, S. A. (2019). Machine learning and complex biological data. *Genome Biol.* 20, 76. doi:10.1186/s13059-019-1689-0

Yang, P., Ho, J. W., Yang, Y. H., and Zhou, B. B. (2011). Gene-gene interaction filtering with ensemble of filters. *BMC Bioinforma.* 12, S10–S10. doi:10.1186/1471-2105-12-S1-S10

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics* 26, 976–978. doi:10.1093/bioinformatics/btq064

Yu, G. (2012). Reactome pathway analysis. *Homo* 1266738, 29.

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118

Zhang, K., Geng, W., and Zhang, S. (2018). Network-based logistic regression integration method for biomarker identification. *BMC Syst. Biol.* 12, 135–122. doi:10.1186/s12918-018-0657-8

Zhang, X.-D. (2020). "Machine learning," in *A matrix algebra approach to artificial intelligence* (Berlin, Heidelberg: Springer), 223–440.

Zolotareva, O., and Kleine, M. (2019). A survey of gene prioritization tools for mendelian and complex human diseases. *J. Integr. Bioinforma.* 16, 20180069. doi:10.1515/jib-2018-0069

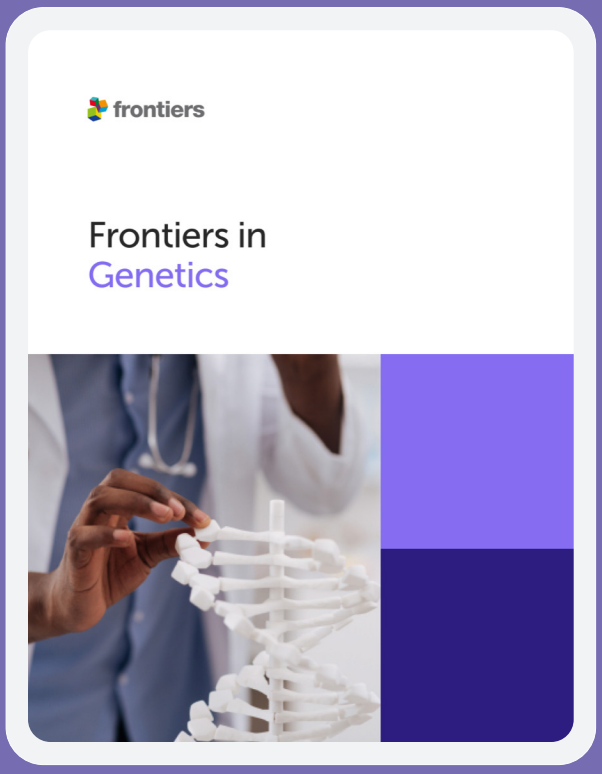
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact