frontiers | Frontiers in Physiology

TYPE Systematic Review
PUBLISHED 30 September 2022
DOI 10.3389/fphys.2022.952709

OPEN ACCESS

EDITED BY
Naveen Aggarwal,
Panjab University, India

REVIEWED BY
Rajesh Kumar Garg,
National Institute of Technology, Hamirpur, India
John Paul,
Universiti Malaysia Pahang, Malaysia

*CORRESPONDENCE
Mohammad Shabaz,
bhatsab4@gmail.com

SPECIALTY SECTION
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

RECEIVED 25 May 2022
ACCEPTED 01 September 2022
PUBLISHED 30 September 2022

CITATION
Gupta S, Gupta MK, Shabaz M and Sharma A (2022), Deep learning techniques for cancer classification using microarray gene expression data. *Front. Physiol.* 13:952709. doi: 10.3389/fphys.2022.952709

Deep learning techniques for cancer classification using microarray gene expression data

Surbhi Gupta[1,2], Manoj K. Gupta[1], Mohammad Shabaz[2]* and Ashutosh Sharma[3]

[1]Department of Computer Science and Engineering Department, SMVDU, Jammu, India, [2]Model Institute of Engineering and Technology, Jammu, India, [3]School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India

Cancer is one of the top causes of death globally. Recently, microarray gene expression data has been used to aid in cancer's effective and early detection. The use of DNA microarray technology to uncover information from the expression levels of thousands of genes has enormous promise. The DNA microarray technique can determine the levels of thousands of genes simultaneously in a single experiment. The analysis of gene expression is critical in many disciplines of biological study to obtain the necessary information. This study analyses all the research studies focused on optimizing gene selection for cancer detection using artificial intelligence. One of the most challenging issues is figuring out how to extract meaningful information from massive databases. Deep Learning architectures have performed efficiently in numerous sectors and are used to diagnose many other chronic diseases and to assist physicians in making medical decisions. In this study, we have evaluated the results of different optimizers on a RNA sequence dataset. The Deep learning algorithm proposed in the study classifies five different forms of cancer, including kidney renal clear cell carcinoma (KIRC), Breast Invasive Carcinoma (BRCA), lung adenocarcinoma (LUAD), Prostate Adenocarcinoma (PRAD) and Colon Adenocarcinoma (COAD). The performance of different optimizers like Stochastic gradient descent (SGD), Root Mean Squared Propagation (RMSProp), Adaptive Gradient Optimizer (AdaGrad), and Adaptive Momentum (AdaM). The experimental results gathered on the dataset affirm that AdaGrad and Adam. Also, the performance analysis has been done using different learning rates and decay rates. This study discusses current advancements in deep learning-based gene expression data analysis using optimized feature selection methods.

KEYWORDS

artificial intelligence, cancer, deep learning, gene expression, Rna-sequences

1 Introduction

Cancer is one of the deadliest diseases, and with its increasing prevalence, early identification and treatment are critical (Sung et al., 2021) (Schiff et al, 2007; Reid et al, 2011). Lung cancer cases have been surpassed by female breast cancer cases and are one of the most often detected forms of cancer. Figure 1 shows the cancer cases and deaths in 2020.

About two-third of cases are detected at initial stages (Fotouhi et al, 2019; Id et al., 2021, Kashyap et al, 2022). The classification and identification of gene expression using DNA microarray data is an effective tool for cancer diagnosis and prognosis for specific cancer subtypes. AI-based learning algorithms are vital tools and the most often used way to achieve significant features of gene expression data and play an essential part in gene categorization. This article will give a review of some of those strategies from the literature and information on the various datasets on which these techniques are applied and their associated benefits and drawbacks. The most classic variants of deep learning, such as Convolution Neural Networks, Artificial Neural Networks, and Autoencoders, have been established as essential tools for clinical oncology research and can be used to drive decision-making regarding disease diagnosis and therapy. As time passes, sickness in general, and cancer in particular, grow increasingly complex and challenging to identify, analyze, and treat. Cancer research is a prominent topic of study in the medical world.

1.1 Distribution of articles

The selected articles for analysis have been published in last 5-years. Most of the research articles explored in this study have been published in 2018 and 2019. The articles that have explored gene expression data for cancer diagnosis/survival/stage prediction have been included in this study. Figure 2 presents the year-wise distribution of articles.

1.2 Contributions of study

The study contributes in a number of ways. Following are the significant contributions made by the study:

- This article reviews recent developments in deep learning-based feature selection techniques for gene expression data interpretation and offers an extensive review of Deep Learning architectures that have demonstrated success across a wide range of industries and are now used to help doctors identify various chronic conditions.
- In this work, we have compared the outcomes of several optimizers on a dataset of RNA sequences. The study's deep learning system categorizes five types of cancer: colon cancer, lung adenocarcinoma, prostate cancer, invasive breast carcinoma, and kidney clear cell carcinoma (COAD).
- The efficiency of several optimizers, including adaptive gradient optimization (AdaGrad), stochastic gradient descent (SGD), root mean square propagation (RMSProp), as well as adaptive momentum (Adam). AdaGrad and Adam are more precise, according to the experimental findings discovered in the dataset. The performance of a variety of learning and decay rates was explored in the performance study.

1.3 Organization of paper

This paper is organized in a way that boosts the comprehensibility of the article. Second section gives the description of the significance of gene-expression analysis in cancer research. Section 2 gives description of search strategy used to select the articles for this study. Further Section 3 presents an overview of deep learning approaches where conventional approaches are discussed. Section 4 illustrates the importance of deep learning techniques in Cancer Prediction. Further, Section 5 embraces the literature of recent studies that have explored the deep learning strategies for gene section or survival prediction from microarray gene expression datasets. The article is discussed and concluded in Section 6 and Section 7, respectively. This study reviews and presents a comparative analysis of the previous studies. This article aims to analyze the concepts underlying deep learning-based classification algorithms used in healthcare.

2 Search strategy

The search strategy used in this paper is Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) strategy. All the research studies selected for this systematic review have been extracted from databases like PubMed, Web of Science, EBSCO, and EMBASE. All the research articles that have been published before 2016 are excluded from the analysis. The keywords used for extraction of articles include "Deep Learning", "Artificial Intelligence", "Cancer", "Micro-array analysis", "gene-expression", and combination of these keywords. The research articles that have focused on the optimization of gene selection using deep learning techniques have been included in the study. Figure 3 shows the PRISMA strategy flowchart.

3 Deep learning

The Artificial intelligence is the idea of making innovative and intelligent machines. Machine learning is an artificial

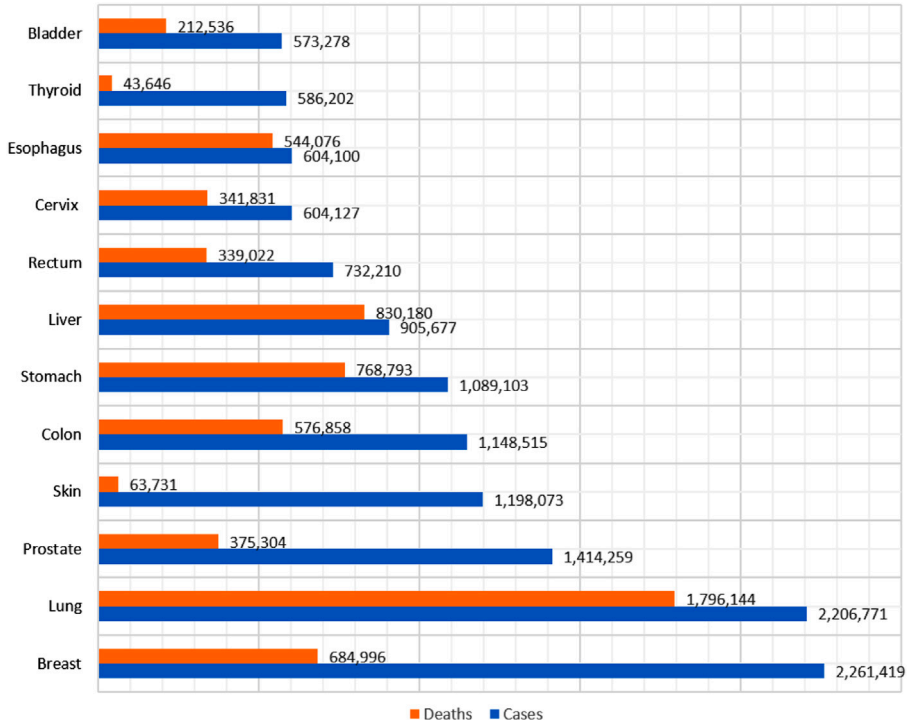

FIGURE 1
Cancer cases and deaths in 2020.

intelligence subset that aids in developing AI-driven applications. Deep learning is a subtype of machine learning that trains a model using large amounts of data and advanced methods. Figure 4 shows hierarchy of AI, Machine Learning, Deep Learning.

The significant differences between deep learning approaches and traditional learning are summarized in Table 1.

- Artificial Neural Networks: One of the most often used data modeling algorithms in medicine is neural networks. In the early 20th century, neural networks were developed (Daoud and Mayo, 2019). The primary goal of employing neural networks is to recognize patterns and conduct classification tasks. A human brain is used to represent the neural network system. The human brain is made up of millions of neurons that are all linked together. Figure 5 shows the representation of an artificial neural network.

Similarly, a neural network represents multiple neurons with a weight assigned to each link. These neurons act in parallel. During the learning stage, the network updates the weights for prediction of proper input to produce the output function (Gupta and Gupta, 2021b). Different optimization tasks are done by neural networks using different optimization techniques. Sigmoid optimization is mathematically given in Equation 1.

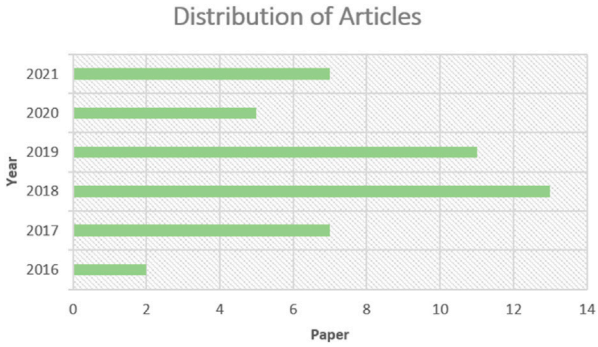

FIGURE 2
Year-wise Distribution of articles.

$$Sigmoid\,(a) = \frac{1}{1+e^{-a}} \tag{1}$$

The mathematical working of Hyperbolic Tangent (*Tanh*) optimization technique is given in Equation 2.

$$tanh\,(a) = \frac{2}{1+e^{-2a}} - 1 \tag{2}$$

The working of Rectilinear Unit (*Relu*) optimization technique is expressed in Eq. 3.

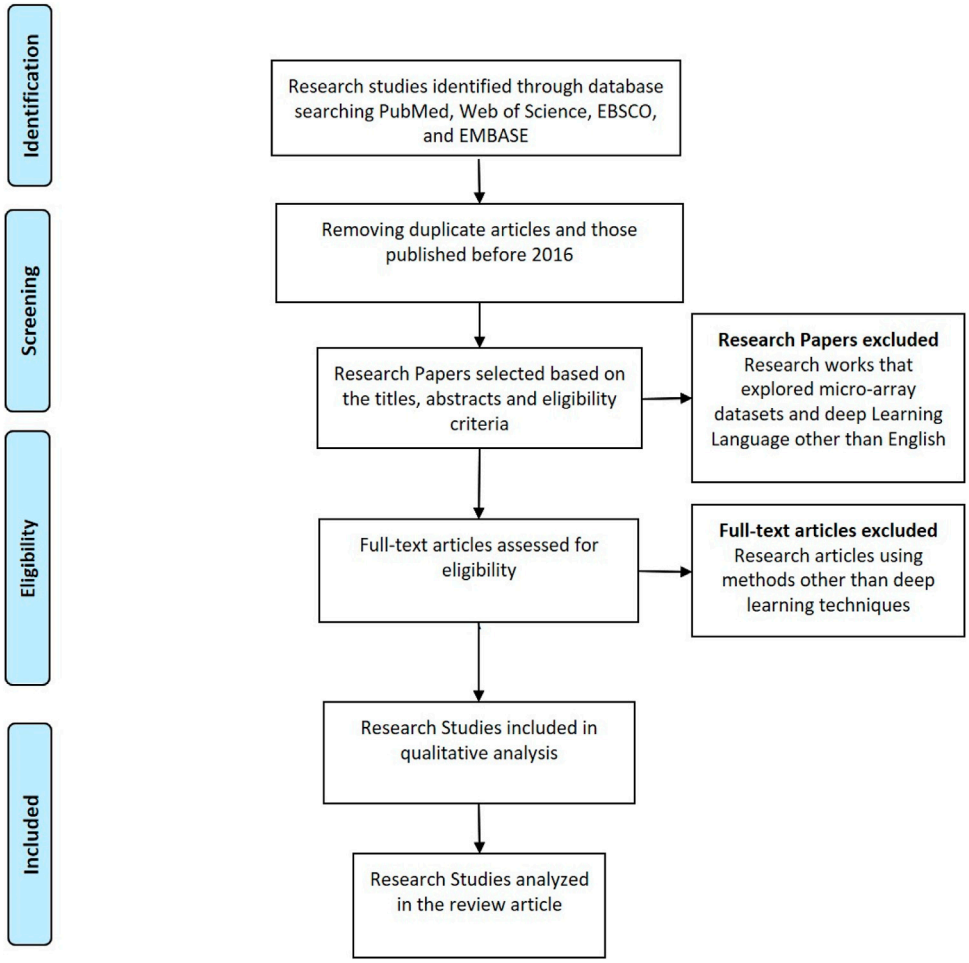

FIGURE 3
Prisma search strategy.

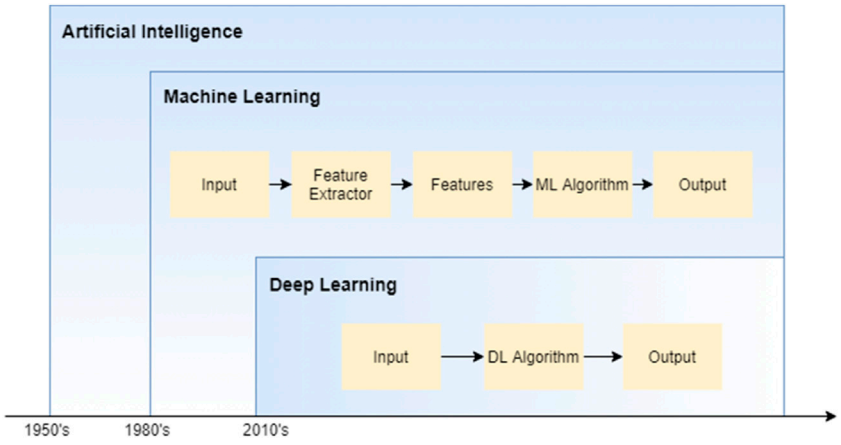

FIGURE 4
Artificial intelligence and sub-parts.

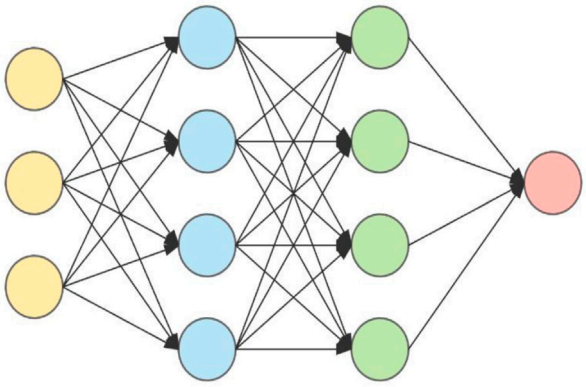

FIGURE 5
Artificial intelligence.

$$relu\,(a) = \max\,(0, a) \tag{3}$$

Because of its adaptive character, altering the weights aids in the minimization of error. In contrast to basic modeling methods, neural networks have the advantage of predicting non-linear relationships. In the study of medical data, neural networks play a significant role such as medication development. The use of a neural network to predict cardiac disease is possible.

- Convolutional Neural Network (CNN): CNN is a multi-layer neural network based on the visual brain of animals. LeCun et al. constructed the first CNN. CNN's major application areas include image processing and character recognition (Akkus et al, 2017; Zahras, 2018). In terms of construction, the initial layer recognizes features, however the intermediate layer recombines features to produce high-level input characteristics, followed by classification. The collected characteristics will then be pooled, which reduces their dimensionality. Convolution and pooling are the following steps, which are then put into a fully connected multi-layer perceptron. The last layer, known as the output layer, recognizes the image's characteristics using back-propagation techniques (Gupta and Gupta, 2021a). Because of its unique properties, such as local connection and shared weights, CNN increases the system's accuracy and performance. It outperforms all other deep learning techniques. In comparison to other types of architecture, it is the most often utilized. Figure 6 shows a convolutional neural network.

The cell's weight can be utilized as a regulating factor. There is a requirement for a training approach known as Backpropagation through time (BPTT) that improves weight. For optimization, the technique requires network output error.

4 Deep learning in cancer prediction

Deep learning has been widely utilized to improve prognosis (Huang et al., 2020). Gene expression profiles, which describe the molecular state, offer enormous promise as a medical diagnostic tool. However, current training data sets have a minimal sample size for classification compared to the number of genes involved, and these training data constraints challenge specific classification techniques. One

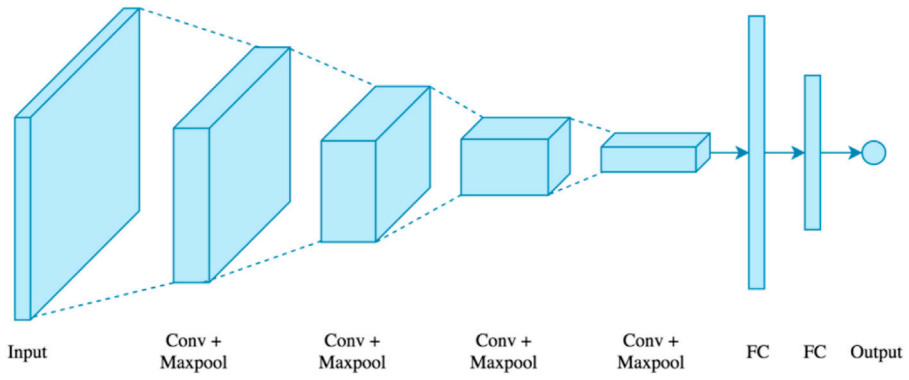

FIGURE 6
Convolutional neural network. ● Long Short-Term Memory (LSTM) Network: Hochreiter and Schimdhuber collaborated to create the LSTM (Lecun et al, 2015), which is utilized in various applications. LSTMs were chosen by IBM primarily for voice recognition. The LSTM employs a memory unit known as a cell that may retain its value for an extended period and aids the device in remembering the most recent computed value. The memory unit, also known as a cell, comprises three gates that regulate the movement of data inside the unit. Figure 7 shows the logical structure of a LSTM model.

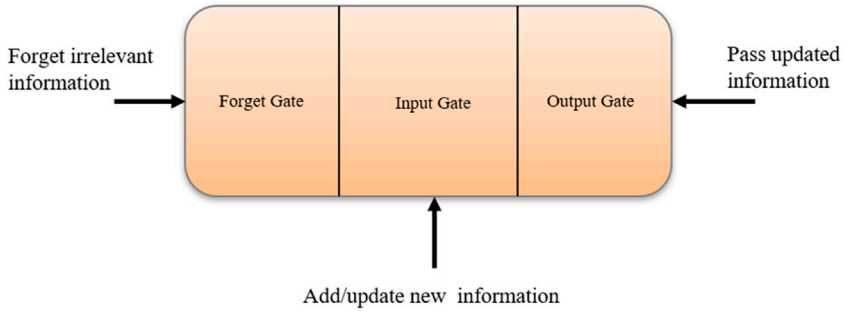

FIGURE 7
Long short-term memory.● The input port, also known as the gate, controls new data flow into the memory. ● The forget gate forgets the irrelevant/unnecessary information. ● The third port must regulate the information stored as output.

TABLE 1 Distinction between deep and traditional learning.

Feature	Traditional learning	Deep learning
Extraction and representation of features	Traditional learning relied on feature vectors that were manually created and were application-specific. In complexity, these characteristics are difficult to model	Deep learning approaches can learn characteristics from raw sensor data and determine the best pattern for enhancing recognition accurateness
Diversity and Generalization	Traditional learning relied on sensor data that had been tagged. Also, use dimensionality reduction strategies to focus on feature selection	Deep learning allows the extraction of intricate properties from complex data
Data preparations	Traditional learning derives features from sensors based on their appearance and active windows	Pre-processing and standardization of data are not required in deep learning
Changes in Activities' Temporal and Spatial Dimensions	In traditional learning, handcrafted features are ineffective and unsuitable for resolving inter-class variability and inter-class linkages	Handcrafted characteristics with intra-class variability can be solved by using hierarchical features and translational invariant features
Model Training and Execution Time	In traditional training, small-sized data can also train the model and reduced computation time and space usage	Deep learning requires a vast amount of sensor datasets to avoid overfitting. It is accelerated using a graphics processing unit (GPU). It is also utilized to speed up computations

of the most important new clinical applications of microarray data is abnormality detection. Because of the high dimensionality, gene selection is a crucial step in enhancing the classification performance of expression data. As a result, better approaches for selecting functional genes for cancer prediction and detection are required. Microarray studies yield a massive quantity of gene-expression information from a single sample. The quantity of gene-expressions (features) to cases (samples) ratio is highly skewed, resulting in the well-known curse-of-dimensionality issue. In a single experiment, microarray technology generates hundreds of gene expressions. However, comparing the quantity of characteristics, the quantity of samples/patients is significantly lower (up to a few hundred) (several thousand). The limited number of samples (training data) provided is insufficient to create an efficient model from the given data. This is referred to as data scarcity.

Processing microarray gene expression data is a diverse field of computer science that includes graph analysis, machine learning, clustering, and classification. Microarray technology allows for the measurement of thousands of gene expressions in a single experiment. Gene expression levels aid in identifying linked genes and disease development, which aids in the early detection and prognosis of many forms of cancer.

5 Literature work

Using microarray gene expression patterns (Dwivedi, 2016), develop a framework of supervised machine learning approaches for discriminating acute lymphoblastic leukemia from acute myeloid leukemia. This classification was accomplished using an artificial neural network (ANN) (Tumuluru and Ravi, 2017). Using microarray gene expression patterns develop a framework of supervised machine learning approaches for discriminating acute lymphoblastic leukemia from acute myeloid leukemia. This classification was accomplished using an artificial neural

TABLE 2 Research analysis.

Study	Cancer dataset	Objective	Technique	Acc
Dwivedi, (2016)	Leukemia	Cancer classification employing microarray gene-expression data using deep learning	ANN	98%
Yuan et al. (2016)	12 selected types of cancer	Cancer type classification using deep learning and somatic point mutations	DeepGene	94%
Motieghader et al., 2017	6 different cancer	Cancer classification using microarray data using genetic algorithm	Genetic Algorithm	94%
Aziz et al., 2017	5 gene microarray datasets	Microarray data classification using novel hybrid method	Artificial Bee Colony (ABC)	95%
Tumuluru and Ravi, (2017)	Colon and Leukemia data	Implementing deep neural networks for cancer classification	GOA-based DBN	95%
Extraction, (2017)	Breast cancer	Integrated deep neural networks to predict breast cancer	Deep-SVM	70%
Danaee et al. (2017)	Breast cancer	Relevant gene identification for better cancer classification	Stacked Denoising Autoencoder (SDAE)	98%
Urda and Moreno, (2017)	3 cancer databases	Investigating RNA-sequence gene expression data utilizing deep learning	Regularized linear model (standard LASSO) and two deep learning models	75%
Salman, (2018)	TCGA	Analyzing the Effect of meta heuristic iteration on the neural networks in cancer data	GA and FWA	98%
Ching et al. (2018)	TCGA RNA-Sequence data	Evaluating deep learning technique for tumor detection	Cox-nnet.	--
Cho et al. (2018)	TCGA LUAD	Examining the relationship between specific gene mutations and lung cancer survival	Information gain, chi-squared test	--
Xiao et al. (2018b)	RNA-sequence data sets of three cancers	Analyzing deep learning technique to predict cancer employing RNA sequence data	Sparse Auto-Encoder (SSAE)	98%
Lin et al. (2018)	TCGA Leukemia	Introduced deep learning to Predicting Prognosis of Leukemia	Stacked Autoencoders	83%
Alomari et al. (2018)	10 microarray datasets	Implementing a novel strategy for gene selection based on a hybrid technique	Hybrid Bat-inspired Algorithm	100%
Parvathavardhini and Manju, (2020)	Gene expression data of liver cancer	Cancer gene recognition using neuro-fuzzy approach	Neuro-Fuzzy method	96%
Ahn and Lee, (2018)	TCGA	Recognition of cancer tissues using RNA-Sequence data	Deep neural network (DNN)	99.7%
Kong and Yu, (2018)	Two RNA-seq expression datasets	Extracting features for RNA-Sequence data classification	Forest Deep Neural Network (fDNN)	90.4%
Guo et al., (2018)	Multiple Cancer datasets	Cancer subtype classification using RNA-Sequence gene expression data	BCD Forest	92.8%
Chen et al. , (2018)	mRNA datasets from the GDC repository	Cancer type recognition using neural network	Deep Learning models	98%
Sevakula et al. (2018)	36 datasets from the GEMLeR repository	Implemented transfer learning for molecular cancer classification	Sparse Autoencoders on gene expression data	98%
Joshi and Park, (2019)	LUAD	Lung Cancer Subtype Classification using deep learning model	Sparse Cross-modal Superlayered Neural Network	99%
Gao et al., (2019)	gene expression data	Cancer subtype prediction using gene expression data	Deep cancer subtype classification (DeepCC)	90%
Basavegowda and Dagnew, (2020)	8 microarray cancer datasets	Deep neural networks for classifying microarray cancer data	7-layer deep neural network architecture	90%
Huynh et al., (2019)	TCGA	Developed hybrid approach for classifying RNA-sequence data	Deep convolutional neural network (DCNN)	95%
Xu et al., (2019a)	RNA-seq gene expression data	Cancer subtype classification	Deep flexible neural forest (DFNForest)	76%
Xiao et al., (2018a)	LUAD, BRCA, and STAD	Cancer type prediction using deep learning model	ensemble-based approach	97%
Guia, (2019)	RNA-sequence data from Pan-Cancer Atlas	Cancer type Classification using RNA-sequence data	DeepGx Convolutional neural network (CNN)	95.65%
Huang et al. (2020)	TCGA cancers	Cancer survival prediction from RNA-sequence data	AECOX (AutoEncoder with Cox regression network)	--
Shon et al. (2021)	TCGA stomach cancer dataset		CNN	96%

(Continued on following page)

TABLE 2 (*Continued*) Research analysis.

Study	Cancer dataset	Objective	Technique	Acc
		Stomach cancer prediction using gene expression data		
García-díaz et al., (2019)	5 types of cancer	Multiclass cancer classification of gene expression RNA-Sequence data	Extreme Learning Machine algorithm	98.81%
Kim et al., (2020)	TCGA	Cancer prediction using gene expression data	NN, SVM, KNN, RF	94%
Jerez et al., (2020)	31 Tumor types	Prediction of cancer survival using gene-expression data	Transfer learning with CNN	73%
Panda, (2017)	10 most common UCI Cancer datasets	Analyzed microarray cancer data using deep neural networks	Elephant search optimization based deep learning approach	92%
He and Luo, (2020)	15 different cancer types	Prediction of the tissue-of-origin of cancer types on basis of RNA-sequence data	a novel NN model	80%
Abdollahi et al., (2021)	Diabetes, heart, cancer dataset	Disease prediction model for the healthcare system	neural network-based ensemble learning	100%
Torkey et al., (2021)	RNA-seq data of three datasets	Cancer survival analysis for microarray dataset.	AutoCox and AutoRandom	98%
Wessels et al., (2021)	prostate cancer patients	Prediction of lymph node metastasis straight from tumor histology in prostate malignancy	convolutional neural network	62%
Chaunzwa et al., (2021)	311 NSCLC patients at Massachusetts General Hospital	Tumor detection using CT images	convolutional neural network	71%
Gupta, (2021)	cervical cancer dataset	Prediction of Cervical Cancer risk factors	Ensemble model	99.7%
Gupta and Gupta, (2021a)	five benchmark datasets	Cancer diagnosis analysis along with imbalanced classes	Stacked Ensemble Model	98%

network (ANN). In 2020, prostate cancer (Surbhi Gupta, 2021) was predicted using Multi-layer perceptrons and explored multiple data balancing techniques. Another recent study in 2021 (Gupta and Gupta, 2021b) predicted mesothelioma with 96% accuracy using ANN (Tumuluru and Ravi, 2017). presented an approach for cancer categorization based on gene-expression data. The logarithmic transformation pre-processed the gene expression data to reduce the classification's complexity, while the Bhattacharya distance identified the most informative genes. The weight update in Deep Belief Neural Networks has estimated the average error using GOA and Gradient Descent.

The experimentation with colon and leukemia data demonstrates the proposed cancer classification's efficacy. The accuracy rate of the proposed classification approach employing gene expression data is 0.9534, and 0.9666 detection rate.

Despite decades of research, clinical diagnosis of cancer and the identification of tumor-specific markers remain unknown (Danaee et al., 2017). offered a deep learning technique for cancer detection and identifying critical genes for breast cancer diagnosis using autoencoders. The error rates are computed using log loss function given in Equation 4.

$$Logloss = \sum J(k)log(L(m)) + (1 - J(k))(log(1 - L(m))) \quad (4)$$

In the above equation, $J(k)$ and $L(m)$ represent prediction and target values (Cho et al., 2018). applied automated learning to search for survival-specific gene mutations in patients with lung adenocarcinoma (LUAD) using data from TCGA. Distinct feature selection methods were utilized to find survival-specific mutations in response to particular clinical variables. Kaplan-Meier survival analysis was performed on the extracted LUAD survival-specific mutations individually or in groups. Patient death was strongly associated with mutations in MMRN2 and GMPPA, whereas patient survival was associated with mutations in ZNF560 and SETX. In addition, DNAJC2 and MMRN2 mutations were associated with a substantial negative correlation with overall survival, but ZNF560 mutations were associated with a significant positive correlation with overall survival (Lin et al, 2018). tested the proposed SSAE model on three public RNA-seq data sets of three types of cancers.

A retrospective study (Lin et al., 2018) investigated the use of Deep Learning (DL) to predict acute myeloid leukemia (AML) prognosis. This study used 94 AML cases from the TCGA database. Age, ten common cytogenetic mutations, and the 23 most common mutations have been used as input data. Also, the results suggested feasible applications of deep learning (DL) in the prognostic prediction utilizing next-generation sequencing (NGS) data as proof-of-concept research.

Research work (Parvathavardhini and Manju, 2020) proposed a Neuro-Fuzzy approach for interpreting gene-expression data from microarray experiments. The analysis enabled the detection and classification of cancer, hence facilitating treatment selection and development. The proposed strategy was evaluated against three publicly

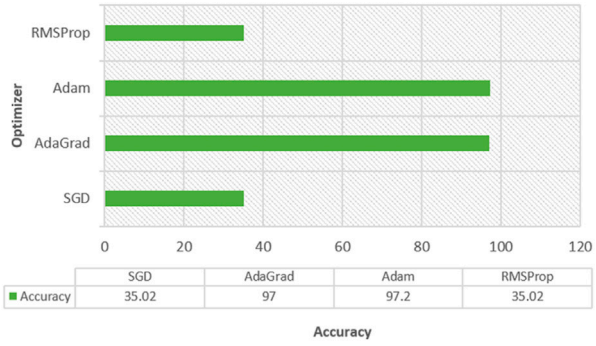

FIGURE 8
Accuracy of multiple optimizers.

available datasets of cancer gene expression. Also (Sevakula et al, 2018), proposed a cancer-verification transfer learning process in combination with autoencoders. The cross entropy function is used for optimizing the neural models. The cross entropy (CE) is calculated using Equation 5.

$$CE = \frac{1}{k}\sum_{i=1}^{k} Y_i log\left(X_i\right) + \left(1 - Y_i\right)\left(log\left(1 - X_i\right)\right) \tag{5}$$

The term X_i denotes the probability for i^{th} instance and Y_i represents all the truth values for k instances. The algorithm's performance was evaluated on the GEMLeR repository dataset, and hence has significant implications for precision medicine.

(Xu et al., 2019b) employed numerous computational methods for classifying cancer subtypes have been presented. However, the majority of them create the model only using gene expression data. 2019 (Huynh et al, 2019). proposed a new support vector machine (SVM) classification model for gene expression based on features collected from a deep convolutional neural network (DCNN). The Equation 6 illustrates the working of CNN.

$$K[x, y] = (a*b)[x, y] = \sum_{j}\sum_{k} b[j, k]a[x - j, y - k] \tag{6}$$

Here a *and b* denote the input data and kernel respectively. Also, $[x, y]$ denote the row and column indexes of resultant matrix

Nonetheless, it is characterized by highly high-dimensional data, which results in an over-fitting problem for the classifying model (Lin et al, 2018). purposed a novel way for incorporating deep learning into an ensemble approach that included numerous machine learning models. First, the study provided valuable gene data to five distinct categorization models using differential gene expression analysis. Then outputs of the five classifiers are then combined using a deep learning algorithm.

Significant bioinformatics research (Shon et al, 2021) has been undertaken in cancer research, and bioinformatics methodologies may aid in developing methods and models

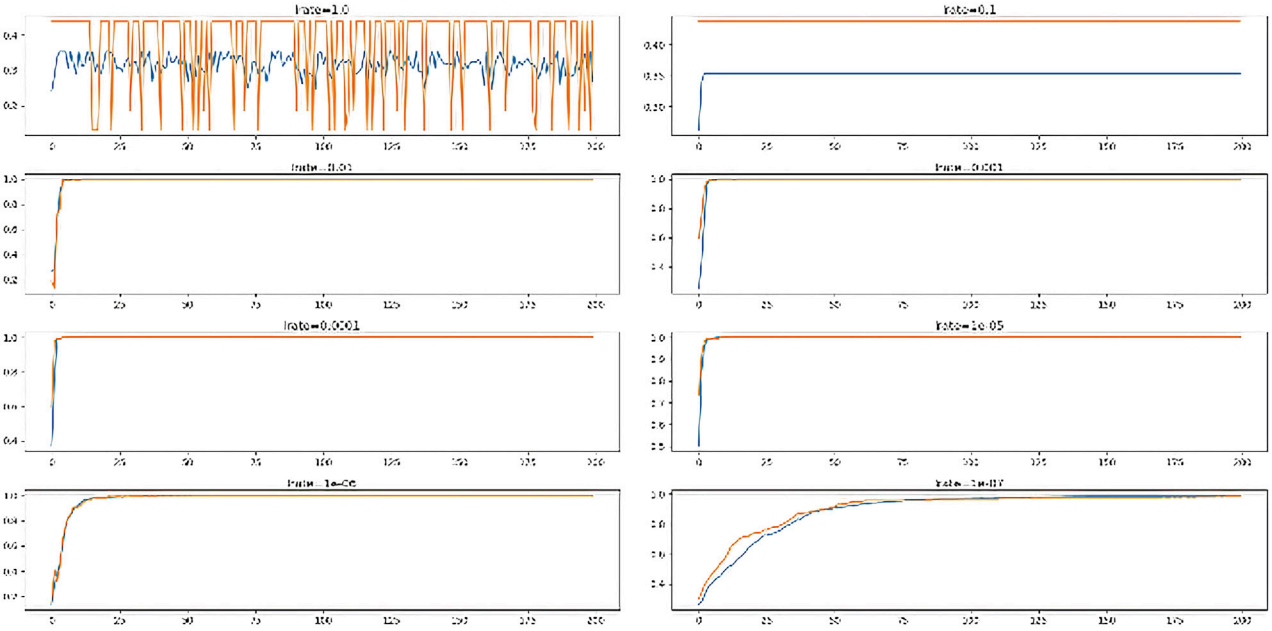

FIGURE 9
Performance of multiple learning rates.

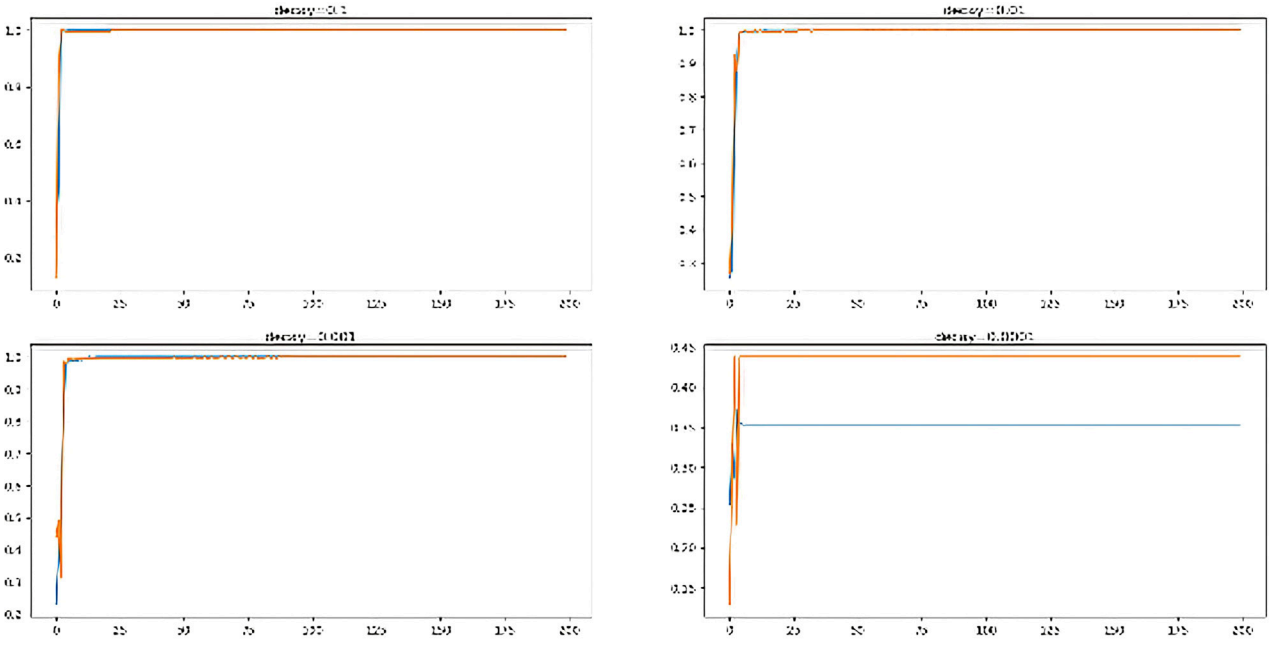

FIGURE 10
Performance of multiple decay rates.

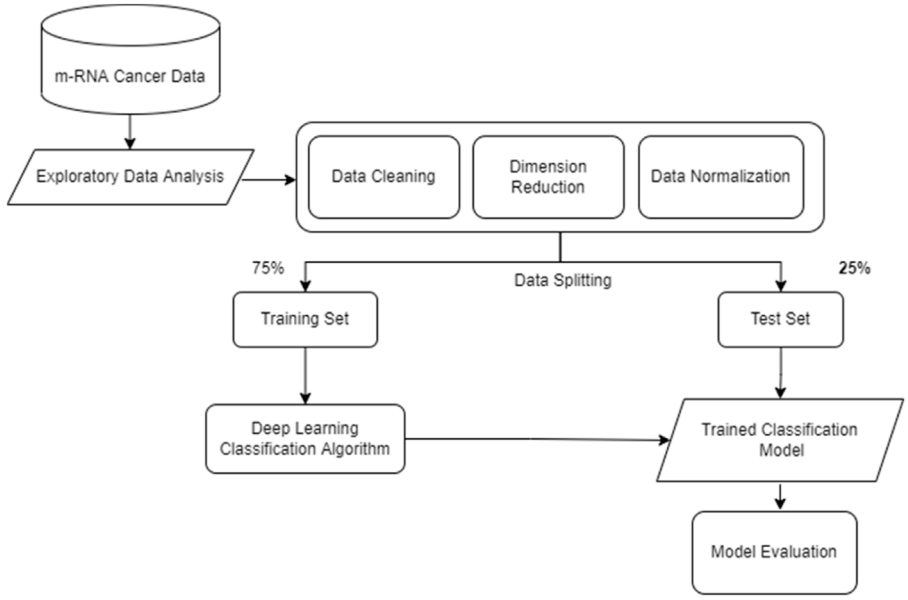

FIGURE 11
Deep learning for Cancer Classification.

for early prediction of stomach cancer. This study aimed to build a CNN algorithm to analyze TCGA data. This study merged RNA-seq, and clinical data looked for and assessed potential genes employing the CNN model. In addition, this study performed learning and evaluated the status of cancer patients. The proposed model acquired an accuracy of 95.96 percent and a critical status accuracy of 50.51 percent. Despite overfitting due to the small sample size, reasonably accurate results for the sample type were achieved. This method can be used to forecast the diagnosis of stomach cancer, which comes in various forms and has a variety of underlying causes.

(Gupta and manoj, 2021) discovered that group algorithms for chronic disease diagnosis could be more effective than baseline algorithms. Additionally, it outlines many impediments to furthering the use of machine learning classification to detect illness. The proposed strategy achieved 98.5, 99, and 100% accuracy in this study. The disease datasets used in the study includes Diabetes, Cardiovascular Disease, and Breast Cancer. The algorithms used for the disease prediction are Group Algorithms, Stacked, and Neural Network.

(Abdollahi et al, 2021) proposed a novel strategy for reducing the number of features by utilizing an autoencoder. Each gene's weight is determined as a consequence of our autoencoder model. The weights indicate the magnitude of each gene's effect on survival probability. Our approach enhances survival analysis by speeding up the procedure, increasing prediction accuracy, and decreasing the calculated survival probability's error rate. The error rates are computed using root mean squared error (*RMSE*). The mathematical formula of RMSE is given in Equation 7) where A and O represent actual and observed values respectively.

$$RMSE = \sqrt{\frac{\sum((A) - (O))^2}{N}} \tag{7}$$

5.1 Comparative analysis

Multiple studies aimed to investigate cancer prediction models. Table 2 presents the research analysis table.

6 Experimental results

This section holds the simulation results achieved using ANN model along with multiple optimizers like Stochastic gradient descent (SGD), Root Mean Squared Propagation (RMSProp), Adaptive Gradient Optimizer (AdaGrad), and Adaptive Momentum (AdaM). Also, the performance analysis has been done using different learning rates and decay rates.

6.1 Dataset analysis

TCGA dataset is available at https://archive.ics.uci.edu/ml/datasets/gene + expression + cancer + RNA-Seq. This dataset comprises data on five different forms of cancer, including kidney renal clear cell carcinoma (**KIRC**), Breast Invasive Carcinoma (**BRCA**), lung adenocarcinoma **(LUAD**), Prostate Adenocarcinoma (**PRAD**) and Colon Adenocarcinoma (**COAD**). The dataset consists of 20,531 attributes of 801 patients.

6.2 Optimization with multiple optimizers

The performance of multiple optimizers is analyzed and shown in Figure 8. From Figure 8, it is clear that both "Adam" and "Adagrad" performed the best on training and testing data.

The ANN model using SGD and rmsprop optimizer attained 35.3% on training data and 43.8% on test data. Both the Adam approaches performed well. Hence, we considered analyzing the performance of different parameters like learning rates and decay rates.

6.3 Optimization with learning rates

The performance of ***ADAM*** optimizer using different learning rates is analyzed and shown in Figure 9.

From the figure, it is clear that learning rate ('0.01', '0.001', '0.0001', '$1e^{-05}$) performed the best on training and testing data. The ANN models performed worst (35% on train and 43.8% on test set) with slowest (lrate = "1.0", "0.1").

6.4 Optimization with decay rates

The technique of learning rate decay (lrDecay) is used to train current neural networks. It begins with a high rate of learning and then decays several times. It has been demonstrated empirically to aid in both optimization and generalization. The performance of *ADAM* optimizer using different decay rates is investigated and revealed in Figure 10.

From the figure, it is clear that decay rate ("0.1", "0.001") performed the best on training and testing data. The ANN models performed worst (35.3% on train and 43.8% on test set) and (63.5% on train and 68.7% on test set) with decay rates "0.01" and "0.0001" respectively.

7 Discussion

Several strategies for gene selection in cancer categorization have been proposed in prior studies. The advent of deep learning has profoundly affected a wide variety of machine learning applications and research. Few of such studies (Gupta and Gupta, 2021a), (Gupta and Gupta, 2021b), (Gupta and Gupta, 2021c) are described in this section. The work flow used for classification of cancer data is shown in Figure 11.

Initially, the exploration of data is done and termed as "exploratory data analysis". Further, data preprocessing steps are used like cleaning data, reducing dimension (feature reduction), normalizing the data. Further the next stage splits the preprocessed data into sets. The deep learning classification algorithm is trained on the training set for classification of data. The trained classification model is further evaluated on the test set. The evaluation of the data can express the accurateness of the model. The number of cancer cases is rapidly increasing. It is difficult to diagnose because the illness is frequently asymptomatic in its early stages. Early detection can increase the odds of a patient's recovery and cure. Cancer is notoriously difficult to diagnose in its early stages and is prone to recurrence after treatment. Cancer classification is a crucial topic. One of the most effective methods for cancer classification is gene selection (Gupta and Gupta, 2021d). The task of choosing a set of genes that enhances classification accuracy is NP-Hard. Furthermore, making accurate and specific cancer diagnostic forecasts is quite tricky. Because of the nonspecific symptoms and imprecise scans, certain tumors are more challenging to diagnose in their early stages. As a result, improving the prediction model in diagnostic cancer research is vital. Furthermore, most cancer research articles have increased dramatically, particularly those that use deep learning methodologies (Shimizu and Nakayama, 2020). Again, the present research shows that traditional analysis techniques (Akkus et al., 2017; Ronoud and Asadi, 2019; Chaunzwa et al., 2021) aid in improving the prediction accurateness and is frequently applied in healthcare sector. Its success is since it enables the discovery of highly complicated non-linear correlations between characteristics; and the extraction of information from unlabeled data unrelated to the situation at hand. Statistical studies demonstrate that deep learning models outperform numerous widely used cancer categorization algorithms.

Several academics have investigated automated learning methodologies; however, these approaches still have several flaws that make cancer classification difficult. Specific machine learning algorithms have been found incapable of exploiting unstructured data in cancer classification. CNNs are particularly appropriate for analyzing a wide range of unstructured data. This capability enabled deep learning algorithms to take an active role in the early diagnosis of cancer through data classification. Deep learning approaches have achieved high accuracy and other statistical characteristics. Deep Learning has succeeded in various domains, including image, video, audio, and text processing. Deep Learning faces a unique problem in gene expression analysis for various cancer detection and prediction tasks to define appropriate biomarkers for different cancer subtypes. Despite several research studies on multimodal treatment approaches, survival times remain short. The gathering of significant genes that can increase accuracy can provide adequate guidance in early cancer detection. Cancer can be classified into several subgroups. However, it is a complex task because of the vast number of genes and the comparatively few experiments in gene expression data (Kumar et al, 2021). Cancer identification from microarray gene expression data presents a significant difficulty due to the small sample size, high dimensionality, and complexity of the data (Dargan et al, 2020). There is a need for rapid and computationally efficient methods to address such issues. This study briefly explores the research studies that employed deep learning architectures that selected the most relevant genes for cancer prediction using gene expression data. Although Deep Learning has had success in various domains, it has yet to be thoroughly explored in genomics, notably in genomic cancer.

8 Conclusion

Cancer has become one of the top causes of death worldwide in recent years. As a result, increasing research is being done to determine the most effective diagnosing and treating cancer. However, cancer treatment faces numerous obstacles, as possible causes of cancer include genetic problems or epigenetic modifications in the cells. RNA sequencing is a substantial approach for assessing gene expression in model organisms and can provide information for bio-molecular cancer diagnosis. Microarray gene expression profiles can be used to classify tumors efficiently and effectively. Predicting various tumors is a significant problem, and offering accurate predictions would be highly beneficial in delivering better therapy to patients. The advent of deep learning approaches is critical for improving patient monitoring, as it can aid clinicians in making decisions regarding deadly diseases. Furthermore, Gene expression data are utilized to develop a classification model that will help cancer treatment. Classification of cancer subtypes is critical for effective diagnosis and individualized cancer treatment. The article concludes that the recent advances in high-throughput sequencing technology have resulted in the quick generation of multi-omics data from the same cancer sample. Thus, deep learning-based molecular

illness classification holds considerable promise in the realm of genomics, particularly concerning gene microarray data.

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

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

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

Abdollahi, J., Nouri-Moghaddam, B., and Ghazanfari, M. (2021). Deep Neural Network Based Ensemble learning Algorithms for the healthcare system diagnosis of chronic diseases. ArXiv Preprint Available at: https://ArXiv/org/abs.2103.08182.

Ahn, T., and Lee, C. (2018). Deep learning-based identification of cancer or normal tissue using gene expression data.In Proceeding IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Madrid Spain. 03-06 December 2018. IEEE, 1748–1752. doi:10.1109/BIBM.2018.8621108

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., and Erickson, B. J. (2017). Deep learning for brain MRI segmentation : State of the art and future directions. *J. Digit. Imaging.* 30 (4), 449–459.

Alomari, O. A., Khader, A. T., Al-Betar, M. A., and Awadallah, M. A. (2018). A novel gene selection method using modified MRMR and hybrid bat-inspired algorithm with β-hill climbing. *Appl. Intell. (Dordr).* 48 (11), 4429–4447. doi:10.1007/s10489-018-1207-1

Aziz, R., Verma, C. K., and Srivastava, N. (2017). A novel approach for dimension reduction of microarray. *Comput. Biol. Chem.* 71, 161–169. doi:10.1016/j.compbiolchem.2017.10.009

Basavegowda, H. S., and Dagnew, G. (2020). Deep learning approach for microarray cancer data classification. *CAAI Trans. Intell. Technol.* 5, 22–33. doi:10.1049/trit.2019.0028

Chaunzwa, T. L., Hosny, A., Xu, Y., Shafer, A., Diao, N., Lanuti, M., et al. (2021). Deep learning classification of lung cancer histology using CT images. *Sci. Rep.* 1, 5471. doi:10.1038/s41598-021-84630-x

Chen, X., Xie, J., and Yuan, Q. (2018). A method to facilitate cancer detection and type classification from gene expression data using a deep autoencoder and neural network. *Mach. Learn.* Available at: https://arXiv/org/abs1812.08674.

Ching, T., Zhu, X., and Garmire, L. X. (2018). Cox-nnet : An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput. Biol.* 14, e1006076–18. doi:10.1371/journal.pcbi.1006076

Cho, H., Lee, S., Ji, Y. G., Hyeon, D., and Id, L. (2018). Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. *PLoS One* 13, e0207204. doi:10.1371/journal.pone.0207204

Danaee, P., Ghaeini, R., and Hendrix, D. A. (2017). A deep learning approach for cancer detection and relevant gene identification. *Pac. Symp. Biocomput.* 22, 219–229. doi:10.1142/9789813207813_0022

Daoud, M., and Mayo, M. (2019). A survey of neural network-based cancer prediction models from microarray data. *Artif. Intell. Med.* 97, 204–214. doi:10.1016/j.artmed.2019.01.006

Dargan, S., Kumar, M., Rohit, M., and Gulshan, A. (2020). A survey of deep learning and its applications : A new paradigm to machine learning. *Arch. Comput. Methods Eng.* 27 (4), 1071–1092. doi:10.1007/s11831-019-09344-w

Dwivedi, A. K. (2016). Artificial neural network model for effective cancer classification using microarray gene expression data. *Neural comput. Appl.* 29, 1545–1554. doi:10.1007/s00521-016-2701-1

Extraction, S. F. (2017). "Prognosis prediction of human breast cancer by integrating deep neural network and support vector machine supervised feature extraction and classification for breast cancer prognosis prediction," in Proceeding International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai China, 14-16 October 2017 (IEEE). doi:10.1109/CISP-BMEI.2017.8301908

Fotouhi, S., Asadi, S., and Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J. Biomed. Inf.* 90, 103089. doi:10.1016/j.jbi.2018.12.003

Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., et al. (2019). DeepCC : A novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* 8, 44. doi:10.1038/s41389-019-0157-8

García-díaz, P., Sánchez-berriel, I., Martínez-, J. A., and Diez-pascual, A. M. (2019). Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data. *Genomics* 112, 1196. doi:10.1016/j.ygeno.2019.11.004

Guia, J. M. De. (2019). "DeepGx : Deep learning using gene expression for cancer classification," in Proceeding IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver BC Canada, 27-30 August 2019 (IEEE), 913–920. doi:10.1145/3341161.3343516

Guo, Y., Liu, S., Li, Z., and Shang, X. (2018). BCDForest : A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. *BMC Bioinforma.* 19 (5), 118–213. doi:10.1186/s12859-018-2095-4

Gupta, G., and Manoj, G. (2021). "Deep learning for brain tumor segmentation using magnetic resonance images," in Proceeding IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Melbourne, Australia, 13-15 October 2021 (IEEE), 1–6. doi:10.1109/CIBCB49929.2021.9562890

Gupta, S. (2021). *Computational prediction of cervical cancer diagnosis using ensemble-based classification algorithm*. doi:10.1093/comjnl/bxaa198

Gupta, S., and Gupta, M. (2021c). "Deep learning for brain tumor segmentation using magnetic resonance images," in ProceedingIEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, Melbourne, Australia, 13-15 October 2021 (IEEE). doi:10.1109/CIBCB49929.2021.9562890

Gupta, S., and Gupta, M. K. (2021d). A comparative analysis of deep learning approaches for predicting breast cancer survivability. *Archives Comput. Methods Eng.*, 1

Gupta, S., and Gupta, M. K. (2021a). A comprehensive data-level investigation of cancer diagnosis on imbalanced data. *Comput. Intell.* 38, 156–186. doi:10.1111/coin.12452

Gupta, S., and Gupta, M. K. (2021b). Computational model for prediction of malignant mesothelioma diagnosis. *Comput. J.* doi:10.1093/comjnl/bxab146

He, B., Luo, H., Zhou, Z., Wang, B., Liang, Y., Lang, J., et al. (2020). A neural network framework for predicting the tissue-of-origin of 15 common cancer types

based on RNA-seq data. *Front. Bioeng. Biotechnol.* 8 (8), 737–811. doi:10.3389/fbioe.2020.00737

Huang, Z., Johnson, T. S., Han, Z., Helm, B., Cao, S., Zhang, C., et al. (2020). Deep learning-based cancer survival prognosis from RNA-seq data : Approaches and evaluations. *BMC Med. Genomics* 13 (5), 41–12. doi:10.1186/s12920-020-0686-1

Huynh, P., Nguyen, V., and Do, T. (2019). Novel hybrid DCNN–SVM model for classifying RNA-sequencing gene expression data. *J. Inf. Telecommun.* 3, 533–547. doi:10.1080/24751839.2019.1660845

Id, J. L., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., et al. (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review. *PLoS One* 16, e0250370–23. doi:10.1371/journal.pone.0250370

Jerez, M., Franco, L., Veredas, F. J., and Lo, G. (2020). Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *Plos One* 15, e0230536–24. doi:10.1371/journal.pone.0230536

Joshi, P., and Park, T. (2019). "Cancer subtype classification based on superlayered neural network," in Proceeding IEEE International Conference on Bioinformatics and Biomedicine, San Diego, CA, USA, 18-21 November 2019 (IEEE), 1988–1992. doi:10.1109/BIBM47256.2019.8983343

Kashyap, D., Pal, D., Sharma, R., Garg, V. K., Goel, N., Koundal, D., et al. (2022). Global increase in breast cancer incidence: Risk Factors and preventive Measures. *Biomed. Res. Int.* 2022, 9605439. doi:10.1155/2022/9605439

Kim, B., Yu, K., and Lee, P. C. W. (2020). Cancer classification of single-cell gene expression data by neural network. *Bioinformatics* 36, 1360–1366. doi:10.1093/bioinformatics/btz772

Kong, Y., and Yu, T. (2018). A deep neural network model using random forest to extract feature representation for gene expression data classification. *Sci. Rep.* 8 (1), 16477. doi:10.1038/s41598-018-34833-6

Kumar, Y., Gupta, S., Singla, R., and Chen, Y. (2021). A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. *Arch. Comput. Methods Eng.* 29 (4), 2043–2070. doi:10.1007/s11831-021-09648-w

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553), 436–444. doi:10.1038/nature14539

Lin, M., Jaitly, V., Wang, I., Hu, Z., Chen, L., Wahed, M., et al. (2018). Application of deep learning on predicting prognosis of acute myeloid leukemia with cytogenetics age and mutations. *Mach. Learn.* Available at: https://arXiv/org/abs1810.13247.

Motieghader, H., Najafi, A., Sadeghi, B., and Masoudi-nejad, A. (2017). A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. *Inf. Med. Unlocked* 9, 246–254. doi:10.1016/j.imu.2017.10.004

Panda, M. (2017). Elephant search optimization combined with deep neural network for microarray data analysis. *J. King Saud Univ. - Comput. Inf. Sci.* 32, 940–948. doi:10.1016/j.jksuci.2017.12.002

Parvathavardhini, S., and Manju, S. (2020). Cancer gene detection using Neuro fuzzy classification algorithm. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* 3 (3), 2456

Reid, A., Klerk, N. De, and Musk, A. W. B. (2011). Does exposure to asbestos cause ovarian cancer ? A systematic literature review and meta-analysis. *Cancer Epidemiol. Biomarkers Prev.* 20, 1287–1295. doi:10.1158/1055-9965.EPI-10-1302

Ronoud, S., and Asadi, S. (2019). An evolutionary deep belief network extreme learning-based for breast cancer diagnosis. *Soft Comput.* 23. 13139–13159. doi:10.1007/s00500-019-03856-0

Salman, I., Ucan, O., Bayat, O., and Shaker, K. (2018). Impact of metaheuristic iteration on artificial neural network structure in medical data. *Process. (Basel).* 6, 57. doi:10.3390/pr6050057

Schiff, M., Castle, P. E., Jeronimo, J., Rodriguez, A. C., and Wacholder, S. (2007). Human papillomavirus and cervical cancer. *Clin. Microbiol. Rev.* 16, 1–17. doi:10.1128/CMR.16.1.1-17.2003

Sevakula, R. K., Singh, V., Member, S., Kumar, C., and Cui, Y. (2018). Transfer learning for molecular cancer classification using deep neural networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 5963, 2089–2100. doi:10.1109/TCBB.2018.2822803

Shimizu, H., and Nakayama, K. I. (2020). Artificial intelligence in oncology. *Cancer Sci.* 111 (5), 1452–1460. doi:10.1111/cas.14377

Shon, H. S., Yi, Y., Kim, K. O., Cha, E., and Kim, K. (2021). Classification of stomach cancer gene expression data using CNN algorithm of deep learning. *J. Biomed. Transl. Res.* 20 (1), 15–20. doi:10.12729/jbtr.2019.20.1.015

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020 : GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Surbhi Gupta, M. G. (2021). "Prostate cancer prognosis using multi-layer perceptron and class balancing techniques," in Proceeding 2021 Thirteenth International Conference on Contemporary Computing (IC3-2021) (IC3 '21), August 05-07, 2021 (New York NY USA: ACM), 1.

Torkey, H., Atlam, M., El-fishawy, N., and Salem, H. (2021). A novel deep autoencoder based survival analysis approach for microarray dataset. *Peer Comput. Sci.* 1, e492. doi:10.7717/peerj-cs.492

Tumuluru, P., and Ravi, B. (2017). "Goa-Based DBN : Grasshopper optimization algorithm-based deep belief neural networks for cancer classification Goa-based DBN : Grasshopper optimization algorithm-based deep belief neural networks for cancer classification," in Proceeding International Journal of Applied Engineering Research, 14218–14231.

Urda, D., and Moreno, F. (2017). "Deep learning to analyze RNA-seq gene expression data," in *International work-conference on artificial neural networks*, 50–59. Springer: Cham. doi:10.1007/978-3-319-59147-6

Wessels, F., Schmitt, M., Krieghoff-henning, E., Jutzi, T., Worst, T. S., Waldbillig, F., et al. (2021). Deep learning approach to predict lymph node metastasis directly from primary tumor histology in prostate cancer. *BJU Int.* 128, 352. doi:10.1111/bju.15386

Xiao, Y., Wu, J., Lin, Z., and Zhao, X. (2018a). A deep learning-based multi-model ensemble method for cancer prediction. *Comput. Methods Programs Biomed.* 153, 1–9. doi:10.1016/j.cmpb.2017.09.005

Xiao, Y., Wu, J., Lin, Z., and Zhao, X. (2018b). A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data. *Comput. Methods Programs Biomed.* 166, 99–105. doi:10.1016/j.cmpb.2018.10.004

Xu, J., Wu, P., Chen, Y., Meng, Q., Dawood, H., and Dawood, H. (2019a). 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., Wu, P., Chen, Y., Meng, Q., Dawood, H., and Khan, M. M. (2019b). A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data. *IEEE Access* 7, 22086–22095. doi:10.1109/ACCESS.2019.2898723

Yuan, Y., Shi, Y., Li, C., Kim, J., Cai, W., Han, Z., et al. (2016). DeepGene : An advanced cancer type classifier based on deep learning and somatic point mutations. *BMC Bioinforma.* 17 (17), 476. doi:10.1186/s12859-016-1334-9

Zahras, D. (2018).Cervical cancer risk classification based on deep convolutional neural network, In ProceedingInternational Conference on Applied Information Technology and Innovation, Padang, Indonesia, 03-05 September 2018,. IEEE, 149–153. doi:10.1109/ICAITI.2018.8686767