

# Adoption of artificial intelligence in human and clinical genomics

**Edited by**

Deepak Kumar Jain, Li Zhang, Guangming Zhang  
and Piyush Shukla

**Published in**

Frontiers in Genetics  
Frontiers in Molecular Biosciences



## FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714  
ISBN 978-2-8325-2184-7  
DOI 10.3389/978-2-8325-2184-7

## 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](https://frontiersin.org/about/contact)

# Adoption of artificial intelligence in human and clinical genomics

## Topic editors

Deepak Kumar Jain — Chongqing University of Posts and Telecommunications, China

Li Zhang — University of London, United Kingdom

Guangming Zhang — University of Texas Health Science Center at Houston, United States

Piyush Shukla — Rajeev Gandhi Technical University, India

## Citation

Jain, D. K., Zhang, L., Zhang, G., Shukla, P., eds. (2023). *Adoption of artificial intelligence in human and clinical genomics*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2184-7

## Table of contents

- 04 **Hsa\_circ\_0062270 Promotes Tumorigenesis of Melanoma by Stabilizing the Linear Transcript Cell Division Cycle Protein 45**  
Cuie Wei, Wentao Sun, Changhai Liu, Fanjun Meng, Lele Sun and Xiangsheng Ding
- 11 **Effect of Normobaric Oxygen Inhalation Intervention on Microcirculatory Blood Flow and Fatigue Elimination of College Students After Exercise**  
Yong Peng, Liang Meng, Huan Zhu, Li Wan and Fen Chen
- 19 **Research on Adoption Behavior and Influencing Factors of Intelligent Pension Services for Elderly in Shanghai**  
Juan Luo and Lingqi Meng
- 28 **Effect of Aerobic Exercise on Lipid Metabolism in Rats With NAFLD**  
Tongxi Zhou, Mengfan Niu, Ruichen Liu and Li Li
- 34 **Performance of Problem-Based Learning Based Image Teaching in Clinical Emergency Teaching**  
Xiaohong Xu, Yingcui Wang, Suhua Zhang and Fengting Liu
- 45 **Looking for the Genes Related to Lung Cancer From Nasal Epithelial Cells by Network and Pathway Analysis**  
Noman Qureshi, Jincheng Chi, Yanan Qian, Qianwen Huang and Shaoyin Duan
- 54 **Mathematical model and genomics construction of developmental biology patterns using digital image technology**  
Shiwei Ni, Fei Chen, Guolong Chen and Yufeng Yang
- 69 **Knowledge structure and emerging trends in the application of deep learning in genetics research: A bibliometric analysis [2000–2021]**  
Bijun Zhang and Ting Fan
- 82 **Study on complications of osteoporosis based on network pharmacology**  
Zhijing Song, Haoling Zhang, Yuhang Jiang, Rui Zhao, Xuedong Pei, Haochi Ning, Hailiang Chen, Jing Pan, Yanlong Gong, Min Song and Wei Wang
- 97 **Semi-supervised segmentation of metastasis lesions in bone scan images**  
Qiang Lin, Runxia Gao, Mingyang Luo, Haijun Wang, Yongchun Cao, Zhengxing Man and Rong Wang
- 110 **A comprehensive survey on computational learning methods for analysis of gene expression data**  
Nikita Bhandari, Rahee Walambe, Ketan Kotecha and Satyajeet P. Khare





# Hsa\_circ\_0062270 Promotes Tumorigenesis of Melanoma by Stabilizing the Linear Transcript Cell Division Cycle Protein 45

Cuie Wei<sup>1†</sup>, Wentao Sun<sup>2†</sup>, Changhai Liu<sup>1</sup>, Fanjun Meng<sup>1</sup>, Lele Sun<sup>1</sup> and Xiangsheng Ding<sup>1\*</sup>

<sup>1</sup>Department of Plastic Surgery, The First People's Hospital of Lianyungang, Lianyungang, China, <sup>2</sup>Department of Plastic Surgery, The First Affiliated Hospital of Kunming Medical University, Kunming, China

## OPEN ACCESS

### Edited by:

Deepak Kumar Jain,  
Chongqing University of Posts and  
Telecommunications, China

### Reviewed by:

Jian Tang,  
Nanjing Medical University, China  
Weiguo Wang,  
Affiliated Hospital of Shandong  
University of Traditional Chinese  
Medicine, China

### \*Correspondence:

Xiangsheng Ding  
ding888364035140@126.com

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Human and Medical Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 16 March 2022

**Accepted:** 25 April 2022

**Published:** 10 May 2022

### Citation:

Wei C, Sun W, Liu C, Meng F, Sun L  
and Ding X (2022) Hsa\_circ\_0062270  
Promotes Tumorigenesis of Melanoma  
by Stabilizing the Linear Transcript Cell  
Division Cycle Protein 45.  
Front. Genet. 13:897440.  
doi: 10.3389/fgene.2022.897440

**Background:** To elucidate the potential biological function of hsa\_circ\_0062270 in the malignant process of melanoma and its potential target.

**Methods:** Quantitative real-time polymerase chain reaction (qRT-PCR) was conducted to examine relative level of hsa\_circ\_0062270 in melanoma tissues and normal skin tissues. The diagnostic and prognostic potentials of hsa\_circ\_0062270 in melanoma were evaluated. The regulatory effect of hsa\_circ\_0062270 on the expression of linear transcript Cell division cycle protein 45 (CDC45) was also examined.

**Results:** Hsa\_circ\_0062270 was up-regulated in melanoma samples and cell lines, which displayed certain diagnostic and prognostic potentials in melanoma. Inhibition of hsa\_circ\_0062270 attenuated the proliferative, migratory and invasive functions. Hsa\_circ\_0062270 could stabilize the expression of linear transcript CDC45, and thus participated in the malignant process of melanoma.

**Conclusion:** Hsa\_circ\_0062270 promotes proliferative, migratory and invasive functions of melanoma cells via stabilizing the linear transcript CDC45. Hsa\_circ\_0062270 can be used to diagnosis and treatment of melanoma.

**Keywords:** hsa\_circ\_0062270, melanoma, Cdc45, proliferation, metastasis

## INTRODUCTION

Melanoma is a highly malignant skin cancer derived from melanocytes. Melanoma accounts for more than 70% of skin cancer deaths (Testori et al., 2017). More seriously, estimated numbers of new cases and death cases of melanoma are on the rise. Surgical resection combined lymph node dissection is an effective treatment for early stage melanoma (Chattopadhyay et al., 2016). Nevertheless, most of melanoma patients cannot be operated because of metastasis, leading to a poor 5-year survival (20%) (Shain and Bastian, 2016). Molecular mechanisms underlying melanoma process and metastasis remain largely unclear.

CircRNAs are novel noncoding RNAs to be widely analyzed. They are extensively involved in various fields of life sciences (Hsiao et al., 2017; Chen and Huang, 2018; Han et al., 2018). Since circRNAs do not have 3' end, 5' end and poly (A) tail, they can escape degradation from exonucleases. CircRNAs are more conservative and stable than linear RNAs. Moreover, they are evolutionarily conserved in different species, and specifically expressed in different tissues and

developmental stages (Dong et al., 2019; Patop et al., 2019). According to the origins, circRNAs are classified to circular exonic RNAs that only contain reverse-splicing exons, circular intronic RNAs (ciRNAs) that only contain reverse-splicing introns, and exon-intron circRNAs (ElciRNAs) that introns are remained in circular exons (Salzman, 2016). The following mechanisms explain the biological functions of circRNAs: 1) CircRNAs inhibit miRNA activities through exerting the miRNA sponge effect; 2) CircRNAs interact with RNA-binding proteins as protein sponges; 3) CircRNAs regulate Pol II transcription of parent genes; 4) CircRNAs regulate linear splicing through competitively targeting splicing sites of pre-mRNAs; 5) CircRNAs have protein-encoding ability and can translate proteins (Kristensen et al., 2019). Abundant evidences have proven the vital functions of circRNAs in physiological and pathological processes (Salzman, 2016; Qu et al., 2017).

Hsa\_circ\_0062270 is located on chromosome 22: 19496052-19502571, and its gene symbol is CDC45. Evidence has showed that hsa\_circ\_0062270 is obviously up-regulated in melanoma (Hao et al., 2021). A previous study have demonstrated that downregulation of hsa\_circ\_0062270 can inhibit the progression of melanoma, however, the mechanism still remains unclear (Hao et al., 2021). The aim of the current research was to explore the biological effect of hsa\_circ\_0062270 on malignant phenotypes of melanoma and its potential target.

## PATIENTS AND METHODS

### Subjects and Specimens

The normal skin tissues and melanoma tissues of 50 patients with melanoma in our hospital were selected. The ethics committee of The First People's Hospital of Lianyungang approved our study. Signed written informed consents were obtained from all participants before the study.

### Cell Culture

Melanoma cells (SKMEL1, A375, A2058 and A875) and normal human epidermal melanocytes (NHEM) were provided by Cell Bank of Type Culture Collection (Shanghai, China). Cells were cultivated in DMEM containing 10% fetal bovine serum (FBS), 100 U/mL penicillin and 100 µg/ml streptomycin at 5% CO<sub>2</sub>, 37°C. Cell transfection was performed using Lipofectamine 3,000 as per the protocols. Cell proliferation was determined by EdU (Beyotime, Shanghai, China).

### Quantitative Real-Time Polymerase Chain Reaction

RNAs isolation was done with TRIzol and were then reversely transcribed into cDNAs. U6 and GAPDH were used as the internal controls with the method of  $2^{-\Delta\Delta C_t}$ . Primers used were shown below: hsa\_circ\_0062270: Forward: 5'-AGGATGGCTCAGGGACAGAT-3', reverse: 5'-AGGCCATGGTACAGCTTGTC-3'; CDC45: Forward: 5'-TTCGTGTCCGATTTCGCAAA-3', reverse: 5'-TGGAACCAGCGTATATTGCAC-3'; GAPDH: Forward: 5'-CGGAGTCAACGGATTTGGTCG

TAT-3', reverse: 5'-AGCCTTCTCCATGGTGGTGAAGAC-3'; U6: Forward: 5'-GCTGAGGTGACGGTCTCAAA-3', reverse: 5'-GCCTCCCAGTTTCATGGACA-3'.

### Actinomycin D and Rnase R Assays

A375 cells were exposed to Actinomycin D (3 µg/ml). They were collected for isolating total RNAs. Expressions of hsa\_circ\_0062270 and CDC45 were detected by Quantitative real-time polymerase chain reaction. Cellular RNA (4 mg) was treated either with RNase R (10 U/µg) at 37°C for 30 min or not, followed by purification using RNeasy MinElute (Qiagen, Hilden, Germany).

### Cell Transwell Assay

Cells were seeded into the top chamber and bottom chamber. After 48-h incubation, cells in the bottom were fixed, dyed in crystal violet and captured. Migratory cells were counted in five randomly selected fields per sample. Invasion assay was conducted using transwell chamber precoated with 100 µL of Matrigel (Corning, Corning, NY, United States). In detail, Matrigel was diluted in serum-free medium at 1:3, which was coated on the top of a chamber.

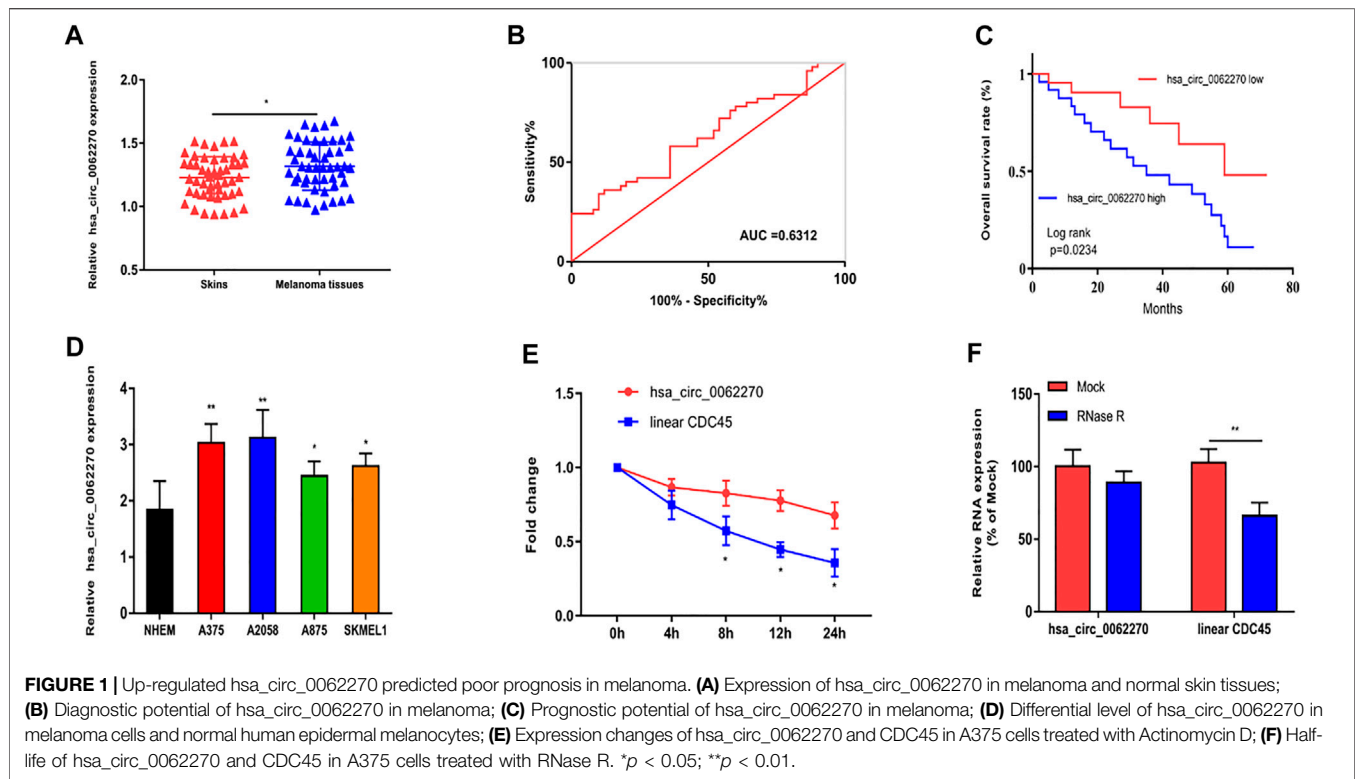
### Statistical Analysis

Data were expressed as mean ± SD (standard deviation) and they were processed using Statistical Product and Service Solutions (SPSS) 20.0 (IBM, Armonk, NY, United States). Prognostic value of hsa\_circ\_0062270 in melanoma were evaluated by Kaplan-Meier and receiver operating characteristic (ROC) method, respectively. The correlation between hsa\_circ\_0062270 and CDC45 levels was assessed through Pearson correlation test. A significant difference was set at  $p < 0.05$ .

## RESULTS

### Up-Regulated hsa\_circ\_0062270 Predicted Poor Prognosis of Melanoma

Firstly, we explored the expression of hsa\_circ\_0062270 in melanoma tissues and the normal skin tissues. Results revealed that the expression hsa\_circ\_0062270 in melanoma tissues was significantly higher than that in normal ones (**Figure 1A**). Then, receiver operating characteristic (ROC) curves depicted that the AUC of hsa\_circ\_0062270 was 0.6312 (**Figure 1B**). Kaplan-Meier analysis found that highly expressed hsa\_circ\_0062270 predicted a poor prognosis for melanoma patients ( $p = 0.0234$ ) (**Figure 1C**). Hsa\_circ\_0062270 significantly increased in melanoma cell lines as well (**Figure 1D**). We furthermore examined the stability of hsa\_circ\_0062270 by detecting mRNA levels of hsa\_circ\_0062270 and CDC45 in Actinomycin D-treated A375 cells. Compared with that of CDC45 (<12 h), the half-life of hsa\_circ\_0062270 was over 24 h (**Figure 1E**). RNase R induction did not affect expression level of hsa\_circ\_0062270, but markedly down-regulated CDC54 (**Figure 1F**). Therefore, we have verified that hsa\_circ\_0062270 was highly stable in melanoma cells.



## Hsa\_circ\_0062270 Promoted Melanoma to Proliferate, Migrate and Invade

To investigate the effects of hsa\_circ\_0062270 on melanoma cell proliferation, migration and invasion, cells were treated with hsa\_circ\_0062270 siRNA. Results indicated that transfection of hsa\_circ\_0062270 siRNA markedly down-regulated hsa\_circ\_0062270 level in A375 and A2058 cells (**Figure 2A**). Knockdown of hsa\_circ\_0062270 reduced EdU-positive rate in melanoma cells (**Figure 2B**). In addition, hsa\_circ\_0062270 siRNA markedly reduced migratory and invasive rates (**Figures 2C,D**).

## Hsa\_circ\_0062270 Stabilized CDC45 Expression

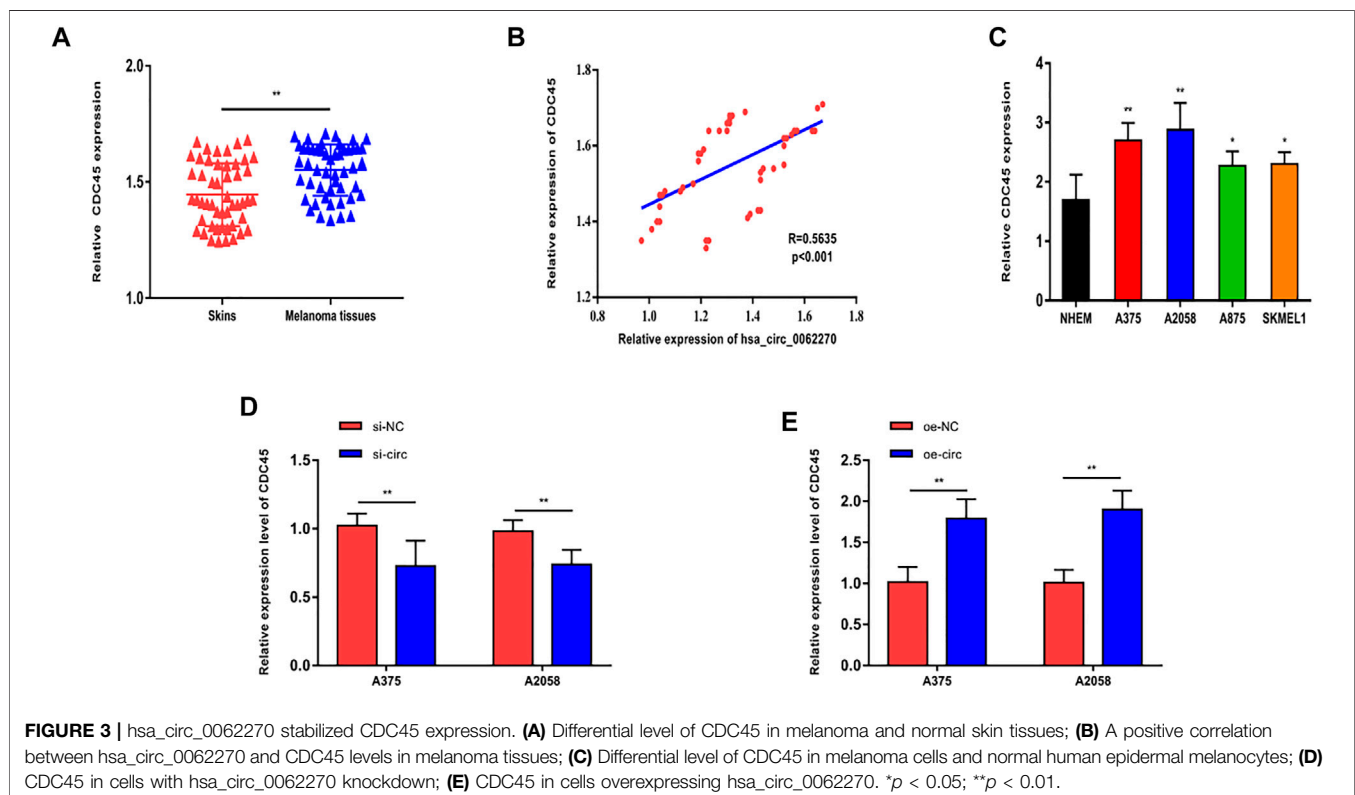
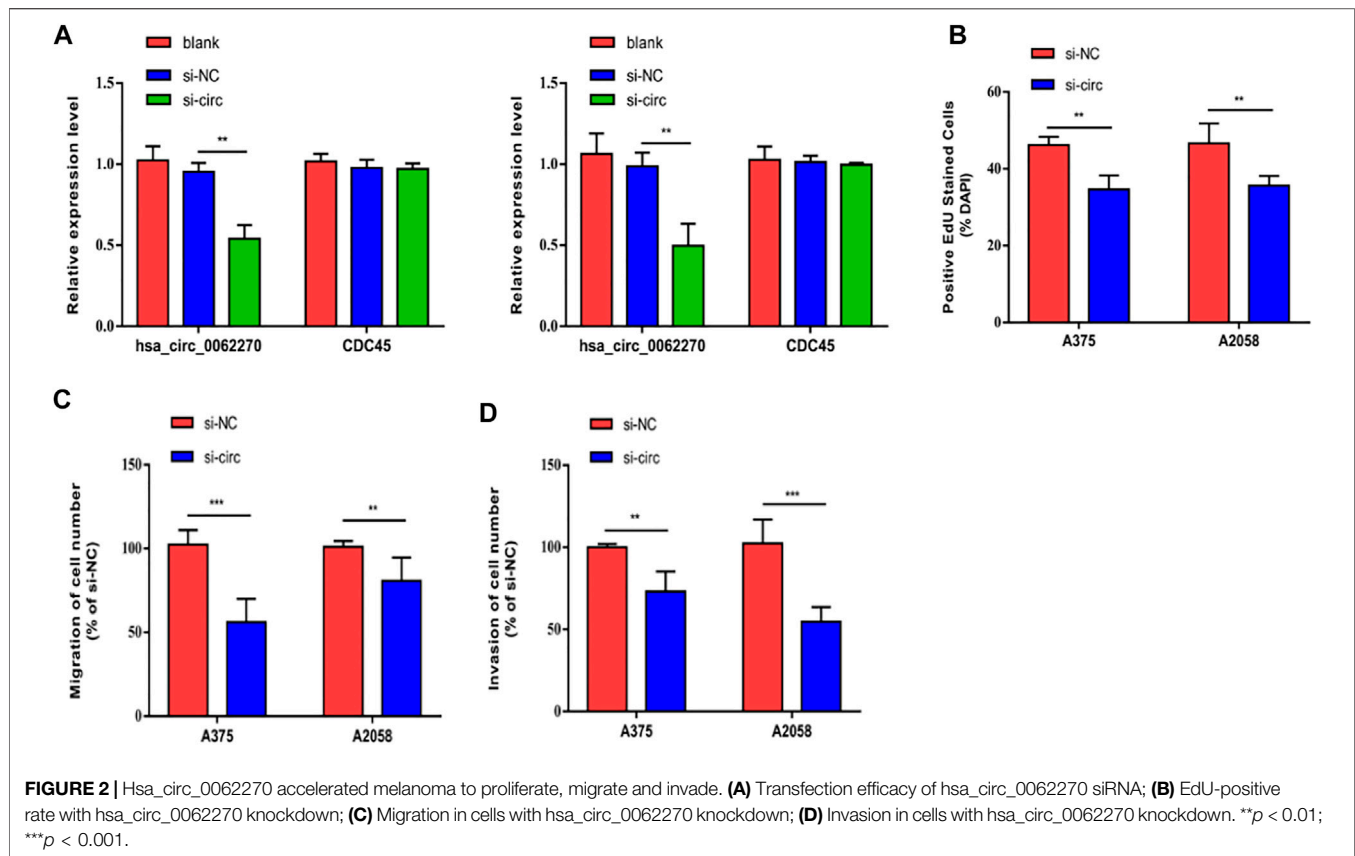
Then we focused on the potential target of hsa\_circ\_0062270 in the regulation of phenotypes of melanoma. CircRNAs are involved in pathological process *via* mediating expression levels of their linear transcripts. CDC45 was the linear transcript of hsa\_circ\_0062270 (**Figure 3A**) and positively correlated to hsa\_circ\_0062270 level (**Figure 3B**). Identically, CDC45 was highly expressed in melanoma cell lines (**Figure 3C**). Knockdown of hsa\_circ\_0062270 could downregulate CDC45 and as expected, CDC45 was up-regulated in A375 and A2058 cells overexpressing hsa\_circ\_0062270 (**Figures 3D,E**). It is concluded that hsa\_circ\_0062270 could stabilize the expression of CDC45.

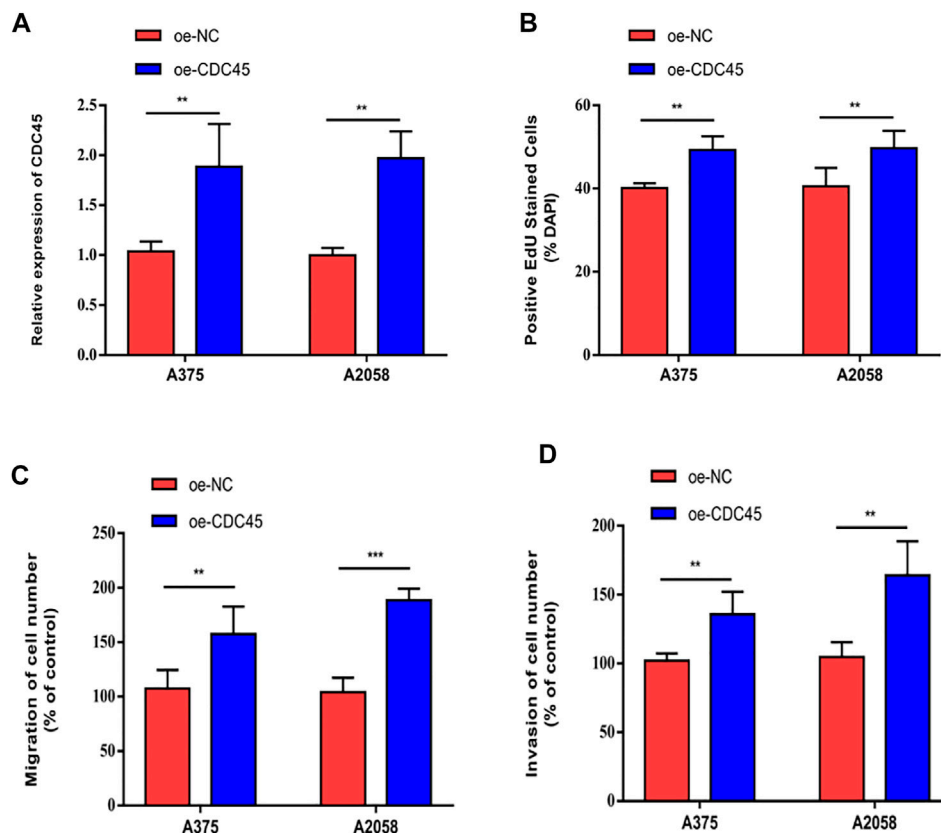
## CDC45 Promoted Melanoma to Proliferate, Migrate and Invade

To further elucidate the effects of CDC45 on melanoma phenotypes, we established the overexpression models of CDC45. Transfection of overexpressed plasmid of CDC45 effectively up-regulated CDC45 in A375 and A2058 cells (**Figure 4A**). In melanoma cells overexpressing CDC45, EdU-positive rate increased (**Figure 4B**). Moreover, migratory and invasive potentials of melanoma were promoted by overexpressed CDC45 (**Figures 4C,D**). Therefore, CDC45 could stimulate the malignant process of melanoma.

## Hsa\_circ\_0062270 Stimulated the Malignant Process of Melanoma by Stabilizing CDC45

To uncover the co-regulation of hsa\_circ\_0062270 and CDC45, cells were co-transfected using si-CDC45 and overexpressed plasmid of hsa\_circ\_0062270 (**Figure 5A**). Overexpression of hsa\_circ\_0062270 could enhance the down-regulated CDC45 in melanoma cells transfected with si-CDC45. Knockdown of CDC45 reduced EdU-positive rate, migratory cell number and invasive cell number, which were reversed by overexpressed hsa\_circ\_0062270 (**Figures 5B–D**). All above indicated that hsa\_circ\_0062270 may play a carcinogenic role in melanoma by stabilizing its linear transcription CDC45.





**FIGURE 4 |** CDC45 promoted melanoma to proliferate, migrate and invade. **(A)** Transfection efficacy of overexpressed plasmid of CDC45; **(B)** EdU-positive rate in cells overexpressing CDC45; **(C)** Migration in cells overexpressing CDC45; **(D)** Invasion in cells overexpressing CDC45. \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ .

## DISCUSSION

The incidence of melanoma is not high, covering 4–5% of malignant tumors. Family history, multiple atypical moles and dysplastic moles are risk factors that trigger the carcinogenesis of melanoma (Pavri et al., 2016). In addition, ultraviolet rays can induce melanoma by damaging DNA repair genes.

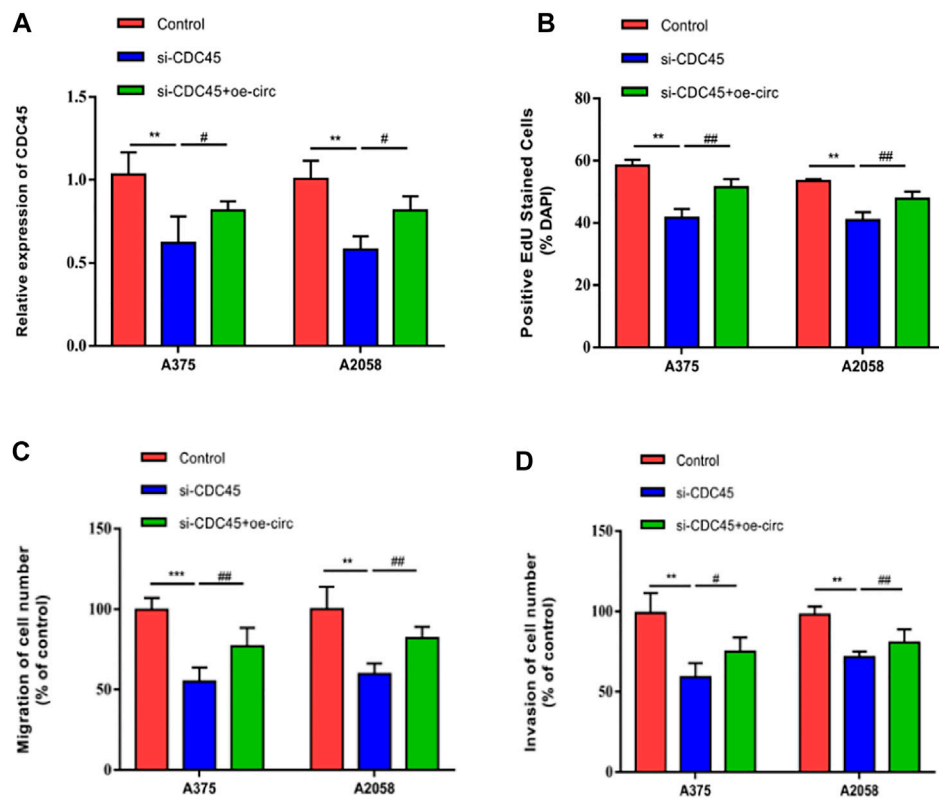
CircRNAs, as a type of emerging noncoding RNAs, have been well concerned because of their unique structure and vital functions (Meng et al., 2017; Chen and Huang, 2018). They can be utilized as molecular biomarkers for diagnosing tumors and evaluating their prognosis (Bian et al., 2018; Zhou et al., 2018; Chen et al., 2020). The involvement of circRNAs in melanoma has been reported through literature review.

We previously found that the expression of hsa\_circ\_0062270 in melanoma was up-regulated. The diagnostic and prognostic potentials of hsa\_circ\_0062270 in melanoma were verified through depicting ROC and Kaplan-Meier curves, respectively. *In vitro* experiment results illustrated that knockdown of hsa\_circ\_0062270 remarkably suppressed proliferative, migratory and invasive functions of melanoma cells.

Recent studies have demonstrated that circRNAs have an important role in disease progression by mediating expressions of their linear transcripts (Wu et al., 2020; Tang et al., 2021;

Mecozzi et al., 2021; Peng and Wang, 2021). Through qRT-PCR and rescue experiments, we found that hsa\_circ\_0062270 was able to stabilize the expression of its linear transcript CDC45. Knockdown of CDC45 blocked proliferative, migratory and invasive functions of melanoma cells, which could be reversed by overexpression of hsa\_circ\_0062270. The binding of CDC45 to chromatin is regulated by the convergent effect of CDKs and DDK coincides with the time point of the start of DNA replication, suggesting that CDC45 is crucial in regulating the initiation of DNA replication (Köhler et al., 2016; Szambowska et al., 2017; Can et al., 2019).

Collectively, our present study was the first attempt to reveal that hsa\_circ\_0062270 was up-regulated in melanoma specimens and correlated to its prognosis. Hsa\_circ\_0062270 stimulated malignant process of melanoma by stabilizing its linear transcript CDC45. Our findings provide a new aspect for developing diagnostic and therapeutic strategies for melanoma. Several limitations of our study should be pointed out. First of all, *in vivo* role of hsa\_circ\_0062270 in melanoma is not explored. Secondly, how hsa\_circ\_0062270 regulates CDC45 remains unclear. Thirdly, other cell phenotypes of melanoma, including apoptosis, epithelial-mesenchymal transition and cell cycle progression affected by hsa\_circ\_0062270 are not clear.



**FIGURE 5 |** hsa\_circ\_0062270 stimulated the malignant process of melanoma by stabilizing CDC45. **(A)** Co-transfection of si-CDC34 and overexpressed plasmid of hsa\_circ\_0062270; **(B)** Edu-positive rate co-regulated by hsa\_circ\_0062270 and CDC45; **(C)** Migrative ability regulated by hsa\_circ\_0062270 and CDC45; **(D)** Invasion in A375 and A2058 cells co-regulated by hsa\_circ\_0062270 and CDC45. \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ ; # $p < 0.05$ ; ## $p < 0.01$ .

## CONCLUSION

Hsa\_circ\_0062270 promotes proliferative, migrative and invasive functions in melanoma cells *via* stabilizing the linear transcript CDC45. These findings provided strong evidence that hsa\_circ\_0062270 could be a novel promising therapeutic target used to diagnosis and treatment of melanoma.

## 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.

## REFERENCES

- Bian, D., Wu, Y., and Song, G. (2018). Novel Circular RNA, Hsa\_circ\_0025039 Promotes Cell Growth, Invasion and Glucose Metabolism in Malignant Melanoma via the miR-198/CDK4 axis. *Biomed. Pharmacother.* 108, 165–176. doi:10.1016/j.biopha.2018.08.152
- Can, G., Kauerhof, A. C., Macak, D., and Zegerman, P. (2019). Helicase Subunit Cdc45 Targets the Checkpoint Kinase Rad53 to Both Replication Initiation and Elongation Complexes after Fork Stalling. *Mol. Cell* 73 (3), 562–573. doi:10.1016/j.molcel.2018.11.025

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of the First People's Hospital of Lianyungang. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

CW, WS, and XD designed the study and performed the experiments, CL and FM collected the data, CL, FM, and LS analyzed the data, CW, WS, and XD prepared the manuscript. All authors read and approved the final manuscript.

- Chattopadhyay, C., Kim, D. W., Gombos, D. S., Oba, J., Qin, Y., Williams, M. D., et al. (2016). Uveal Melanoma: From Diagnosis to Treatment and the Science in between. *Cancer* 122 (15), 2299–2312. doi:10.1002/cnrc.29727
- Chen, B., and Huang, S. (2018). Circular RNA: An Emerging Non-coding RNA as a Regulator and Biomarker in Cancer. *Cancer Lett.* 418, 41–50. doi:10.1016/j.canlet.2018.01.011
- Chen, Z., Chen, J., Wa, Q., He, M., Wang, X., Zhou, J., et al. (2020). Knockdown of Circ\_0084043 Suppresses the Development of Human Melanoma Cells through miR-429/tribbles Homolog 2 axis and Wnt/ $\beta$ -Catenin Pathway. *Life Sci.* 243, 117323. doi:10.1016/j.lfs.2020.117323



- Dong, W., Dai, Z.-h., Liu, F.-c., Guo, X.-g., Ge, C.-m., Ding, J., et al. (2019). The RNA-Binding Protein RBM3 Promotes Cell Proliferation in Hepatocellular Carcinoma by Regulating Circular RNA SCD-circRNA 2 Production. *Ebiomedicine* 45, 155–167. doi:10.1016/j.ebiom.2019.06.030
- Han, B., Chao, J., and Yao, H. (2018). Circular RNA and its Mechanisms in Disease: From the Bench to the Clinic. *Pharmacol. Ther.* 187, 31–44. doi:10.1016/j.pharmthera.2018.01.010
- Hao, T., Yang, Y., He, J., Bai, J., Zheng, Y., and Luo, Z. (2021). Knockdown of Circular RNA Hsa\_circ\_0062270 Suppresses the Progression of Melanoma via Downregulation of CDC45. *Histol. Histopathol.*, 18412. doi:10.14670/HH-18-412
- Hsiao, K.-Y., Sun, H. S., and Tsai, S.-J. (2017). Circular RNA - New Member of Noncoding RNA with Novel Functions. *Exp. Biol. Med. (Maywood)* 242 (11), 1136–1141. doi:10.1177/1535370217708978
- Köhler, C., Koalick, D., Fabricius, A., Parplys, A. C., Borgmann, K., Pospiech, H., et al. (2016). Cdc45 is Limiting for Replication Initiation in Humans. *Cell Cycle* 15 (7), 974–985. doi:10.1080/15384101.2016.1152424
- Kristensen, L. S., Andersen, M. S., Stagsted, L. V. W., Ebbesen, K. K., Hansen, T. B., and Kjems, J. (2019). The Biogenesis, Biology and Characterization of Circular RNAs. *Nat. Rev. Genet.* 20 (11), 675–691. doi:10.1038/s41576-019-0158-7
- Mecozi, N., Vera, O., and Karreth, F. A. (2021). Squaring the Circle: circRNAs in Melanoma. *Oncogene* 40 (37), 5559–5566. doi:10.1038/s41388-021-01977-1
- Meng, X., Li, X., Zhang, P., Wang, J., Zhou, Y., and Chen, M. (2017). Circular RNA: An Emerging Key Player in RNA World. *Brief. Bioinform* 18 (4), bbw045–57. doi:10.1093/bib/bbw045
- Patop, I. L., Wüst, S., and Kadener, S. (2019). Past, Present, and Future of Circ RNA. *S. Embo J.* 38 (16), e100836. doi:10.15252/embj.2018100836
- Pavri, S. N., Clune, J., Ariyan, S., and Narayan, D. (2016). Malignant Melanoma: Beyond the Basics. *Plastic Reconstr. Surg.* 138 (2), 330e–340e. doi:10.1097/PRS.0000000000002367
- Peng, Q., and Wang, J. (2021). Non-coding RNAs in Melanoma: Biological Functions and Potential Clinical Applications. *Mol. Ther. Oncol.* 22, 219–231. doi:10.1016/j.omto.2021.05.012
- Qu, S., Zhong, Y., Shang, R., Zhang, X., Song, W., Kjems, J., et al. (2017). The Emerging Landscape of Circular RNA in Life Processes. *RNA Biol.* 14 (8), 992–999. doi:10.1080/15476286.2016.1220473
- Salzman, J. (2016). Circular RNA Expression: Its Potential Regulation and Function. *Trends Genet.* 32 (5), 309–316. doi:10.1016/j.tig.2016.03.002
- Shain, A. H., and Bastian, B. C. (2016). From Melanocytes to Melanomas. *Nat. Rev. Cancer* 16 (6), 345–358. doi:10.1038/nrc.2016.37
- Szambowska, A., Tessmer, I., Prus, P., Schlott, B., Pospiech, H., and Grosse, F. (2017). Cdc45-induced Loading of Human RPA onto Single-Stranded DNA. *Nucleic Acids Res.* 45 (6), gkw1364. doi:10.1093/nar/gkw1364
- Tang, K., Zhang, H., Li, Y., Sun, Q., and Jin, H. (2021). Circular RNA as a Potential Biomarker for Melanoma: A Systematic Review. *Front. Cell Dev. Biol.* 9, 638548. doi:10.3389/fcell.2021.638548
- Testori, A., Ribero, S., and Bataille, V. (2017). Diagnosis and Treatment of In-Transit Melanoma Metastases. *Eur. J. Surg. Oncol.* 43 (3), 544–560. doi:10.1016/j.ejso.2016.10.005
- Wu, X., Xiao, Y., Ma, J., and Wang, A. (2020). Circular RNA: A Novel Potential Biomarker for Skin Diseases. *Pharmacol. Res.* 158, 104841. doi:10.1016/j.phrs.2020.104841
- Zhou, M.-y., Yang, J.-M., and Xiong, X.-d. (2018). The Emerging Landscape of Circular RNA in Cardiovascular Diseases. *J. Mol. Cell. Cardiol.* 122, 134–139. doi:10.1016/j.yjmcc.2018.08.012

**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.

Copyright © 2022 Wei, Sun, Liu, Meng, Sun and Ding. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Effect of Normobaric Oxygen Inhalation Intervention on Microcirculatory Blood Flow and Fatigue Elimination of College Students After Exercise

Yong Peng<sup>1,2</sup>, Liang Meng<sup>3\*</sup>, Huan Zhu<sup>1</sup>, Li Wan<sup>1</sup> and Fen Chen<sup>1</sup>

<sup>1</sup>School of Physical Education, Hubei Minzu University, Enshi, China, <sup>2</sup>Graduate Schools, Moscow State Academy of Physical Culture, Malakhovka, Russia, <sup>3</sup>Sports Department, Suzhou University of Science and Technology, Suzhou, China

## OPEN ACCESS

### Edited by:

Deepak Kumar Jain,  
Chongqing University of Posts and  
Telecommunications, China

### Reviewed by:

Kangjie Sun,  
Shanghai Jiaotong University, China  
Xuetao Li,  
Hubei University of Automotive  
Technology, China

### \*Correspondence:

Liang Meng  
mengliang@usts.edu.cn

### Specialty section:

This article was submitted to  
Human and Medical Genomics,  
a section of the journal  
Frontiers in Genetics

Received: 22 March 2022

Accepted: 06 May 2022

Published: 01 June 2022

### Citation:

Peng Y, Meng L, Zhu H, Wan L and  
Chen F (2022) Effect of Normobaric  
Oxygen Inhalation Intervention on  
Microcirculatory Blood Flow and  
Fatigue Elimination of College Students  
After Exercise.  
Front. Genet. 13:901862.  
doi: 10.3389/fgene.2022.901862

**Objective:** To explore the effect of normobaric oxygen inhalation intervention on microcirculatory blood flow of college students after exercise and the impact of the elimination of exercise-induced fatigue, to provide a theoretical and methodological reference for the rapid elimination of fatigue of college students after endurance exercise.

**Methods:** Forty-eight male non-sports majors of Hubei University for nationalities were randomly divided into the control group ( $n = 24$ ) and intervention group ( $n = 24$ ). The subjects in both groups completed the same exercise program twice (running 3,000 m on the treadmill at maximum speed). After running, the issues in the intervention group inhaled portable oxygen for 30 min, and the control group recovered naturally. Microcirculatory blood flow (MBP), blood flow velocity (AVBC), blood flow concentration (CMBC), muscle oxygen saturation (SmO<sub>2</sub>), heart rate (HR), blood lactic acid (BLA), blood urea (BU), and creatine kinase (CK) were measured before exercise, immediately after exercise and 30 min after exercise.

**Results:** 1) MBP and AVBC had interaction between groups and time before and after exercise, MBP and AVBC immediately after exercise in the intervention group were significantly higher than those before exercise and 30 min after exercise, and 30 min after exercise in the intervention group were significantly higher than those in the control group. 2) SmO<sub>2</sub>, HR, BLA, BU, and CK had interaction between groups and time, and SmO<sub>2</sub> immediately after exercise in the intervention group was significantly lower than that before exercise and 30 min after exercise, but significantly higher than that in the control group at 30 min after exercise. The HR and BLA immediately after exercise in the intervention group were significantly higher than those before exercise and 30 min after exercise, but significantly lower than those in the control group at 30 min after exercise, and the BU and CK in the intervention group were significantly higher than those before exercise, but significantly lower than those in the control group at 30 min after exercise.

**Conclusion:** Normobaric oxygen inhalation for 30 min after exercise can delay the decrease of microcirculatory blood flow, increase muscle oxygen saturation, and



promote the recovery of heart rate, blood lactic acid, blood urea and creatine kinase. Therefore, normobaric oxygen inhalation for 30 min after exercise can be used as an effective means to promote the elimination of exercise-induced fatigue after endurance running.

**Keywords:** atmospheric oxygen inhalation, microcirculatory blood flow, muscle oxygen saturation, exercise-induced fatigue, fatigue elimination

## 1 INTRODUCTION

In recent years, with the continuous decline of college students' physical health, the probability of adverse cardiovascular events after exercise has gradually increased. The fatigued state of high stress after high intensity exercise is the main factor leading to sports injury and damaging cardiovascular disease of college students, especially when students' physique is poor, high intensity exercise brings more significant risks. Therefore, the rapid recovery of physical function and the immediate elimination of fatigue after high intensity exercise is of great significance to reduce the occurrence of adverse sports injuries of college students. Atmospheric oxygen inhalation, as an effective method to promote the elimination of fatigue after exercise, is not only practical, but also convenient and straightforward to operate, and has been widely used in competitive sports. Yang Yantao's study found that rapid oxygen inhalation after competition can significantly reduce the levels of serum creatine kinase and lactate dehydrogenase and increase the content of immunoglobulin (Yang et al., 2016). Oxygen inhalation before and after the match can promote the clearance of lactic acid after the match, improve the ability of antioxidation and accelerate the elimination of exercise-induced fatigue (Xie et al., 2016). Other studies have shown that oxygen inhalation in middle-and long-distance running can reduce students' cardiovascular stress levels, which has a positive effect on students' psychology (Zhang, 2019). Although oxygen inhalation contributes to the elimination of fatigue after exercise, the timely elimination of metabolic waste after exercise and the rapid transport of energy and oxygen to skeletal muscle and other organs are essential factors that determine the effect of fatigue elimination. Microcirculation is the only place for material and energy exchange, and its blood perfusion level is an essential condition to complete this process. The previous study of the author's team shows a close relationship between microcirculation blood flow and the occurrence of exercise-induced fatigue, and increasing microcirculation blood flow is helpful to reduce the occurrence of fatigue (Zhu and Binghong, 2016; Zhang et al., 2017; Zhu and Gao, 2019). However, at present, there are few studies on the effect of oxygen inhalation on microcirculation blood flow after exercise, and it is not clear whether oxygen inhalation can improve microcirculation blood flow after exercise. Based on this, through the intervention of rapid oxygen inhalation of college students after 3000-m running, this study discusses the effect of oxygen inhalation on microcirculatory blood flow and fatigue elimination after exercise, to provide a theoretical and methodological reference for the rapid elimination of fatigue after endurance exercise.

## 2 RESEARCH OBJECTS AND METHODS

### 2.1 Research Object

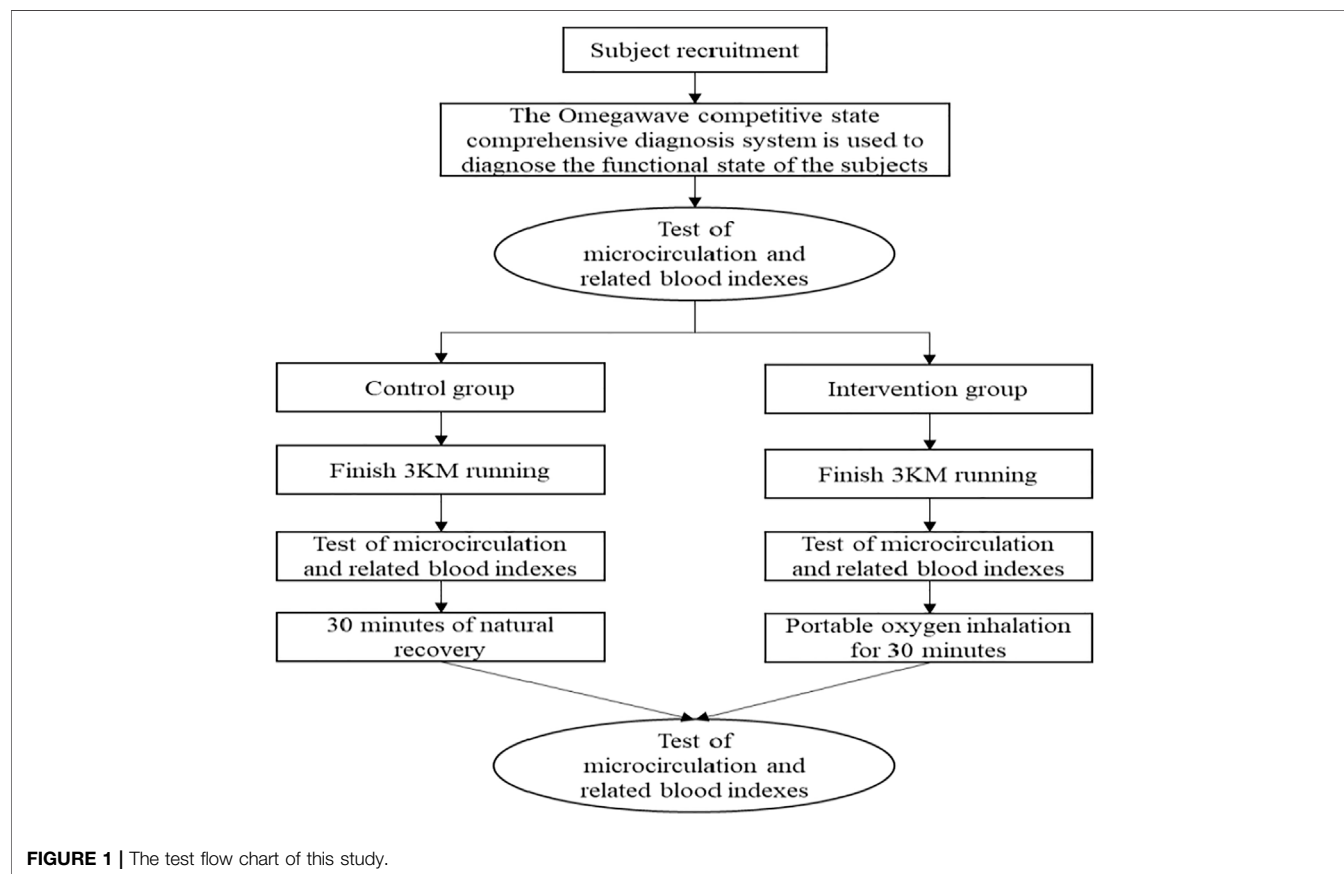
Forty-eight non-sports male students of Hubei University for nationalities were selected as the research subjects, which met the following criteria: 1) after the unified physical examination of the research group (the physical examination center of the affiliated Hospital of Hubei University for nationalities), there were no cardiovascular diseases such as hypertension and coronary heart disease, respiratory diseases, chronic kidney disease, and liver disease; 2) BMI <24; 3) no history of operation, a limb movement disorder. 4) exercise at least 1–2 times a week recently, and be able to bear a certain exercise load; 5) voluntarily participate in and sign a written consent form after being informed of the test process and purpose. Forty-eight subjects were randomly divided into control group (n = 24) and intervention group (n = 24), The ratio of male to female in the experimental group and the control group was the same, and there was no gender difference. This study was approved by the Biomedical Ethics Committee of Hubei University for nationalities (approval document No.: HBMZDX2022042). The basic conditions of the two groups are shown in **Table 1** below.

### 2.2 Test Method Design

The subjects sat in the lab 30 min earlier and wore a team heart rate meter. After sitting for 30 min, the subjects' microcirculatory blood flow, muscle oxygen saturation, heart rate, blood lactic acid, blood urea, creatine kinase (venous blood drawing), and other indexes were collected. The issues then ran 3,000 m (the subjects ran 3,000 m at the fastest speed on the treadmill). After the end of running, the above-mentioned indexes of the subjects were collected again. After the end of index collection, the issues in the intervention group inhaled portable oxygen for 30 min, the control group recovered naturally; after the end of 30 min, the above-mentioned indexes were collected again. Before the experiment, immediately after the experiment and 30 min after the experiment, the collection of microcirculation index and blood index were carried out at the same time, and there was no time difference. To ensure the stability of the operating condition of the issues, the American Omegawave competitive state comprehensive diagnosis system was used to diagnose the functional state of the subjects before exercise to ensure that the operating conditions of the two groups were the same before exercise. Environment is an important factor affecting vascular blood flow, so this study strictly controls the subjects' ambient temperature, humidity and indoor air flow in the test process to ensure the consistency of the test environment. The specific test process is shown in **Figure 1**.

**TABLE 1 |** Basic information of subjects ( $\bar{x} \pm s$ ,  $n = 48$ ).

Group	Age (years)	Height (m)	Weight (kg)	BMI (kg/M <sup>2</sup> )
Control group	21.07 $\pm$ 0.59	1.73 $\pm$ 0.05	66.80 $\pm$ 4.89	22.18 $\pm$ 1.02
Intervention group	21.41 $\pm$ 0.63	1.72 $\pm$ 0.06	65.40 $\pm$ 6.39	22.09 $\pm$ 1.27



## 2.3 Oxygen Suction Equipment

In this study, a portable oxygen-making machine made by Shandong Jing'an Medical Devices Co.Ltd., was used (model: P2MurW). The oxygen concentration was (90–96)%, the flow rate was 1L/min, and the oxygen discharge mode was pulse-type (equivalent to ordinary 5L/min oxygen machine effect). The oxygen supply of the instrument is stable and can meet the oxygen supply-demand when the respiratory rate is between 10 and 40 beats per minute.

## 2.4 Test Index and Instrument

### 2.4.1 Microcirculatory Blood flow、Muscle Oxygen Saturation

PF6010 dual-channel laser Doppler blood flow detector made in Sweden was used to measure the blood flow of the skin of the middle muscle of the quadriceps femoris of the right lower limb of the subjects. The test indexes included mean movement velocity of blood cells (AVBC), exercise blood cell concentration (CMBC), and microvascular blood perfusion (MBP), including MBP =

AVBC\*CMBC/100. The instrument for measuring muscle oxygen saturation (SmO<sub>2</sub>) is Moxy near-infrared wireless muscle oxygen testing system, and the test site is the same as microcirculatory blood flow.

### 2.4.2 Heart rate、Blood Lactic acid、Blood urea、Creatine Kinase

The heart rate (HR) acquisition instrument is the heart rate meter of the Polar team made in Finland. Blood lactic acid (BLA) was collected by German EKFLactateScout4 portable blood lactic acid analyzer. Blood urea (BU) and creatine kinase (CK) were tested in the laboratory. After venous blood was drawn, the blood was centrifuged, and then the serum was stored in a refrigerator of minus 80°, and then sent to the hospital for testing.

## 2.5 Statistical Method

Use SPSS25.0 to count the data. The normal distribution of the data was tested. The time and energy consumption of the

**TABLE 2 |** Comparison of time and energy consumption of two groups of subjects in running 3,000 m ( $\bar{x} \pm s$ ,  $n = 48$ ).

Group	Exercise time (s)	Energy consumption (kcal)
Control group	997.13 $\pm$ 81.55	186.87 $\pm$ 5.77
Intervention group	975.33 $\pm$ 84.06	192.25 $\pm$ 6.53

subjects running 3,000 m before and after running twice were tested by independent sample *t*-test, and the other indexes were compared by the analysis of repeated measurement variance.

## 3 RESULT

### 3.1 The Time and Energy Consumption of the Two Groups of Subjects Running 3,000 m

There was no significant difference in the time and calories consumed in running 3,000 m between the intervention group and the control group, as shown in Table 2.

### 3.2 Comparison of Microcirculation Perfusion in Different Periods Between Two Groups of Subjects

Table 3 showed that MBP and AVBC had interaction between groups and time before and after the test. Simple effect analysis showed that the group had different effects on MBP and AVBC, the MBP and AVBC immediately after exercise in the control group were significantly higher than those before exercise and 30 min after exercise, and the MBP and AVBC immediately after exercise in the intervention group were significantly higher than those before exercise and 30 min after exercise, and the MBP and AVBC before exercise were substantially lower than those at 30 min after exercise. Time had a different effect on MBP and AVBC. In the intervention group, MBP and AVBC before exercise were significantly lower than 30 min after exercise ( $p < 0.05$ ). Before and after the experiment, CMBC had no interaction between group and time ( $p > 0.05$ ).

### 3.3 Comparison of SmO<sub>2</sub>, HR, BLA, BU, and CK Between Two Groups of Subjects in Different Periods

Table 4 showed that SmO<sub>2</sub>, HR, BLA, BU, and CK had interaction between groups before and after the test. Simple effect analysis showed that the group had different effects on SmO<sub>2</sub>, HR, BLA, BU, and CK. In the control group, SmO<sub>2</sub> immediately after exercise was substantially lower than that before exercise and 30 min after exercise. At the same time, HR, BLA, BU, and CK immediately after exercise were substantially higher than those before exercise. HR and BLA immediately after exercise were substantially higher than those at 30 min after exercise. SmO<sub>2</sub> before exercise was substantially higher than at 30 min after exercise, while HR, BLA, BU, and CK before exercise were substantially lower than those at 30 min after exercise. In the intervention group, SmO<sub>2</sub> immediately after exercise was substantially lower than before exercise and 30 min after exercise. At the same time, HR, BLA, BU, and CK immediately after exercise were substantially higher than those before exercise and 30 min after exercise. BLA, BU, and CK before exercise were substantially lower than those at 30 min after exercise. Time had separate effects on SmO<sub>2</sub>, HR, BLA, BU, and CK. 30 min after exercise, SmO<sub>2</sub> in the control group was substantially lower than that in the intervention group. In contrast, HR, BLA, BU, and CK in the control group were substantially higher than those in the intervention group.

## 4 DISCUSS

After high intensity exercise, metabolic wastes such as lactic acid, carbon dioxide, and adenosine can be eliminated in time, and energy substances (oxygen, glucose, etc.) can be quickly transported to cells, thus promoting the rapid elimination exercise-induced fatigue and the quick recovery of body function. As the only place of material and energy metabolism, microcirculation is the key to completing this process. Insufficient microcirculatory blood perfusion after exercise will not only slow down the elimination of metabolic waste after exercise, but also cause the exchange of energy materials, resulting in the accumulation of metabolic waste in cells, thus unable to obtain sufficient energy materials and oxygen in time. In addition, the lack of microcirculation blood perfusion

**TABLE 3 |** Comparison of microcirculation perfusion in different periods between two groups of subjects ( $\bar{x} \pm s$ ,  $n = 48$ ).

Index	Group	Before exercise	Immediately after exercise	30 min after exercise	Interaction	
					F	P
MBP(PU)	Control group	6.05 $\pm$ 1.86*	37.25 $\pm$ 8.51	6.34 $\pm$ 2.14* <sup>&amp;</sup>	4.066	0.028
	Intervention group	6.19 $\pm$ 1.91* <sup>#</sup>	39.83 $\pm$ 5.92	10.13 $\pm$ 2.16*		
CMBC	Control group	62.08 $\pm$ 18.18	214.80 $\pm$ 66.12	65.57 $\pm$ 13.82	0.032	0.969
	Intervention group	64.57 $\pm$ 19.82	220.93 $\pm$ 62.06	68.07 $\pm$ 9.09		
AVBC	Control group	9.93 $\pm$ 2.60*	18.91 $\pm$ 7.32	10.23 $\pm$ 2.38* <sup>&amp;</sup>	3.959	0.031
	Intervention group	9.79 $\pm$ 2.03* <sup>#</sup>	19.81 $\pm$ 6.06	15.49 $\pm$ 4.24*		

Note: \* the difference is statistically significant compared with the corresponding index immediately after exercise in the group; # the difference between the corresponding index in the group and 30 min after exercise in the group is statistically significant; & the difference is statistically significant compared with the index at the same time in the group.

**TABLE 4 |** Comparison of SmO<sub>2</sub>, HR, BLA, BU and CK between two groups of subjects in different periods ( $\bar{x} \pm s$ ,  $n = 48$ ).

Index	Group	Before exercise	Immediately after exercise	30 min after exercise	Interaction	
					F	P
SmO <sub>2</sub>	Control group	67.91 $\pm$ 9.61 <sup>*#</sup>	48.24 $\pm$ 12.00	56.89 $\pm$ 9.32 <sup>*&amp;</sup>	3.987	0.025
	Intervention group	68.32 $\pm$ 9.44 <sup>*</sup>	46.04 $\pm$ 9.98	64.15 $\pm$ 8.87 <sup>*</sup>		
HR	Control group	76.87 $\pm$ 8.03 <sup>*#</sup>	174.20 $\pm$ 11.53	91.40 $\pm$ 7.23 <sup>*&amp;</sup>	3.693	0.038
	Intervention group	77.13 $\pm$ 7.37 <sup>*</sup>	175.47 $\pm$ 5.94	82.40 $\pm$ 7.47 <sup>*</sup>		
BLA	Control group	2.99 $\pm$ 0.50 <sup>*#</sup>	15.87 $\pm$ 2.54	6.08 $\pm$ 1.66 <sup>*&amp;</sup>	4.771	0.017
	Intervention group	2.88 $\pm$ 0.42 <sup>*#</sup>	16.14 $\pm$ 2.44	4.45 $\pm$ 1.07 <sup>*</sup>		
BU	Control group	4.23 $\pm$ 0.59 <sup>*#</sup>	6.20 $\pm$ 0.89	6.22 $\pm$ 0.85 <sup>&amp;</sup>	5.244	0.025
	Intervention group	4.28 $\pm$ 0.51 <sup>*#</sup>	6.46 $\pm$ 1.06	5.48 $\pm$ 0.87 <sup>*</sup>		
CK	Control group	122.27 $\pm$ 28.02 <sup>*#</sup>	324.60 $\pm$ 49.22	342.47 $\pm$ 64.38 <sup>&amp;</sup>	11.147	0.000
	Intervention group	125.67 $\pm$ 27.07 <sup>*#</sup>	325.73 $\pm$ 49.06	270.27 $\pm$ 51.29 <sup>*</sup>		

Note: \* the difference is statistically significant compared with the corresponding index immediately after exercise in the group; # the difference between the corresponding index in the group and 30 min after exercise in the group is statistically significant; & the difference is statistically significant compared with the index at the same time in the group.

will also lead to the disturbance of material and energy metabolism during exercise, which will lead to insufficient oxygen supply during exercise, induce exercise-induced fatigue and reduce exercise performance. Therefore, the amount of microcirculatory blood perfusion is closely related to the recovery of physical function and exercise-induced fatigue of athletes after exercise. Increasing the amount of microcirculatory blood perfusion is helpful to improve the exchange capacity of oxygen and nutrients between blood and tissues, increase the efficiency of systemic blood circulation, and promote the rapid recovery of the body after exercise and the elimination of fatigue.

In this study, the microcirculatory blood flow of the two groups increased significantly after the end of 3000-m running. The increase of microcirculation blood flow during exercise is mainly related to the release of endogenous NO. In the process of exercise, the blood flow in the blood vessel increases, which increases the fluid shear stress of the blood to the blood vessel wall, and the shear stress can promote the vascular endothelial cells to produce endogenous NO, promote vasodilation and increase blood flow (Wang and Cai, 2012; Zhu and Gao, 2020). In addition, the production of hydrogen ions, adenosine, and other metabolites during exercise can also stimulate vascular endothelial cells and improve vasodilation ability and blood perfusion level (Binggeli et al., 2003; Cracowski et al., 2007; Lorenzo and Minson, 2007). The increase of microcirculation blood flow during exercise can provide energy and oxygen for skeletal muscle and other organs to meet the metabolic needs of the body; at the same time, the metabolic waste produced during exercise can be removed from the body in time. After the exercise, due to the excessive oxygen consumption, the subjects are still in high intensity operation, and metabolic wastes such as lactic acid continue to be produced. Therefore, keeping the microcirculation blood flow at a certain level after exercise is helpful to the elimination of metabolic waste and the transport of energy and oxygen, accelerating the elimination of fatigue and promoting the recovery of cell function. In this study, the microcirculation blood flow of the subjects in the intervention group was significantly higher than that in the control group after 30 min of portable oxygen inhalation, indicating that inhaling

high concentration oxygen after exercise could slow down the decline of microcirculation blood flow. The effect of oxygen inhalation on microcirculation blood flow may be related to the release of NO. Studies have shown that oxygen inhalation during exercise can significantly increase the levels of eNOS and NO in blood after exercise, and increase the level of blood perfusion after exercise (Cui et al., 2005). In addition, the effect of oxygen inhalation on microcirculation blood flow after exercise may be related to the deformability of red blood cells. After strenuous exercise, many intracellular Ca<sup>2+</sup> gathered and activated the K<sup>+</sup> ion channel on the erythrocyte membrane, which led to the deformation of red blood cells. Inhaling a high oxygen concentration after exercise could promote the morphological changes of red blood cells to be double flattened, increase the contact area with oxygen and improve the movement speed of blood cells (Xiao et al., 2002).

Muscle oxygen saturation mainly refers to the dynamic balance of oxygen supply and oxygen consumption in arteries, veins, and capillaries of skeletal muscle and the overall effect of myoglobin content, equal to (oxygenated hemoglobin + oxygenated myoglobin)/(hemoglobin + myoglobin). After oxygen enters the blood, it combines with hemoglobin to form oxygenated hemoglobin, which is then transported to cells (especially muscle cells) with the help of capillaries to be used by cells. At the same time, myoglobin in muscle will also combine part of oxygen to be stored for cell activity to form oxygenated myoglobin. The combination of oxygen and myoglobin directly affects the ability of myoglobin to store oxygen. Under normal circumstances, the blood oxygen saturation can reach more than 98%. Still, the muscle oxygen saturation (the comprehensive reaction of hemoglobin and myoglobin to bind oxygen) is lower, indicating that under normal circumstances, the ability of myoglobin oxygenation is much less than that of hemoglobin oxygenation. Myoglobin oxygenation ability mainly depends on the ability of capillaries to transport oxygen to muscle cells. In this oxygen, partial pressure difference is the power to promote oxygen from veins to muscle cells. The greater the oxygen partial pressure of microvessels, the greater the pressure difference between microvessels and muscle cells, the greater the rate of oxygen from capillaries into muscle cells, the higher the

content of muscle, and the stronger the oxygenation ability of myoglobin. The higher the partial pressure of oxygen, the greater the pressure difference between the oxygen and muscle cells, which accelerates the rate of oxygen from capillaries to muscle cells, so the oxygen content in muscle increases and the oxygenation ability of myoglobin increases. Therefore, compared with blood oxygen saturation, muscle oxygen saturation can better reflect the functional recovery of skeletal muscle cells after exercise.

In the exercise process, due to the massive consumption of oxygen, the muscle oxygen saturation decreased significantly. After exercise, with the establishment of oxygen supply-aerobic balance, muscle oxygen saturation gradually recovers, so muscle oxygen saturation can objectively reflect the change of oxygen supply capacity of the body after exercise. In this study, the muscle oxygen saturation of the subjects in the intervention group was significantly higher than that in the control group after portable oxygen inhalation for 30 min, indicating that oxygen inhalation after exercise can improve the oxygen supply capacity of the body and promote the recovery of muscle oxygen saturation, which is similar to the research results of relevant scholars. Studies have shown that oxygen supplementation during exercise can significantly prolong the endurance time and the lowest muscle oxygen saturation of mice (Marillier et al., 2021). The intervention effect of oxygen inhalation on muscle oxygen saturation is closely related to the change of microcirculatory blood flow. As mentioned earlier, the oxygen partial pressure difference between capillaries and muscle cells is the driving force to increase the oxygen diffusion rate and distance. The microcirculation blood flow is the carrier to carry and transport oxygen: the greater the microcirculation blood flow, the greater the capillary oxygen partial pressure. Therefore, the oxygen partial pressure difference between capillaries and muscle cells will also increase.

Heart rate is the most direct and straightforward index to reflect the exercise intensity and exercise recovery of the human body. It is found that inhaling high oxygen for 30 min after exercise can accelerate the recovery speed of heart rate and lactic acid (Tao, 2011). Other studies have shown that although oxygen inhalation during exercise has no significant effect on the oxygen consumption of rowers, oxygen inhalation after exercise can significantly reduce the level of blood lactic acid and heart rate of rowers before the next exercise (Pelvan and Çotuk, 2016). This study found that the heart rate of the subjects increased significantly after 3000-m running, but inhaling hyperoxia for 30 min immediately after exercise could accelerate the recovery of heart rate, which was related to the change of sympathetic nerve function. Oxygen inhalation after exercise can enhance the function of the vagus nerve and heart rate variability, quickly recover heart rate, and reduce cardiac oxygen consumption (Xie, 2016). The content of blood lactic acid in the human body is low in a quiet state, and a large amount of lactic acid will be produced after high intensity exercise. The lack of timely decomposition or oxidation of lactic acid in the body will lead to the decline of cell function (inhibiting the activity of enzymes) and induce exercise-induced fatigue. It is found that the elimination of lactic acid can be accelerated by inhaling atmospheric pressure and high

concentration of oxygen after a 400-m race (Xie et al., 2016). In addition, studies have found that inhaling hyperoxia after exercise can increase oxygen content in the blood, alleviate the blood oxygen supply after strenuous exercise, accelerate the speed of intracellular H<sup>+</sup> movement, and promote the elimination of lactic acid (Tao, 2011). In addition, studies have shown that oxygen inhalation during exercise can improve endurance exercise performance and reduce the concentration of lactic acid after exercise (Silva et al., 2021). In this study, the concentration of blood lactic acid increased significantly after 3000-m running. Still, the clearance of lactic acid was accelerated by inhaling high oxygen for 30 min immediately in the intervention group. The main pathways of lactic acid after exercise are direct oxidative decomposition, gluconeogenesis for saccharification storage, synthesis of fatty acids and other substances, and directly removed from the body with urine and body fluids, etc. Still, all ways must be completed with the help of blood circulation. Therefore, increasing the blood flow of microcirculation after exercise can accelerate the oxidative decomposition of lactic acid and promote the recovery of cell function.

Blood urea is the product of protein decomposition and is an important index to evaluate exercise-induced fatigue. Blood urea level is mainly affected by renal function, protein intake, and catabolism. When athletes are in a state of exhaustion, the energy supply of fat oxidation decreases, the proportion of protein-energy supply increases, and the decomposition rate increases, increasing blood urea content. Creatine kinase is a standard index to evaluate exercise intensity. Under the stimulation of hypoxia, mechanical stress, and traction, the permeability of skeletal muscle cell membrane will change, resulting in the flow of creatine kinase from cells to the blood, resulting in an increase in the content of creatine kinase in the blood. The results showed that after 30-s maximum power cycling and exercise-induced fatigue in rats, the renal function was restored, and the blood urea level decreased significantly (Teng et al., 2013; Xu, 2014; Guoyin, 2019). In addition, other studies have shown that a high concentration of oxygen inhalation after high intensity exercise can reduce the contents of creatine kinase, creatine kinase isoenzyme, and lactate dehydrogenase, and prevent the occurrence of sports injury (Xie et al., 2016; Yang et al., 2016). In this study, after 30 min of high concentration oxygen supplementation, the levels of blood urea and creatine kinase in the intervention group decreased significantly, indicating that high concentration oxygen supplementation after high intensity exercise can promote the recovery of cellular functions such as liver and skeletal muscle, and reduce the occurrence of exercise-induced fatigue. In the process of high intensity exercise, renal hypoxia leads to the decrease of glomerular filtration ability, resulting in the accumulation of blood urea in the body. At the same time, oxygen inhalation intervention improves the occurrence of renal hypoxia by increasing the blood flow of renal tissue, so that the renal function can be recovered. The urea accumulated in the blood is removed. In addition, a high concentration of oxygen supplementation after exercise can



also promote the aerobic metabolism of fat and sugars, reduce the decomposition of protein, and reduce the content of urea in the blood. Therefore, the intervention of high concentration oxygen supplementation after exercise-induced fatigue can promote the recovery of blood urea levels. In addition, oxygen supplementation after exercise can promote the healing of skeletal muscle cell membrane permeability caused by hypoxia and make skeletal muscle cells maintain normal function. However, some studies have shown that the change of creatine kinase level after high intensity exercise mainly reflects delayed muscle soreness. High concentration oxygen supplementation can not improve muscle soreness and creatine kinase level (Shelina et al., 2003; Bennett, 2014; Navid et al., 2020). The authors believe that the effect of a high concentration of oxygen supplementation on creatine kinase after high intensity exercise is related to the changes in muscle cell membrane permeability. When the muscle fiber structure is physiologically damaged by eccentric exercise, the intervention effect of high concentration oxygen supplementation may not be good. Still, when hypoxia leads to the increase of muscle membrane permeability, high concentration oxygen supplementation can promote the recovery of cell function by improving the level of skeletal muscle blood perfusion and oxygen supply capacity. In this study, the subjects took the 3000-m race as the exercise content (without centrifugal exercise). Hypoxia caused changes in cell membrane permeability after exercise, while a high concentration of oxygen supplementation promoted the recovery of cell membrane permeability and decreased the level of creatine kinase.

## 5 CONCLUSION

Normobaric oxygen inhalation for 30 min after exercise can delay the decrease of microcirculatory blood flow, increase muscle oxygen saturation, and promote the recovery of heart rate, blood lactic acid, blood urea, and creatine kinase. Therefore, normobaric oxygen inhalation for 30 min after exercise can be used as an effective means to promote the elimination of exercise-induced fatigue after endurance running.

## REFERENCES

- Bennett, M. H. (2014). Hyperbaric Medicine and the Placebo Effect. *Diving Hyperb. Med.* 44 (4), 235–240.
- Binggeli, C., Spieker, L. E., Corti, R., Sudano, I., Stojanovic, V., Hayoz, D., et al. (2003). Statins Enhance Postischemic Hyperemia in the Skin Circulation of Hypercholesterolemic Patients. *J. Am. Coll. Cardiol.* 42 (1), 71–77. doi:10.1016/s0735-1097(03)00505-9
- Cracowski, J. L., Lorenzo, S., and Minson, C. T. (2007). Effects of Local Anaesthesia on Subdermal Needle Insertion Pain and Subsequent Tests of Microvascular Function in Human. *Eur. J. Pharmacol.* 559 (2–3), 150–154. doi:10.1016/j.ejphar.2006.11.069
- Cui, J., Zhang, X., and Zhan, X. (2005). etc. Effects of Oxygen Inhalation on Nitric Oxide and Nitric Oxide Synthase during Exhaustive Exercise and Evaluation of Liquid Oxygen Application at High Altitude. *Clin. rehabilitation China* 1(16), 192–193.
- Guoyin, L. I. (2019). Study on the Effect of Hyperbaric Oxygen on Biochemical Blood Indexes of Rats with Exercise-Induced Fatigue. *J. Jiangxi Normal Univ. Sci. Technol.* 1(06), 101–105.
- Lorenzo, S., and Minson, C. T. (2007). Human Cutaneous Reactive Hyperaemia: Role of BKCa channels and Sensory Nerves. *J. Physiol.* 585, 295–303. doi:10.1113/jphysiol.2007.143867
- Marillier, M., Bernard, A. C., Moran-Mendoza, S. O., O'Donnell, D. E., and Neder, J. A. (2021). Oxygen Supplementation during Exercise Improves Leg Muscle Fatigue in Chronic Fibrotic Interstitial Lung Disease. *Thorax* 76, 672–680. doi:10.1136/thoraxjnl-2020-215135

## 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

The research subjects selected voluntarily participated in and signed a written consent form after being informed of the test process and purpose.

## AUTHOR CONTRIBUTIONS

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

## ACKNOWLEDGMENTS

We want to thank all those who participated in this study, especially the co-first author LM from Jiangsu University of Science and Technology, HZ of the Sports Physiology Laboratory of Hubei University for nationalities, Professor Xiaoli Liu of Hubei University for nationalities, and all the subjects for their support. Thank you for the high-level training project of Hubei University for nationalities, “Construction and empirical Research on exercise intervention Model of Common chronic Diseases of the elderly in Community under the background of healthy China 2030” (Project No.: PY21002), and “Research on the effectiveness and Mechanism of oxygen inhalation intervention on the Elimination of exercise-induced fatigue in competitive ethnic Sports” (Project No.: D20201901), which is the critical project of the Science and Technology Plan of Hubei Provincial Department of Education in 2020, and “Application of Heart Rate Variability in the Monitoring of Exercise Fatigue Before Competitive National Sports Athletes Prepare for Major Competitions” (Project No.: B2021152), a guiding project of the Scientific Research Program of Hubei Provincial Department of Education in 2021.

- Navid, M., Michinari, H., Lindsey, R., Levine, B. D., and Guiliod, R. (2020). Hyperbaric Oxygen Therapy in Sports Musculoskeletal Injuries. *Med. Sci. Sports Exerc* 52 (6), 1420–1426. doi:10.1249/MSS.0000000000002257
- Pelvan, S. O., and Çotuk, B. (2016). Investigation of the Effect of Inhalation of Hyperoxic Gas between High Intensity Rowing Exercises on the Muscle Tissue Oxygen Consumption Level Using Functional Near Infrared Spectroscopy. *Clin. Exp. Health Sci.* 1. 1.
- Shelina, B., Rhodes Edward, C., Taunton Jack, E., and Lepawsky, M. (2003). Effects of Intermittent Exposure to Hyperbaric Oxygen for the Treatment of an Acute Soft Tissue Injury. *Clin. J. Sport Med.* 13 (3), 138–147. doi:10.1097/00042752-200305000-00003
- Silva, T. C., Aidar, F. J., Zanona, A. d. F., Matos, D. G., Pereira, D. D., Rezende, P. E. N., et al. (2021). The Acute Effect of Hyperoxia on Onset of Blood Lactate Accumulation (OBLA) and Performance in Female Runners during the Maximal Treadmill Test. *Ijerp* 18 (9), 4546. doi:10.3390/ijerp18094546
- Tao, X. (2011). An Empirical Study on the Effect of Hyperoxia on the Recovery of Male Kayakers after Wingate Test. *J. Beijing Univ. Phys. Educ.* 34 (08), 63–65. doi:10.19582/j.cnki.11-3785/g8.2011.08.017
- Teng, J., Yuan, C., and Guo, Y. (2013). etc. Protective Effect of Hyperbaric Oxygen on Renal Injury in Fatigue Rats. *Chin. J. Sports Med.* 32 (06), 525–528. doi:10.16038/j.1000-6710.2013.06.009
- Wang, Y., and Cai, J. (2012). Effects of Shear Stress and Tensile Stress on Vascular Endothelial Cells. *Chin. J. Med. Front.* 4 (08), 38–41.
- Xiao, G., Jia, H., and Qiu, Z. (2002). etc. Effect of Oxygen Inhalation on Erythrocyte Morphology and Hemorheological Properties after High Intensity Exercise. *Chin. J. Sports Med.*, 1. 37–40. doi:10.16038/j.1000-6710.2002.01.010
- Xie, W., Xie, X., Yang, Y., and Weibin, M. O. (2016). Effects of Oxygen Inhalation on BLA, SOD, CK, and LDH of 400 Meters Athletes before and after Competition. *J. Shandong Sport Univ.* 32 (02), 79–83. doi:10.14104/j.cnki.1006-2076.2016.02.014
- Xie, Y. (2016). Effect of Oxygen Inhalation Intervention on Sleep Quality of Female Boxers [D]. *Master's degree thesis Cap. Inst. Phys. Educ.* 1. 1.
- Xu, M. (2014). *Study on the Recovery of a 30-second Maximum Power of College Students in Short-Distance Racing in Soft Hyperbaric Oxygen Chamber* [D]. Master's degree thesis Cap. Inst. Phys. Educ.
- Yang, Y., Weibin, M. O., and Xie, W. (2016). Effects of Oxygen Inhalation on Biochemical Indicators and Immunoglobulin of 400m Runners. *Mod. Prev. Med.* 43 (15), 2724–2728.
- Zhang, H. N., Gao, B. H., and Zhu, H. (2017). The Relationship between Reserve Capacity of Microcirculatory Blood Perfusion and Related Biochemical Indices of Male Rowers in Six Weeks' Pre-competition Training. *Chin. J. Appl. Physiol.* 33 (2), 112–116. doi:10.12047/j.cjap.5429.2017.029
- Zhang, Yu. (2019). *Experimental Study on the Effect of Oxygen Inhalation on Cardiovascular Stress in Middle and Long Distance Running*. Master's degree thesis Jiangxi Normal Univ. Sci. Technol. doi:10.27751/d.cnki.gjxkj.2019.000045
- Zhu, H., and Binghong, G. A. O. (2016). The Study of Relationship Between Maximum Reserve Capacity of Microcirculatory Blood Perfusion and Functional Status of M. An Rower During Six Weeks of Altitude Training. *J. Henan Normal Univ. Sci. Ed.* 44 (02), 176–182. doi:10.16366/j.cnki.1000-2367.2016.02.031
- Zhu, H., and Gao, B. (2019). Application of Microvascular Reactivity in the Training of Endurance Athletes. *Chin. J. Sports Med.* 38 (10), 907–914. doi:10.16038/j.1000-6710.2019.10.01
- Zhu, H., and Gao, B. (2020). Research Progress in Effects and Mechanisms of Aerobic Exercise on Human Microvascular Reactivity. *Chin. Bull. Life Sci.* 32 (08), 855–863. doi:10.13376/j.cbls/2020107

**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.

Copyright © 2022 Peng, Meng, Zhu, Wan and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Research on Adoption Behavior and Influencing Factors of Intelligent Pension Services for Elderly in Shanghai

Juan Luo<sup>1,2</sup> and Lingqi Meng<sup>2\*</sup>

<sup>1</sup>School of Social Development and Public Policy, Fudan University, Shanghai, China, <sup>2</sup>Shanghai University of Engineering Science, Shanghai, China

## OPEN ACCESS

### Edited by:

Deepak Kumar Jain,  
Chongqing University of Posts and  
Telecommunications, China

### Reviewed by:

Gabriel Gomes,  
State University of Campinas, Brazil  
Xuetao Li,  
Hubei University of Automotive  
Technology, China  
Shuai Li,  
Anyang Normal University, China  
Rose Inawaty Ibrahim,  
Islamic Science University of Malaysia,  
Malaysia

### \*Correspondence:

Lingqi Meng  
mlq\_kiki@163.com

### Specialty section:

This article was submitted to  
Human and Medical Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 28 March 2022

**Accepted:** 03 May 2022

**Published:** 15 June 2022

### Citation:

Luo J and Meng L (2022) Research on  
Adoption Behavior and Influencing  
Factors of Intelligent Pension Services  
for Elderly in Shanghai.  
Front. Genet. 13:905887.  
doi: 10.3389/fgene.2022.905887

With the rapid development of artificial intelligence and Internet-of-Things technology, the traditional pension service mode has changed, and intelligent pension services have become a new direction of development. Descriptive statistical analysis is conducted on the supply status and demand of intelligent pension services. It is believed that the current intelligent pension services are still in the initial stage of development, and the contradiction between supply and demand is prominent. The demand for intelligent pension of the elderly is high, but the user acceptance and satisfaction are not high. On this basis, variables were selected from individual characteristics, family situation, economic status, education level, living conditions, and other indicators for multivariate unconditional logistic regression analysis. It was found that the adoption behavior of intelligent pension service users was most significantly affected by age, number of children, living conditions, service cost, service docking channel, and equipment operation difficulty. Based on the conclusion, this article puts forward some suggestions such as taking the government as the center to realize the multi-governance of intelligent pension services, improving the supply of intelligent pension service-related facilities guided by demand, optimizing the service mode based on the platform to realize dynamic combination, and taking talents as the core to promote the high-quality development of intelligent pension services.

**Keywords:** artificial intelligence, intelligent pension, pension model, influencing factors, home care

## 1 RESEARCH REVIEW AND QUESTIONS

Population aging has become an inevitable trend of social development. The National Office for Aging released a research report on the prediction of population aging trend in China, and it is expected that China will enter the accelerated aging stage from 2021 and the severe aging stage from 2051, and the aging trend will gradually increase. The modernization of society has spawned profound social changes. With the vigorous development of Internet-of-Things technology and big data, artificial intelligence has gradually been applied to the field of pension services and has become an important driving force for the transformation of traditional pension models. The issuance of Opinions on Promoting the Development of Pension Service by the General Office of the State Council in 2019 and the deployment requirements of the Outline of Healthy China 2030 in China have clarified the important role of improving the quality of pension service development in the whole process of social development. In 2017, the Ministry of Industry and Information Technology,



the Ministry of Civil Affairs, and the Health and Family Planning Commission jointly issued “Smart Health Care Industry Development Action Plan (2017–2020),” raising the development of the smart and healthy pension industry to the national strategy. As an international metropolis, Shanghai has relatively high financial support capacity. Perfect network infrastructure construction has become a favorable condition for carrying out community intelligent pension services. The development of preliminary pilot work has also accumulated experience for the later operation of intelligent pension services. Shanghai has certain typicality and representativeness in the development of home intelligent pension services.

Combined with the current domestic research, many scholars have defined the concept and explained the connotation of intelligent pension. Most of them believe that intelligent pension is a new pension service model that relies on Internet technology and uses terminal equipment and data platform to provide effective pension services (Zhu and Tang, 2020). Intelligent pension services focus on the integration of pension and medical resources, and top-level design has strong comprehensive requirements (Wu, 2019). Compared with the traditional pension service mode, intelligent pension has the characteristics of high efficiency, which can maximize the action ability of the elderly, expand the rights of the elderly, and effectively reduce the pressure of social support (Jain et al., 2021). Relevant research has been conducted on the influencing factors of the demand for intelligent pension services. From the perspective of the elderly’s own life security needs, Bai Mei et al. (2015) found through a survey of Wuhan that the elderly at different ages will have different willingness to adopt smart pension services due to the influence of factors such as living conditions, economic conditions, and physical conditions. Mao and Li (2015) proposed eight influencing factors based on the UTAUT model, including performance expectation, convenience condition, perceived security, and perceived trust. The study found that convenience condition was the key factor affecting the demand for intelligent pension services. Gao et al. (2019), through the empirical research on the development of smart pension in Bincheng District, put forward the influencing factors of smart pension service demand such as cultural education, children’s situation, and their own quality. From the perspective of service providers, Zhang Lei (2017) believes that the future market development prospect of smart pension is optimistic, but there are many problems in technical means and standard formulation (Zhang and Han, 2017). These factors have become an important factor restricting the promotion of the smart pension mode.

At the same time, some scholars believe that the behavior choice of intelligent pension service users is also affected by external risk factors. The conflict of interest between the supply and demand of smart pension services is the fundamental reason for the difficulties in its popularization and promotion (Zhang and Li, 2019). At present, the overall problem of smart pension services in China is the shortage of total supply and the structural contradiction between supply and demand, which forms the urgency of the supply-side structural reform of smart pension services (Liao, 2019). The contradiction between supply and

demand structures of intelligent pension services is manifested in the mismatch between the supply and demand of products and services and the imbalance between regional supply and demand. In terms of services and products, China’s intelligent pension services are less involved and do not match the diversified and personalized needs of pension services. Therefore, the research on the demand side of the intelligent pension service is the basis of supply side structural reform and has important research value and significance.

On the whole, the current research on the demand and influencing factors of the intelligent pension service has made some achievements, but the empirical research on the demand side is still less, especially the lack of relevant research on the influence of subsequent elderly users’ adoption behavior. On the basis of relatively fully grasping the influencing factors of intelligent pension service demand, it is also necessary to conduct more in-depth research on the influencing factors that determine the final adoption behavior of user groups so as to better transform demand into practical application. The implementation of smart pension services in Shanghai is relatively considerable, and it is representative and exemplary in China. On the basis of household registration, salary level, education level, and other existing characteristics, this article incorporated children, living conditions, and other factors into the analysis variables and explored the relevant situation of the users’ adoption behavior of smart pension services in Shanghai in many aspects, aiming to find an effective path for the development of smart pension services.

## 2 DATA SOURCES OF INTELLIGENT PENSION

### 2.1 Defining the Scope of Investigation

The survey covers 15 districts in Shanghai, and Chongming District is not covered due to its remote geographical location. The selected streets in each district are mainly “smart healthcare demonstration street” and “smart pension community” in Shanghai.

### 2.2 Subjects of the Survey

The subjects of this survey were the elderly population over 60 years old in Shanghai.

### 2.3 Survey Methods and Tools

In 2019, the total number of elderly people over 60 years of age in the Shanghai Bureau of Statistics was 518.12 million. According to the proportion of 0.01%, a total of 518 elderly people were to be investigated. In order to ensure the quality of the questionnaire and eliminate invalid questionnaires, the number of respondents is now increased to 550. In a self-designed questionnaire, the questionnaire content included four main parameters: demographic data, intelligent pension service supply status, intelligent pension service supply satisfaction, and intelligent pension service demand status. Demographic data include gender, age, educational level, income, number of children, and living conditions; the supply status of intelligent pension services

**TABLE 1 |** Descriptive statistics of demographic characteristics.

Variable	Classification	Number of people (n)	Percentage (%)
Sexuality	Males	232	43.7
	Females	299	56.3
Age	60–69 years	265	50.3
	70–79 years	216	41.0
	80–89 years	44	8.3
	Above 90 years old	2	0.4
Educational level	Uneducated	17	3.2
	Primary school	129	24.3
	Junior high school	163	30.7
	Technical secondary school	95	17.9
	High school	66	12.4
	Junior college	44	8.3
	Undergraduate	17	3.2
Number of children	Five or more	14	2.6
	Four	26	4.9
	Three	38	7.2
	Two	151	28.4
	One	298	56.1
	Nil	4	0.8
Income	Less than 2,000 yuan	19	3.6
	2,001–3,000 yuan	43	8.1
	3,001–4,000 yuan	109	20.5
	4,001–5,000 yuan	130	24.5
	More than 5,000 yuan	230	43.4
Revenue sources	Income from work	20	2.90
	Retirement pay	472	67.30
	Children give	174	24.80
	Government subsidies	34	4.90
	Other	1	0.10
Living condition	Individual living alone	27	5.1
	With lover	257	48.4
	With children	150	28.2
	With lover and children	95	17.9
	With the elderly	2	0.4

includes the cognition of users of intelligent pension services, the equipment of intelligent pension services, the use of intelligent pension services, and the sources of information acquisition of intelligent pension services. The satisfaction degree of intelligent pension service includes the satisfaction degree and influencing factors of users to the existing intelligent pension service; the demand for intelligent pension services includes user preference for the types of intelligent pension services, the degree of demand for intelligent pension services, and influencing factors. A total of 550 questionnaires were distributed, and 531 valid questionnaires were recovered, with an effective rate of 96.5%.

## 2.4 Research on Demand and Influencing Factors of the Intelligent Pension Service

The collected questionnaires were collated and coded, and the relevant data were analyzed by SPSS statistical software. Demographic data and the supply and demand status of intelligent pension services were analyzed by descriptive statistics, and the influencing factors of user adoption behavior

of intelligent pension services were studied by the multivariate unconditional logistic regression analysis.

## 3 SUPPLY, DEMAND, AND INFLUENCING FACTORS OF THE INTELLIGENT PENSION SERVICE

### 3.1 Descriptive Statistical Analysis of Demographic Data

In this survey, the proportion of females in intelligent pension service users is slightly higher than that of males, and the age distribution is mainly young elderly (60–69 years old), accounting for about 50.3%. The educational level of the survey users is generally concentrated in the primary and junior high school stages. Combined with the analysis of the background of the times, most of the elderly groups were born before the 1950s. Affected by social and economic factors, the cultural level is generally low. The majority of children are less than three, and the pressure to support the elderly is high. About 43.4% of the intelligent pension service users' income is more than

**TABLE 2 |** Degree of understanding of intelligent elderly care services.

Understanding of existing smart elderly care services	Frequency	Percentage (%)	Effective percentage (%)	Cumulative percentage (%)
Know very well	2	0.4	0.4	0.4
Better understanding	31	5.8	5.8	6.2
Basically understood	124	23.4	23.4	29.6
Not well understood	315	59.3	59.3	88.9
Never heard	58	10.9	10.9	99.8

**TABLE 3 |** Satisfaction with smart elderly care services.

Satisfaction with existing smart elderly care services	Frequency	Percentage (%)	Effective percentage (%)	Cumulative percentage (%)
Very satisfied	54	10.2	10.2	10.2
Quite satisfied	6	1.1	1.1	11.3
General	86	16.2	16.2	27.5
Less satisfied	256	48.2	48.2	75.7
Very dissatisfied	45	8.5	8.5	84.2

5,000 yuan; 67.3% of the intelligent pension service users' funds mainly rely on pension, and 24.8% are given by their children. The overall income of the elderly group is relatively considerable, but the access to security funds is relatively single, and there are still some elderly people in poverty. In the choice of living style, 48.4% of the elderly live with their lovers, 28.2% of the elderly live with their children, and the proportion of the elderly living alone is about 5.1%. **Table 1** shows some basic descriptive data that have been analyzed.

### 3.2 Status Quo of Smart Elderly Service Supply

The survey results showed that the elderly population has relatively low understanding of smart elderly care services. Among them, 59.3% of the elderly do not know much about smart elderly care services, 10.9% have never heard of smart elderly care services, and 23.4% of the elderly have a basic understanding. The elderly who are relatively and very knowledgeable about intelligent elderly care services account for only 5.8% and 0.4%, respectively. In order to more intuitively show the understanding of the elderly on intelligent pension services, we presented this in **Table 2**.

At present, the overall satisfaction of smart elderly care service users is not high. The proportion of relatively dissatisfied and very dissatisfied users is as high as 56.7%, the generally satisfied users account for 16.2%, and the very satisfied and relatively satisfied users account for only 11.3%. It can be seen that the development of intelligent elderly care services is still in the preliminary stage of development. Users lack relevant cognition and have a low understanding of community intelligent elderly care, which affects the further promotion and application of community intelligent elderly care service, and there is still much room for improvement and development. The satisfaction of the elderly group to intelligent elderly care service is illustrated in **Table 3**.

In terms of the factors affecting the satisfaction of smart elderly care services, users generally believe that the types of services and service personnel are the main factors affecting satisfaction. Among them, 28.6% of the elderly believe that the current smart elderly services provide fewer types of services, and 22.4% of the elderly believe that there are fewer professional service personnel, and second, service charges and service timeliness are also important aspects that affect user satisfaction. In order to make the data results more intuitive, we showed the factors influencing the satisfaction of intelligent pension services in **Table 4**.

### 3.3 Description and Analysis of User Needs and Adoption Behavior of Intelligent Elderly Care Services

#### 3.3.1 Analysis of Needs and Preferences of Smart Elderly Care Users

Through statistical analysis, it is found that the needs of the elderly for intelligent elderly care services are mostly concentrated in the two aspects of life care and medical care, followed by spiritual comfort, and most of them have moderate needs in cultural education and fitness. The demand for legal services is relatively low. Combining with Maslow's hierarchy of needs theory, the current relative basics of the life needs of the elderly are mostly related to the physical and social needs of the fetish and spiritual levels, and the degree of self-realization needs is relatively weak. **Table 5** shows the demand preference analysis of the elderly for intelligent elderly care services.

#### 3.3.2 Analysis of the Adoption Behavior of Smart Elderly Service Users

Contrary to the distribution of demand level surveys, the elderly population generally has a higher level of demand for smart

**TABLE 4 |** Factors affecting satisfaction with smart elderly care services (unit: person).

Factors influencing satisfaction of smart elderly care services	Number of response cases	Percentage (%)	Percentage of cases (%)
Unreasonable fees	115	16.50	26.20
Poor service timeliness	104	14.90	23.70
Cannot provide customized services	99	14.20	22.60
Fewer professional service staff	156	22.40	35.50
Fewer professional service staff	199	28.60	45.30
Other	23	3.30	5.20

**TABLE 5 |** Analysis of demand preference for smart elderly care services (unit: person).

Demand level	Life care	Medical insurance	Spiritual comfort	Cultural education	Physical fitness	Legal service
Not needed at all	44 (8.3%)	8 (1.5%)	64 (12.1%)	75 (14.1%)	79 (14.9%)	137 (25.8%)
Not needed	40 (7.5%)	21 (4.0%)	83 (15.6%)	151 (28.4%)	137 (25.8%)	155 (29.2%)
General	90 (16.9%)	74 (13.9%)	120 (22.6%)	130 (24.5%)	146 (27.5%)	119 (22.4%)
Need	126 (23.7%)	128 (24.1%)	132 (24.9%)	106 (20.0%)	93 (17.5%)	71 (13.4%)
Very necessary	231 (43.5%)	300 (56.5%)	132 (24.9%)	69 (13.0%)	75 (14.1%)	49 (9.2%)

**TABLE 6 |** Provision of smart elderly service facilities.

Provision of intelligent elderly service facilities	Frequency	Percentage (%)	Effective percentage (%)	Cumulative percentage (%)
Much	4	0.8	0.8	0.8
More	27	5.1	5.1	5.8
General	125	23.5	23.5	29.4
A bit less	222	41.8	41.8	71.2
Basically no	152	28.6	28.6	99.8

**TABLE 7 |** Equipped with smart elderly service equipment.

Already equipped with intelligent elderly care equipment	Number of response cases	Percentage (%)	Percentage of cases (%)
Intelligent walkway	45	5.4	9.9
Health self-examination cabin	35	4.2	7.7
Self-help physical examination apparatus	60	7.2	13.2
Physical fitness monitoring station	74	8.8	16.3
Mental health self-help instrument	24	2.9	5.3
Intelligent health monitor	74	8.8	16.3
Smartphone watches	136	16.2	30
Intelligence appliance	230	27.4	50.7
Guided robot	41	4.9	9
Intelligent fitness equipment	38	4.5	8.4
Other	82	9.8	18.1

elderly care services but seldom adopts smart elderly care services. Among them, 41.8% of elderly users are equipped with less smart elderly care service equipment, and 23.5% of the elderly population are equipped with average smart elderly care equipment. There are still 28.6% of the elderly who are basically not equipped with any smart elderly care equipment. The users with much or smarter elderly care equipment are around 0.8% and 5.1%, respectively. The frequency analysis of

providing intelligent elderly care service facilities is presented in **Table 6**.

The types of smart elderly care equipment currently equipped by elderly users are relatively basic; 27.4% are smart home appliances, 16.2% are smart phone watches, and the proportion of other smart elderly care equipment is below 10%. Mental health self-service equipment is particularly weak. There are 2.9%. From this point of view,

**TABLE 8 |** Assignment of related variables.

Relevant factor	Option	Assignment	Relevant factor	Option	Assignment
Gender	Male	1	Living situation	Individual living alone	1
	Female	2		With lover	2
Age	60–69 years	1		With children	3
	70–79 years	2		With lover and children	4
	80–89 years	3		With the elderly	5
	Over 90 years old	4		Individual living alone	6
Education	Uneducated	1	Number of children	Five or more	1
	Primary school	2		Four	2
	Junior high school	3		Three	3
	Technical secondary school	4		Two	4
	High school	5		One	5
	Junior college	6		Nil	6
	Undergraduate	7	Service provision conditions	Expense cost	1
Income situation	Below 2,000 yuan	1		Information security	2
	2,001–3,000 yuan	2		Service docking by the community	3
	3,001–4,000 yuan	3		Combined with traditional service methods	4
	4,001–5,000 yuan	4		Simple equipment operation	5
	Above 5,000 yuan	5			

the current devices adopted by smart elderly care users are mostly concentrated on the auxiliary devices of daily life, and the adoption behavior in medical health and psychological services is relatively lacking. The current situation of elderly people equipped with intelligent elderly care service equipment is shown in **Table 7**.

## 4 LOGISTIC REGRESSION ANALYSIS OF FACTORS INFLUENCING THE ADOPTION BEHAVIOR OF SMART ELDERLY SERVICE USERS

### 4.1 Variable Assignment

Seven factors, including gender, age, education level, income, and living conditions, are included as independent variables. The dependent variable is “whether to use smart pension service” (yes = 1 and no = 0), and use multi-factor unconditional logistic analysis to explore the factors affecting the demand for smart elderly services. For the logistic regression analysis of the factors influencing the adoption behavior of intelligent elderly care service users, we have allocated the relevant variables, as shown in **Table 8**.

### 4.2 Result Analysis of Factors Affecting the Adoption Behavior of Smart Elderly Service Users

The influencing factors of adoption behavior of intelligent home-care users in Shanghai are not the same. This article used the multivariate unconditional logistic regression method to analyze. The adoption behavior of intelligent pension users was set as “Y”, independent variable  $X = (x_1, x_2, \dots, x_7)$ , “P” was the response probability of the model, and the corresponding logistic regression model is as follows:

$$Y_i = \ln\left(\frac{p_i}{p_m}\right) = \beta_0 + \sum_{j=1}^n \beta_j X_j (j = 1, 2, 3, 4, 5, 6, 7).$$

The survey data showed that the correlation of age, number of children, living conditions, service cost, service docking channels, equipment operation difficulty, and other factors are less than 0.05, that is, there is an obvious correlation between each factor and adoption behavior. In other words, it has a significant impact on the adoption behavior of intelligent elderly service users.

Among the demographic characteristics, factors such as age, number of children, and living conditions have a significant impact on the demand for intelligent elderly care services. Among them, the elderly aged 60–79 have a significant positive impact on the demand for smart elderly care, that is, the older the elderly, the higher is the demand for smart elderly care services. Analysis of the reasons shows that as age increases, whether the elderly are in action, there will be more and more needs in terms of medical and health care, life care, and meal services. However, for the elderly over 80 years old, age has no effect on their demand for intelligent elderly care services. The only child is also an important factor affecting the demand for intelligent elderly care services. In order to reduce the pressure of supporting children, the elderly will prefer to choose intelligent elderly care services in order to reduce the pressure of care and supervision in daily life. The living environment is also an influencing factor in the demand for intelligent elderly care services. Elderly people living alone generally lack the physical and psychological care provided by family members. They will encounter more risk problems in their daily life, and the additional risks brought by age increase are increasing. The demand for smart elderly care services is also increasing.

In terms of service provision needs, the elderly care about whether the use of intelligent elderly care services is suitable for the elderly. The difficulty of equipment operation and the connection channels of the services will affect the adoption behavior of the elderly. If the service provision is suitable for

**TABLE 9 |** Logistic regression analysis results of factors influencing the adoption behavior of smart elderly service users.

Variable		B	S.E	Wals	p value	Exp (B)
Age	60–69 years	0.759	0.216	12.319	0.000	0.468
	70–79 years	1.433	0.442	10.498	0.001	0.239
Only child		1.308	0.623	4.412	0.036	0.27
Individual living alone		1.2	0.351	11.727	0.001	3.321
Expense cost		−0.886	0.267	11.016	0.001	0.412
Service docking by the community		0.762	0.2	14.497	0.000	0.467
Simple equipment operation		1.474	0.237	38.62	0.000	0.229

the elderly, the elderly user groups are generally willing to accept intelligent elderly care services. This may be because with the increase in age and the influence of the educational background, the elderly groups are generally weak in accepting and adapting to new things. Insufficient age-appropriateness of intelligent elderly care equipment will cause the elderly to face various problems during their use, which will increase the pressure on the provision of elderly care services. **Table 9** shows the results of the logistic regression analysis of factors influencing the adoption behavior of intelligent elderly service users.

## 5 FACTORS AND PROBLEMS RESTRICTING THE DEVELOPMENT OF INTELLIGENT ELDERLY CARE SERVICES

### 5.1 Insufficient Motivation for the Development of Smart Elderly Care Services, Lack of Guidance, and Supervision Mechanisms

The target of intelligent elderly care services is the elderly, and the intelligent service methods are still relatively unfamiliar to them. At present, the popularity of intelligent elderly care services is relatively low, largely because the elderly have a low awareness of new elderly care methods and methods. Through the survey, it is found that the elderly population has a general understanding of smart elderly care services. Among them, 59.3% of the elderly do not know much about smart elderly care services, 10.9% have never heard of smart elderly care services, and 23.4% of the elderly have basic knowledge. Only 5.8% and 0.4% of the elderly have knowledge and knowledge of intelligent elderly care services, respectively, and the information source channels are also concentrated on community bulletin boards, the Internet, and TV. It can be seen that the government and society have relatively weak publicity for intelligent elderly care services, and the relatively single channel for the elderly to obtain information makes it more difficult to promote intelligent elderly care services. At the same time, as a new type of elderly care service, the government has not established complete legal safeguards. There are still many system loopholes in information security, ethics, and quality supervision. When the elderly are exposed to new things, they will be more cautious. In the absence of the corresponding supervision measures and system protection, the acceptance of the elderly will be greatly affected.

### 5.2 Construction of the Intelligent Elderly Care Service Platform Lacks Linkage, and the Service Supply Application Is Poor for the Elderly

The construction of intelligent elderly care services is still in the initial stage of development, and the items and contents of the service supply are still very basic, resulting in inadequate access to and processing of health data, and a complete sharing mechanism has not yet formed. Each service item is independent and independent of each other. The linkage is poor. If elderly users want to adopt smart elderly care services, they need to be equipped with a variety of smart devices. The cumbersome use process will seriously affect the elderly's sense of use. In terms of the technical design of smart elderly care services, the complicated operation steps of related equipment, small screens, small buttons, and unclear voices will also affect the suitability of smart elderly care services. In addition, children spend less time with children and cannot help the elderly at any time. There remain problems with the use of equipment. Therefore, the poor adaptability of smart elderly care services will seriously reduce the desire to use by the elderly.

### 5.3 Supply and Demand of Smart Elderly Care Services Are Not Matched, and the Product Types Are Severely Homogenized

In the questionnaire survey, it was found that the elderly equipped with more smart products and devices are smart home appliances and smart phone watches, which are less equipped with equipment similar to self-service physical examination equipment and mental health self-service equipment. In the needs survey, it was found that the needs and preferences of the elderly group were concentrated in life care, medical care, and mental health. In related surveys that affect the satisfaction of elderly care services, most people think that the lack of service types is an important factor. From this point of view, the current supply and demand of smart elderly care services are relatively low, and the smart elderly care equipment currently provided cannot meet the diverse needs of the elderly for smart elderly care services. At the same time, the current smart elderly care lacks affordable customized services for elderly people of different ages, different living habits, and different cognitive abilities. The survey found that more than half of the elderly believe that the cost of smart elderly care services should be borne by most government subsidies; the rest is



paid by the individual. Therefore, the price factor is an obstacle to the promotion of intelligent elderly care services.

## 6 COUNTERMEASURES AND SUGGESTIONS FOR THE DEVELOPMENT OF INTELLIGENT ELDERLY CARE SERVICES

### 6.1 Realizing Multiple and Co-Governance of Intelligent Elderly Care Services With the Government as the Center

The supply of smart elderly care services involves multiple responsible entities. In order to eliminate the fragmentation problems in the supply of smart elderly care, it is necessary to unite multiple forces to achieve multiple co-governance. First of all, elderly care services are a kind of public goods, which should be led by the government to allocate resources. At the macro-level, the government conducts a scientific and reasonable top-level design, strengthens the construction of legal norms and systems, formulates complete privacy and security and ethical standards, and conducts system implementation and service quality supervision. It is also necessary to encourage, support, guide, and nurture the participation of other subjects in the process of intelligent elderly care services and actively build a complete intelligent elderly care service system. Second, it is necessary to give play to the enterprise's characteristics of strong flexibility and a high degree of specialization in the process of service provision so as to stimulate market vitality and improve the quality and efficiency of intelligent elderly care services; the government should not only formulate sound preferential and tax policies but also constantly improve the market mechanism and provide institutional guarantee for non-government entities such as enterprises. Finally, it is necessary to emphasize the active role of social members in policy publicity and product promotion. The elderly have a single channel for obtaining information. Information dissemination among group members is an important channel for their information acquisition. This "invisible hand" must be fully utilized to carry out quality supervision and evaluation, information transmission, and promotion.

### 6.2 Improvement in the Supply of Related Facilities for Smart Elderly Care Services Based on Demand

The supply of smart elderly care services is a complete operating system. Basic network facilities, smart elderly service hardware facilities, and community service supporting facilities are closely related to the quality of smart elderly care services. The current Internet operating costs are relatively high, especially for the elderly; the network expenses are relatively high, and basic network facilities should be reduced and speeded up to ensure the application rate of the Internet. It is also necessary to strengthen the upgrading and transformation of intelligent elderly care equipment terminals, fully

understand the living habits, acceptance, learning ability, and information literacy of the elderly, and explore their adaptability transformation paths so as to increase the popularity of intelligent elderly service hardware facilities. In addition, it is necessary to strengthen the training of intelligent applications for the elderly so that the elderly can master the basic application methods of intelligent elderly care equipment, increase publicity, and provide after-sales tracking services.

### 6.3 Optimization of Service Methods Based on the Platform to Achieve Dynamic Integration

The most important link in the supply of smart elderly care services is "service." The supply of smart elderly care equipment is only the first step in service provision. The key to services is to establish a unified intelligent elderly care service platform to unify the health data and life needs of the elderly, process, and make corresponding feedback. In the survey, it is found that the elderly generally believe that to realize the transformation from traditional elderly care services to intelligent elderly care services, it is necessary to implement not only one method but also multiple methods in parallel. The demand of the elderly for smart elderly care service methods is to realize the combination of traditional services and smart services, that is, to use both online and offline service channels flexibly. To realize the expandability and extension of service items and service methods, and ensure the flexible handling of internal service and external interfaces, the basic life needs of the elderly must be fully supplied by the community, and the special life needs must be related to hospitals, governments, etc. Departments do a good job of docking to realize data sharing.

### 6.4 Promotion of the High-Quality Development of Intelligent Elderly Care Services With Talents as the Core

Smart senior care service personnel must possess a variety of professional skills such as daily nursing, health care, and intelligent applications, as well as master platform operation and service processes. From the policy level, it is necessary to formulate a sound personnel training and introduction system, introduce professional skills and service quality standards for intelligent elderly care service personnel, formulate professional training courses and teaching materials, and rely on universities for professional personnel training. At the same time, it is necessary to actively encourage members of the public to participate in smart elderly care services, broaden the source of service personnel, establish a complete talent training system, and make full use of the practical experience of such groups to enhance their ability to provide intelligent services. At the social level, it is necessary to enhance the professional identity of elderly care service personnel through methods such as increasing salary and social status and establish a complete career promotion channel. The quality of intelligent elderly care services is a key issue for the elderly, and the key to improving the quality of services is to have professional service personnel to provide efficient services. The current quality of elderly care services varies, and intelligent applications require

relatively high quality of personnel. Therefore, to ensure the quality of intelligent elderly care services, it is necessary to focus on improving the quality and professionalism of service personnel.

## 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.

## REFERENCES

- Gao, L., Liu, D., and Song, S. (2019). Analysis on the Influencing Factors of the Demand Willingness of the Elderly in the Community for "Smart Endowment". *Health Vocat. Educ.* (12), 123–125.
- Jain, D. K., Jain, R., Lan, X., Upadhyay, Y., and Thareja, A. (2021). Driver Distraction Detection Using Capsule Network. *Neural Comput. Applic* 33 (11), 6183–6196. doi:10.1007/s00521-020-05390-9
- Liao, C. (2019). Solving the Overall Problem of Smart Elderly Care Service and its Realization Path. *Econ. Manag. Rev.* 35 (06), 5–13. doi:10.13962/j.cnki.37-1486/f.2019.06.001
- Mao, Y., and Li, D. L. (2015). Research on Influencing Factors of Smart Elderly Users' Use Behavior Based on UTAUT Model: A Case Study of Wuhan One-Key Link [J]. *E-Government* 11, 99–106. doi:10.16582/j.cnki.dzzw.2015.11.001
- Wang, Z. (2020). Supply and Demand Evaluation and Development Countermeasures of "Internet + Pension Service" Model [J]. *Zhongzhou Academic Journal* 2020 (03), 81–86.
- Wu, Y. (2019). The Development path of "Internet + Smart Pension". *People's Forum* 13, 76–77.
- Wei, Y., and Xu, Y. (2019). Difficulties and Path Development of Smart Industry: A Case Study of Shaanxi Province [J]. *Econ. manag. rev.* 35(03), 37–45. doi:10.19331/j.cnki.jxufe.2020.03.005
- Zhang, L., and Han, Y. L. (2017). The Main Models, Existing Problems and Countermeasures of Smart Elderly Care in China. *Soc. Secur. Res.* 02, 30–37.

## AUTHOR CONTRIBUTIONS

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

## FUNDING

This project is supported by the Key project of National Social Science Foundation of China (No. 21AGL024).

- Zhang, Q., and Li, H. (2019). "What is Possible" to "What is Feasible": Foreign Research Progress and Enlightenment of Intelligent Elderly Care. *Learn. Pract.* 02, 109–118. doi:10.19624/j.cnki.cn42-1005/c.2019.02.013
- Zhu, H., and Tang, C. (2020). Social Risk and Legal System Arrangement of Intelligent Pension. *J. Jishou Univ. Soc. Sci. Ed.* 41 (05), 27–36. doi:10.13438/j.cnki.jdxnb.2020.05.004

**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.

Copyright © 2022 Luo and Meng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Effect of Aerobic Exercise on Lipid Metabolism in Rats With NAFLD

Tongxi Zhou<sup>1</sup>, Mengfan Niu<sup>2</sup>, Ruichen Liu<sup>2</sup> and Li Li<sup>1\*</sup>

<sup>1</sup>College of Sports and Human Sciences, Harbin Sport University, Harbin, China, <sup>2</sup>Graduate Faculty, Harbin Sport University, Harbin, China

## OPEN ACCESS

### Edited by:

Deepak Kumar Jain,  
Chongqing University of Posts and  
Telecommunications, China

### Reviewed by:

Xuetao Li,  
Hubei University of Automotive  
Technology, China  
Kangjie Sun,  
Shanghai Jiaotong University, China  
Jiansong Fang,  
Guangdong Baiyun University, China

### \*Correspondence:

Li Li  
lili@hrbipe.edu.cn

### Specialty section:

This article was submitted to  
Human and Medical Genomics,  
a section of the journal  
Frontiers in Genetics

Received: 22 March 2022

Accepted: 06 May 2022

Published: 16 June 2022

### Citation:

Zhou T, Niu M, Liu R and Li L (2022)  
Effect of Aerobic Exercise on Lipid  
Metabolism in Rats With NAFLD.  
Front. Genet. 13:901827.  
doi: 10.3389/fgene.2022.901827

This work aimed to study the intervention effect of exercise on lipid metabolism in NAFLD rats, provide a more scientific experimental basis for exploring and improving the theoretical system of exercise intervention in NAFLD, and further provide a theoretical research basis for clinical treatment of NAFLD. Forty healthy male Sprague Dawley rats were randomly divided into a blank control group (BC, 14) and a model group (MO, 26). After 6 weeks of modeling, the MO group was randomly divided into the model control group (MC, 12) and the aerobic exercise group (AE, 12). Platform running intervention in group E was conducted at a slope of 0°, a speed of 15 m/min, 1 h/time, once a day, six times a week, and a day of rest, for 8 weeks in total. After the intervention, the liver tissues of rats were taken for pathological sections, and serum was taken and analyzed for TC, TG, LDL-C, HDL-C, and FFA levels. Under the light microscope, the liver tissue structure of rats in the BC group was complete and clear, the structure of liver lobules was clear and normal, the volume of hepatocytes was uniform, the nucleus was in the middle, and the cytoplasm was red-stained, and no steatosis of hepatocytes was found. The liver of rats in the MC group showed diffuse fatty lesions, disordered structure of hepatic lobules, disordered arrangement of hepatic cords, different sizes of hepatocytes, loose cytoplasm, and diffuse lipid droplets of different sizes in the cytoplasm. The accumulation of liver lipid droplets in the AE group was improved compared with the MC group, the number of fat vacuoles in hepatocytes was significantly reduced, and the degree of liver lipid deposition was reduced. Compared with the BC group, the content of TC, TG, LDL-C, and FFA in the serum of the MC group increased significantly ( $p < 0.01$ ), and the content of HDL-C decreased significantly ( $p < 0.01$ ). Compared with the MC group, the content of TC, TG, LDL-C, and FFA in the serum of the AE group decreased significantly ( $p < 0.01/p < 0.05$ ), and the content of HDL-C increased significantly ( $p < 0.01$ ). Therefore, moderate-intensity aerobic exercise has an intervention effect on lipid metabolism in NAFLD rats, which can be used as one of the means to treat NAFLD.

**Keywords:** nonalcoholic fatty liver disease, aerobic exercise, lipid metabolism, effect of aerobic exercise, NAFLD

## 1 INTRODUCTION

As the most common chronic liver disease (NAFLD), nonalcoholic fatty liver disease has become a huge and increasingly serious public health problem. At this stage, China is in a period of an aging population. It is expected that by the end of 2030, the prevalence of NAFLD in China will increase to 22.2%, and the number of patients will reach 314 million (Estes et al., 2018). At the same time, under

the risk factors such as obesity, sedentary lifestyle, and/or susceptible genetic background in the past 15 years, the prevalence of NAFLD in children tends to be higher, and the prevalence of NAFLD in children will reach 13–60%. It has become one of the most common liver diseases in children (Mann et al., 2018), but the pathogenesis of NAFLD has not been fully clarified, and the treatment method is not ideal, which brings a heavy economic and psychological burden to society and families. Therefore, it is necessary to explore the pathogenesis of NAFLD and adopt effective treatment strategies for NAFLD in order to understand the epidemic and reduce the disease burden, which has become one of the essential problems to be solved urgently in front of medical workers.

## 1.1 NAFLD

NAFLD refers to the clinicopathological syndrome diagnosed as primary hepatic steatosis by histological detection or imaging, excluding secondary hepatic fat accumulation factors, such as fatty drugs, genetic diseases, excessive drinking, etc. (Grundy et al., 2005). It is the specific manifestation of metabolic syndrome in the liver. The main pathological change is the imbalance between liver lipid synthesis and oxidation, resulting in abnormal accumulation of triglycerides (TG) and total cholesterol (TC) in hepatocytes and the formation of liver lipid droplets (Donnelly et al., 2005; Malhi and Gores, 2008; Neuschwander-Tetri, 2010). According to the histopathological changes of the liver, the pathological classification of NAFLD ranges from simple hepatic steatosis to Nash with/without fibrosis, to fatty liver cirrhosis and hepatocellular carcinoma (Roeb et al., 2015). Most NAFLD patients only have simple hepatic steatosis, of which about 30% of NAFLD patients will develop into more severe Nash, accompanied by hepatocyte injury and inflammation. Studies have shown that (Vernon et al., 2011) the risk of death from liver disease in nonalcoholic steatohepatitis (NASH) patients with fibrosis and cirrhosis can be increased by 50–80 times, which indicates that the stage of simple hepatic steatosis is not completely benign and has the potential to further develop into hepatic fibrosis. The risk of liver cirrhosis and hepatocellular carcinoma needs to strengthen in prevention and treatment intervention. By the end of 2021, although there are drugs that coexist through the treatment of metabolic syndrome and anti-NAFLD characteristics, no drugs that are completely effective in the treatment of NAFLD have been approved. Lifestyle intervention including diet control and increased physical activity is the first-line treatment of NAFLD (Kraus et al., 2002; Yen et al., 2015).

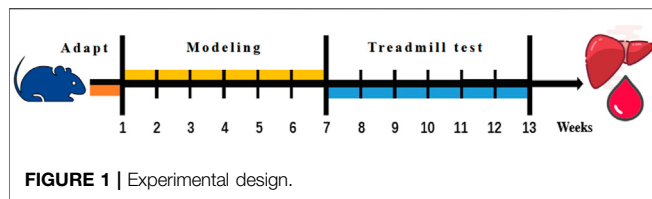
## 1.2 The Pathogenesis of NAFLD

The potential mechanism of NAFLD occurrence and development is caused by many factors. In the beginning, Day and James put forward the “second strike” theory (Qian et al., 2021). The first strike is mainly insulin resistance (IR). The inhibitory effect of insulin on hormone-sensitive lipase is weakened, the ability of peripheral fat decomposition is increased, the content of free fatty acid (FFA) in blood and the intake of FFA by the liver are increased, resulting in lipid

deposition in hepatocytes, which increases the sensitivity of liver to damage mediated by the second strike. Based on this, the hepatic uptake of FFA increases, but the oxidative consumption rate of FFA by mitochondria is limited, resulting in the accumulation of FFA in the liver. At the same time, when mitochondria oxidize a large amount of FFA, they will produce excessive reactive oxygen species (ROS) and lipid peroxide, which will cause mitochondrial dysfunction and oxidative stress in the liver, and cause inflammation, fibrosis, and necrosis of hepatocytes; At the same time, the accumulation of excessive FFA, ROS and lipid peroxide in the liver can induce the occurrence of IR, directly affect the metabolic environment of NAFLD, activate various risk factors and innate immune response, recruit various inflammatory factors, and further lead to hepatocyte injury and form a malignant cycle. With the deepening of research, it is found that many factors related to the occurrence and development of NAFLD will affect and promote each other, causing repeated blows to the liver. Therefore, the “second strike” theory has been unable to fully explain some metabolic disorders and molecular mechanisms in the occurrence and development of NAFLD. The “multiple strike” theory proposed by Tilg and Moschen has gradually replaced the “second strike” theory (Buzzetti et al., 2016). The theory of “multiple blows” clarifies that dietary habits, and environmental and genetic factors can jointly affect the changes of IR, oxidative stress, lipid metabolism disorder, adipocyte proliferation and dysfunction, and intestinal microbiota. The occurrence and development of NAFLD result from the interaction between multiple potential pathways and multiple injuries.

As the central regulator of lipid homeostasis, the liver participates in many essential links in lipid metabolism, including lipid uptake and synthesis, lipid processing, storage, oxidative decomposition and output, and their utilization as energy substrates. When the uptake of FFA by the liver exceeds the oxidation and output of FFA, lipids are deposited in hepatocytes in the form of TG, which eventually leads to lipid peroxidation stress and liver injury. Therefore, TG deposition is the key to the formation of NAFLD, and the mechanism of lipid metabolism disorder is not completely clear, which may be related to the following links: increased lipid uptake by the liver, increased *de novo* lipogenesis by the liver, imbalance of lipid oxidation, disorder of lipoprotein synthesis and output, resulting in an imbalance in the synthesis, degradation, and secretion of liver lipid metabolism, resulting in the decrease of TG transport out of hepatocytes. Finally, lipids are abnormally deposited in hepatocytes.

To sum up, liver fat deposition results from unbalanced TG synthesis and oxidation output. Excessive energy intake is the initiating factor, followed by the increase of FFA in the blood, which results in an increase in liver lipid intake. Excessive carbohydrates in the diet can promote an increase in lipid synthesis. Excessive FFA mobilization causes mitochondria. The oxidation ability is impaired, the production of ROS and lipid peroxide is increased, the liver fat deposition and the activation of the liver necrotizing inflammatory response are increased, and the ability of lipid oxidation is decreased. In



conclusion, the role of liver lipid deposition in NAFLD and its potential mechanism has important basic theoretical significance, and also provide the molecular biological basis and treatment strategies for the prevention and treatment of the disease.

### 1.3 Progress of Exercise Improving NAFLD

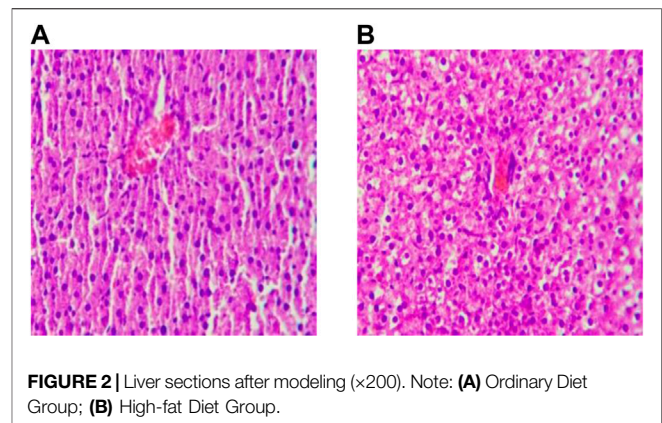
It has become a consensus that a lack of physical exercise can increase the risk of NAFLD in recent years. On the contrary, physical training can reduce the content of liver fat and reduce hepatic steatosis, which is an effective strategy for the treatment of NAFLD. Through the research, it is found that the selection of reasonable exercise mode, the planning of the best exercise intensity, and the formulation of personalized exercise therapy will have a scientific, reasonable, and high curative effect on NAFLD.

#### 1.3.1 Effects of Different Exercise Modes on NAFLD

The selection of NAFLD exercise mode is mainly divided into aerobic exercise and resistance exercise. Both of them can improve the serological indexes and pathological scores of NAFLD. Aerobic exercise is an aerobic metabolic way to consume sugar and fat to provide energy for the body when oxygen is fully supplied to tissues and cells, including jogging, swimming, Tai Chi, and other forms of exercise. In the systematic evaluation study, the relationship between the changes in ALT, AST, IR index, and body mass index and the effects of aerobic exercise training or dietary intervention on NAFLD patients was evaluated. Compared with the observation group, aerobic exercise training plus diet showed a good trend and effect on ALT, AST, IR index, and BMI. In addition, aerobic exercise can regulate and improve the hepatic steatosis and IR status of obese mice by enhancing fat phagocytosis (Li et al., 2021). However, recent studies have shown that the effect of aerobic exercise on the improvement of intrahepatic lipids is not comprehensive, but it can significantly improve IR and lipid droplet size, suggesting that aerobic exercise can delay some aspects of NAFLD, but cannot reduce all metabolic disorders (La Fuente et al., 2019). In addition, in recent years, the prevention and treatment of NAFLD have extended multi-means joint intervention, in order to obtain a better curative effect than simple exercise intervention.

#### 1.3.2 Effect of Different Exercise Intensities on NAFLD

Training intensity may play a key role in enhancing the protective effect of physical exercise on NAFLD. Studies have shown that 12 weeks of moderate-intensity aerobic training can reduce liver fat content in sedentary obese men with NAFLD. Its mechanism may be to regulate lipid metabolism and obesity-related inflammatory state *in vivo* by reducing the gene expression level of lipid synthesis in monocytes (Oh et al., 2017). Low-intensity exercise can also improve NAFLD, but if the load is low,



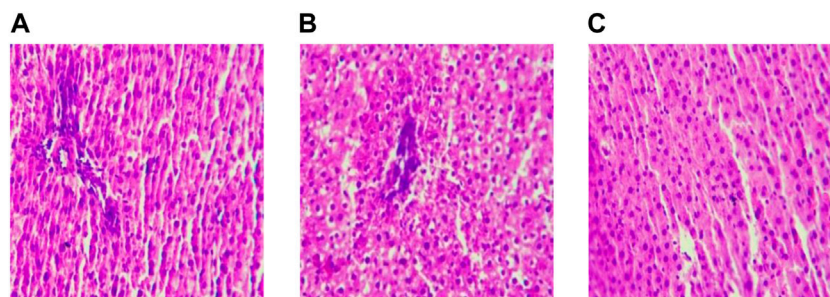
the intervention on liver lipid droplets is not significant. Sufficient load generated during high-intensity exercise can effectively resist and alleviate the formation of liver lipid droplets. Studies have shown that high-intensity training and dietary intervention at 70% VO<sub>2</sub>max intensity for 6 weeks can improve SOD expression level and T-SOD activity, reduce the degree of lipid peroxidation, and then significantly inhibit the progress of hepatic steatosis (Jiang et al., 2021). In addition, the study of 12-weeks high-intensity exercise training in OLETF rats showed that high-intensity training was effective in improving the liver hardness and restoring Kupffer cell function (Linden et al., 2015). Compared with other exercise intensities, high-intensity exercise is a method of getting twice the result with half the effort because of its short exercise duration and exercise cycle. Still, it needs to be carried out under the scientific guidance of professionals.

To sum up, although the best exercise prescription for the prevention and treatment of NAFLD cannot be fully determined, according to the recommendations of Easl-Easd-Easo clinical practice guide (European Association for the Study of The Liver et al., 2016). The physical activities of the NAFLD population need to reach at least 150–200 Min per week, 3–5 times per week and medium intensity; At the same time, considering that resistance exercise and high-intensity intermittent exercise has a clear adaptive population and require standardized training of professionals, compared with people with chronic metabolic diseases such as NAFLD, the risk of exercise under maximum short-term intensity is high, and the actual operation is difficult, which needs to be further evaluated in practical application; Exercise can be selected according to the individual habit and exercise habit. In addition, in recent years, the prevention and treatment of NAFLD have extended multi-means joint intervention, in order to obtain a better curative effect than simple exercise intervention.

## 2 MATERIALS AND METHODS

### 2.1 Animals

Male Sprague Dawley rats (200 ± 20 g, 25–35 days of age) provided by Liaoning Changsheng Biotechnology Co., Ltd.



**FIGURE 3 |** Liver sections after intervention( $\times 200$ ). Note: **(A)** Blank Control Group; **(B)** Model Group; **(C)** Aerobic Exercise Group.

**TABLE 1 |** Serum lipid metabolism indexes ( $\bar{x} \pm S$ ).

Index group	Blank control group (n = 12)	Model group (n = 12)	Aerobic exercise group (n = 12)
TG (mmol/L)	$0.85 \pm 0.10^{**}$	$1.56 \pm 0.20$	$0.92 \pm 0.13^{**}$
TC (mmol/L)	$0.94 \pm 0.20^{**}$	$1.73 \pm 0.10$	$1.21 \pm 0.09^{**}$
HDL-C (mmol/L)	$1.63 \pm 0.18^{**}$	$1.01 \pm 0.12$	$1.48 \pm 0.22^{**}$
LDL-C (mmol/L)	$0.47 \pm 0.08^{**}$	$1.21 \pm 0.09$	$0.78 \pm 0.13^{**}$
FFA (mmol/L)	$0.42 \pm 0.02^{**}$	$0.76 \pm 0.05$	$0.46 \pm 0.05^{*}$

Compared with the model group,  $^{*}p < 0.05$ ,  $^{**}p < 0.01$ .

(production license No. scxk (Liao) 2020-0001) and raised in the animal laboratory of Heilongjiang University of Chinese Medicine (Harbin, Heilongjiang, China). Rats were housed in groups of 6 animals per plastic cage under controlled conditions of light and temperature ( $25 \pm 3^{\circ}\text{C}$ ) and relative humidity (60–70%). Rats were allowed free access to standard laboratory food and tap water, and to adapt to the laboratory for at least 1 week before the onset of the experimental protocol.

## 2.2 Experimental Design

Forty rats were selected for this study and were fed either a normal diet (blank control group, 14) or a high-fat diet (model group, 26) (2% Cholesterol, 0.5% sodium cholate, 5% sucrose, 10% lard and 82.5% basic feed) for 14 weeks. All animal experiments in this study were approved by the animal laboratory of Heilongjiang University of Chinese Medicine.

After six weeks, two rats in each group were killed and the materials were obtained. To determine whether the modeling was successful by observing the results of hematoxylin-eosin (HE) staining of liver tissue. After successful modeling, the model group was randomly divided into the model control group (MC, 12) and the aerobic exercise group (AE, 12).

Referring to T.G.Bedford et al. (Bedford et al., 1979) classic rat treadmill test experiment, the AE group in this experiment adopted the simple speed-up exercise scheme of medium intensity: the treadmill slope was constant  $0^{\circ}$ , the speed in the first stage was 5 m/min for 5 min, and the speed in the second stage was 15 m/min for 55 min. They trained at the same time every Tuesday to Sunday, and rested on Monday. It lasted for 8 weeks. (Figure 1).

## 2.3 Tissue Processing

24hs after the end of the exercise protocol, rats were injected with chloral hydrate 10% (4 ml/kg) in the abdominal cavity. Blood was harvested, and serum was also collected and stored at  $-80^{\circ}\text{C}$  for TC (100,000,220), TG (192061), VDL-C (1,00,020,245), HDL-C (1,00,020,235), and FFA (A042-2-1) analysis. After the completion of blood collection, the liver tissue was dissected quickly and fixed in 4% paraformaldehyde solution for HE staining.

## 2.4 Statistical Analyses

The results were expressed as mean  $\pm$  standard deviation ( $\bar{x} \pm s$ ). The data of each group were tested for normal distribution and homogeneity of variance. An independent-sample t-test was used to compare the differences between the two groups. A value of  $p < 0.01$  was considered more statistically significant.

## 3 RESULTS

### 3.1 Effect of Liver Histomorphology With High-Fat Diet

After feeding for six weeks, the hepatocytes of rats in the BC group were arranged regularly. The cytoplasm, nucleus, hepatic cord, and hepatic sinuses were clear. The staining was uniform, and there was no edema and inflammatory cell infiltration in the liver tissue. The hepatocytes of rats in the MO group were grid-shaped, showing obvious steatosis, the obvious proliferation of connective tissue around large blood vessels, and local scattered inflammatory cells gathered in piles, but there was no obvious



congestion and expansion of capillaries. The results are consistent with the relevant Literature (Wang and Guang, 2007) report, which indicates that the model establishment of NAFLD in rats was successfully established. (Figure 2).

### 3.2 Effect of Liver Histomorphology With Treadmill Exercise

After feeding for eight weeks, the hepatocytes in the MC group still showed a grid shape, the degree of steatosis increased, and the proliferation of connective tissue around large vessels, local inflammatory cell aggregation, and pyknosis of hepatocytes increased significantly. After 8 weeks of aerobic exercise intervention, the liver histopathology of the AE group was improved compared with the MC group. (Figure 3).

### 3.3 Effect of Treadmill Exercise on the Serum Lipid Metabolism Indexes

The levels of TG, TC, LDL-C, and FFA in the serum of the MC group were significantly higher than those of the BC group ( $p < 0.01$ ), while the level of HDL-C was significantly lower than that of the BC group ( $p < 0.01$ ). After aerobic exercise intervention, there was also a very significant difference between the AE group and MC group ( $p < 0.01$ ), and the index results were between the MC group and BC group. The results showed that the lipid deposition was obvious in rats with NAFLD, and aerobic exercise could then improve the serum lipid metabolism imbalance state by regulating the serum lipid metabolism markers. (Table 1).

## CONCLUSION

NAFLD is a syndrome characterized by a disorder of hepatocyte lipid metabolism (hepatic steatosis) without alcohol and other clear factors of liver damage. Generally speaking, lipid metabolism is regulated through a variety of ways, among which *de novo* lipogenesis is one of the main mechanisms leading to the increase of FFA transport and accumulation in the liver. (Linden et al., 2015) (European Association for the Study of The Liver et al., 2016). This study found that compared with the rats in the BC group, the contents of TC, TG, LDL-C, and FFA, increased HDL-C decreased in the MC group. It indicates that NAFLD rats fed a high-fat diet for a long time caused an imbalance of lipid metabolism, excessive accumulation of

hepatocyte lipids, and hepatocyte dysfunction. After aerobic exercise intervention, TC, TC, LDL-C, and FFA in the AE group decreased, HDL-C increased, and the blood lipid index of NAFLD rats improved, which is consistent with the research results of Linden (Linden et al., 2015) and Xu (Xu, 2006). It indicates that regular and appropriate aerobic exercise intervention can inhibit the storage of redundant lipid substances in tissues and cells and improve the disorder of blood lipid metabolism by increasing heat consumption, accelerating lipid oxidation, and reversing lipid deposition, so as to provide a theoretical basis for anti-NAFLD composite targets and joint intervention and provide a reference for the application of aerobic exercise combined with other polysaccharides in the prevention and treatment of NAFLD. Therefore, exercise intervention has unique advantages and certain application prospects.

## 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

All animal experiments in this study were approved by the animal laboratory of Heilongjiang University of Chinese Medicine.

## AUTHOR CONTRIBUTIONS

TZ and MN: editing, data curation, and supervision. RL and LL: writing—original draft preparation.

## FUNDING

This study was supported by the Basic Science Foundation of Laboratory platform project of Harbin Sport University (LAB 2021-5), Heilongjiang Provincial Undergraduate Universities in 2021 (2021KYYWF-PY04), and Discipline Echelon's Construction Foundation of Harbin Sport University (XKB10 and XKL10).

## REFERENCES

- Bedford, T. G., Tipton, C. M., Wilson, N. C., Oppliger, R. A., and Gisolfi, C. V. (1979). Maximum Oxygen Consumption of Rats and its Changes with Various Experimental Procedures. *J. Appl. Physiology* 47 (6), 1278–1283. doi:10.1152/jappl.1979.47.6.1278
- Buzzetti, E., Pinzani, M., and Tsochatzis, E. A. (2016). The Multiple-Hit Pathogenesis of Non-alcoholic Fatty Liver Disease (NAFLD). *Metabolism* 65 (8), 1038–1048. doi:10.1016/j.metabol.2015.12.012
- Donnelly, K. L., Smith, C. I., Schwarzenberg, S. J., Jessurun, J., Boldt, M. D., and Parks, E. J. (2005). Sources of Fatty Acids Stored in Liver and Secreted via Lipoproteins in Patients with Nonalcoholic Fatty Liver Disease. *J. Clin. Invest.* 115 (5), 1343–1351. doi:10.1172/jci23621
- Estes, C., Anstee, Q. M., Arias-Loste, M. T., Bantel, H., Bellentani, S., Caballeria, J., et al. (2018). Modeling NAFLD Disease Burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the Period 2016–2030. *J. Hepatology* 69 (4), 896–904. doi:10.1016/j.jhep.2018.05.036
- European Association for the Study of The Liver; European Association for the Study Of Diabetes; European Association for the Study Of Obesity (2016). EASL-EASD-EASO Clinical Practice Guidelines for the Management of Non-alcoholic Fatty Liver Disease. *J. Hepatol.* 64 (6), 1388–1402. doi:10.1016/j.jhep.2015.11.004

- Grundy, S. M., Cleeman, J. I., Daniels, S. R., Donato, K. A., Eckel, R. H., Franklin, B. A., et al. (2005). Diagnosis and Management of the Metabolic Syndrome. *Circulation* 112 (17), 2735–2752. doi:10.1161/circulationaha.105.169404
- Jiang, T., Liu, M., and Yan, D. (2021). Effect of High Intensity Interval Training on the Formation of Nonalcoholic Fatty Liver in Rats Induced by Dietary. *J. Xinjiang Med. Univ.* 44 (01), 44–48+54.
- Kraus, W. E., Houmard, J. A., Duscha, B. D., Knetzger, K. J., Wharton, M. B., McCartney, J. S., et al. (2002). Effects of the Amount and Intensity of Exercise on Plasma Lipoproteins. *N. Engl. J. Med.* 347 (19), 1483–1492. doi:10.1056/nejmoa020194
- La Fuente, F. P.-d., Quezada, L., Sepúlveda, C., Monsalves-Alvarez, M., Rodríguez, J. M., Sacristán, C., et al. (2019). Exercise Regulates Lipid Droplet Dynamics in Normal and Fatty Liver. *Biochimica Biophysica Acta (BBA) - Mol. Cell Biol. Lipids* 1864 (12), 158519. doi:10.1016/j.bbalip.2019.158519
- Li, H., Dun, Y., Zhang, W., You, B., Liu, Y., Fu, S., et al. (2021). Exercise Improves Lipid Droplet Metabolism Disorder through Activation of AMPK-Mediated Lipophagy in NAFLD. *Life Sci.* 273, 119314. doi:10.1016/j.lfs.2021.119314
- Linden, M. A., Fletcher, J. A., Morris, E. M., Meers, G. M., Laughlin, M. H., Booth, F. W., et al. (2015). Treating NAFLD in OLETF Rats with Vigorous-Intensity Interval Exercise Training. *Med. Sci. Sports Exerc* 47 (3), 556–567. doi:10.1249/mss.0000000000000430
- Malhi, H., and Gores, G. J. (2008). Molecular Mechanisms of Lipotoxicity in Nonalcoholic Fatty Liver Disease. *Semin. Liver Dis.* 28 (4), 360–369. doi:10.1055/s-0028-1091980
- Mann, J. P., Valenti, L., Scorletti, E., Byrne, C. D., and Nobili, V. (2018). Nonalcoholic Fatty Liver Disease in Children. *Semin. Liver Dis.* 38 (1), 1–13. doi:10.1055/s-0038-1627456
- Neuschwander-Tetri, B. A. (2010). Hepatic Lipotoxicity and the Pathogenesis of Nonalcoholic Steatohepatitis: the Central Role of Nontriglyceride Fatty Acid Metabolites. *Hepatology* 52 (2), 774–788. doi:10.1002/hep.23719
- Oh, S., So, R., Shida, T., Matsuo, T., Kim, B., Akiyama, K., et al. (2017). High-Intensity Aerobic Exercise Improves Both Hepatic Fat Content and Stiffness in Sedentary Obese Men with Nonalcoholic Fatty Liver Disease. *Sci. Rep.* 7, 43029. doi:10.1038/srep43029
- Qian, X., Wang, T., Gong, J., Wang, L., Chen, X., Lin, H., et al. (2021). Exercise in Mice Ameliorates High-Fat Diet-Induced Nonalcoholic Fatty Liver Disease by Lowering HMGCS2. *Aging* 13 (6), 8960–8974. doi:10.18632/aging.202717
- Roeb, E., Steffen, H. M., Bantel, H., Baumann, U., Canbay, A., Demir, M., et al. (2015). S2k Guideline Non-alcoholic Fatty Liver Disease. *Z Gastroenterol.* 53 (7), 668–723. doi:10.1055/s-0035-1553193
- Vernon, G., Baranova, A., and Younossi, Z. M. (2011). Systematic Review: the Epidemiology and Natural History of Non-alcoholic Fatty Liver Disease and Non-alcoholic Steatohepatitis in Adults. *Aliment. Pharmacol. Ther.* 34 (3), 274–285. doi:10.1111/j.1365-2036.2011.04724.x
- Wang, Q., and Guang, X.-Q. (2007). Improvement of Induction Method of Non-alcoholic Fatty Liver Model in Rats. *World Chin. J. Dig.* 2007 (11), 1219–1224.
- Xu, S.-S. (2006). Influences of Walking on Some Blood Biochemical Index of NAFLD Patients. *J. Xi'an Phys. Educ. Univ.* 2006 (05), 79–81+101.
- Yen, C.-Y., Hou, M.-F., Yang, Z.-W., Tang, J.-Y., Li, K.-T., Huang, H.-W., et al. (2015). Concentration Effects of Grape Seed Extracts in Anti-oral Cancer Cells Involving Differential Apoptosis, Oxidative Stress, and DNA Damage. *BMC Complement. Altern. Med.* 15, 94. doi:10.1186/s12906-015-0621-8

**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.

Copyright © 2022 Zhou, Niu, Liu and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Performance of Problem-Based Learning Based Image Teaching in Clinical Emergency Teaching

Xiaohong Xu<sup>1</sup>, Yingcui Wang<sup>2</sup>, Suhua Zhang<sup>3\*</sup> and Fengting Liu<sup>4\*</sup>

<sup>1</sup>Department of Rheumatology and Immunology, Qilu Hospital (Qingdao), Cheeloo College of Medicine, Shandong University, Qingdao, China, <sup>2</sup>Department of Education, Qilu Hospital (Qingdao), Cheeloo College of Medicine, Shandong University, Qingdao, China, <sup>3</sup>Department of Geriatrics, Qilu Hospital (Qingdao), Cheeloo College of Medicine, Shandong University, Qingdao, China, <sup>4</sup>Department of Emergency, Qilu Hospital (Qingdao), Cheeloo College of Medicine, Shandong University, Qingdao, China

## OPEN ACCESS

### Edited by:

Deepak Kumar Jain,  
Chongqing University of Posts and  
Telecommunications, China

### Reviewed by:

Ping Wang,  
Jiangxi University of Technology,  
China  
Yifeng He,  
Northwestern Polytechnical  
University, China  
Peilin Chen,  
China University of Labor Relations,  
China

### \*Correspondence:

Suhua Zhang  
bzmuwzk123@stu.bzmc.edu.cn  
Fengting Liu  
pdf@163.com

### Specialty section:

This article was submitted to  
Human and Medical Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 29 April 2022

**Accepted:** 02 June 2022

**Published:** 27 June 2022

### Citation:

Xu X, Wang Y, Zhang S and Liu F  
(2022) Performance of Problem-Based  
Learning Based Image Teaching in  
Clinical Emergency Teaching.  
Front. Genet. 13:931640.  
doi: 10.3389/fgene.2022.931640

At present, with the rapid increase of emergency knowledge and the improvement of people's requirements for medical quality, the traditional teaching mode cannot fully meet the needs of emergency teaching in the new era. PBL is a project-based teaching that allows students to have a deeper understanding of content knowledge and to better apply what they have learned to their lives. This paper aims to improve the clinical emergency teaching mode by PBL teaching method, and improve the comprehensive ability of clinical emergency of medical students. This article proposes a problem-based PBL imaging teaching method, combining the characteristics and content of clinical emergency courses, focusing on students, highlighting the problem-solving process, and improving students' creative thinking ability. To cultivate students' interest in clinical learning, develop their self-learning ability, train their teamwork and communication skills, and cultivate their ability to set, question and solve questions, so as to promote medical students' overall comprehensive ability to integrate specialized knowledge and clinical practice. In this paper, the PBL teaching method and the traditional teaching method of comparative experiments show that the PBL teaching method can more effectively highlight the characteristics of clinical emergency medicine teaching mode, and make full use of the limited emergency teaching resources, so as to improve the quality of clinical emergency teaching. Compared with the traditional teaching mode, the theoretical knowledge and clinical operation skills of medical students under the PBL teaching mode are improved by 13%, Autonomous learning ability, communication ability and creative thinking ability have also been relatively improved.

**Keywords:** PBL teaching method, traditional teaching method, imaging diagnosis, clinical emergency, performance of PBL

## 1 INTRODUCTION

With the rapid evolution of mankind from the large-scale industrial age to the information age and even the age of artificial intelligence, we are increasingly aware of the seriousness of the various challenges that physicians have to face. The unified teaching method based on classroom teaching is the most widely used teaching method in medical colleges and universities in China. The traditional teaching mode focuses on teaching knowledge, although it can cultivate medical students with solid basic theoretical knowledge, but the lack of clinical practice ability and problem-solving ability will

cause the phenomenon that theoretical knowledge and practical operation ability are divorced, so it is difficult to adapt to the clinical practice process. A relatively single teaching method often loses students' interest and independent thinking ability, making teachers' guidance and students' learning burdens heavier and more difficult. Therefore, in clinical emergency education, in addition to transferring knowledge, it is more important for students to learn independent thinking, literature review, and the ability to realize autonomous learning, and to improve students' ability to explore and solve problems, so as to cultivate high-quality suitable for the development of the times Medical talents. At the same time, this teaching adapts to the trend of modern teaching and the development of emergency medicine. Applying this model not only enables medical students to quickly grasp the relevant knowledge points, but also cultivates their learning ability and improves their ability to effectively apply their knowledge when dealing with clinical problems.

Although originally designed as a teaching method for graduate admissions in medicine, problem-based learning (PBL) has been widely used in undergraduate medicine, science, and social science courses. Although it is generally acknowledged that new learners of PBL need support, this does not provide a good description of the undergraduate course, which makes some students feel incomprehensible. In this submission, Moro C provides many extensive considerations and practical suggestions to support the transition of learners to PBL and universities. In a globalized world where the diversity of learners is increasing, this kind of support is particularly important in a learner-centered and socially responsible higher education pattern. But his research is not comprehensive enough, and more in-depth research is needed to prove the feasibility of this method (Moro and Mclean, 2017). The formative assessment of clinical teaching for emergency medicine (EM) teachers is limited. The purpose of this study is to develop a behavior-based tool to assess EM teachers' clinical teaching skills during shift and provide feedback. Dehon e uses a three-stage structured development process. In the first stage, the nominal group technology is used to communicate with a group of teaching staff, and then with residents to generate potential assessment projects. Phase 2 includes an independent focus group and uses improved Delphi technology to evaluate the projects generated in phase 1 with faculty and residents as well as a group of experts. Since then, residents have divided the programs into novice, intermediate and advanced educator skills. Once the items to be included are identified and then ranked, the investigators build them into the tool (phase 3). Results: the final tool "teacher shift card" is a behavior anchored assessment and feedback tool to facilitate feedback to EM teachers on their teaching skills during the shift. However, his method is not supported by specific experimental data and lacks accuracy of experimental results (Dehon et al., 2018). In the past 20 years, social work education has built a bridge between the classroom and the scene by learning clinical skills through customer simulation. Pecukonis EV outlines an innovative simulation model, combined with LS, for teaching motivation interview (MI). In addition, he also discussed the guiding principles and specific steps of using simulation and LS to teach MI. Unfortunately, most

simulation models ignore the method of real-time guidance and monitoring students. In the implementation of clinical simulation, there are few opportunities to correct students' behaviors or practice new skills at the best teachable time in the interview. This instruction must wait until the end of the interview and the beginning of the presentation. With the addition of LS, the students' simulation experience has been enhanced, because the tutor is now an active participant in the interview. With LS, teachers can now direct and even simulate appropriate clinical responses and interventions. However, the research model is too complex to be put into practical application (Yan, 2017; Pecukonis, 2021; Ramadan, 2021).

The innovation of this article lies in the advantages of the PBL teaching method, breaking through the traditional learning method, giving full play to the student-centered teaching method, cultivating students' interest and ability in autonomous learning, and improving the ability of autonomous learning. The article also integrates integrated cognitive activities such as questioning, judging, comparing, choosing and analyzing, synthesizing and generalizing knowledge. ability. Ability to think, solve problems, and solve problems; In order to improve the quality of clinical emergency teaching, it is necessary to cultivate students' teamwork ability, expand their knowledge, and master professional clinical emergency operation skills.

## 2 IMPLEMENTATION OF PROBLEM-BASED LEARNING BASED IMAGE TEACHING METHOD IN CLINICAL EMERGENCY

### 2.1 Problem-Based Learning Teaching Method

PBL is a problem based learning (PBL) which was founded by Howard Barrows, an educator in 1969 at the Medical College of McMaster University in Canada. That is, problem-based learning is a student-centered, group oriented and teacher oriented learning method (Umami, 2018; Liu, 20192020). The PBL model originated in medical education and is an inquiry-based teaching method, that is, a core learning approach for research-based learning, integrated practice courses. This method is in the form of group discussion in a complex and substantive problem situation, so that students can solve problems through independent research and cooperation, and learn and obtain the scientific knowledge behind the problems, so as to make it possible. It is a kind of education method combining basic science and clinical practice. The biggest feature of this teaching method is to change the passive learning of traditional teaching methods into active learning (Petruck, 2017; Gallagher et al., 2017). According to current research, in this teaching method, students are first grouped, and then questions are set. Under the guidance of teachers, each group of students discusses and analyzes to find answers to the problems. In the teaching process, firstly, students learn the professional theoretical knowledge more deeply by searching for answers on their own; secondly, through multiple rounds of group discussion and summarization, they have exercised their learning ability,



communication skills and teamwork ability, and finally students deal with the actual situation. The ability to question can also be greatly improved (Stephanie and Hughes, 2016; Tarling et al., 2016). The traditional teaching method (Lecture-Based Learning), that is, the method of “learning based on teaching,” referred to as LBL (Light and Razak, 2020; Xu et al., 2022), is a traditional theoretical teaching method based on the teacher’s teaching (Ananga, 2020). It is an education-based teacher-centered teaching method that emphasizes the transfer of knowledge in the classroom. Cannot meet the needs of students for knowledge and practice, development skills are not enough to improve students’ ability to explore and solve problems (Dzombak et al., 2020; Posteraro et al., 2020). PBL education is a new educational strategy that allows students to find effective solutions to situational problems. This educational strategy is very consistent with the current theoretical theories and educational principles (Macqueen et al., 2016). In the medical field, the PBL medical teaching model is a process in which medical students work in small groups, with the participation and guidance of a tutor, to formulate, discuss and learn about a complex, multi-scenario, real-world problem-based topic or case.

## 2.2 Interpretation of Low-Quality Medical Images With Deep Learning Methods

### 2.2.1 Image Degradation Mechanism

Assuming that A is a low-quality image and B is a corresponding high-quality image, the image degradation mechanism can be expressed as follows:

$$A = \sigma(B). \quad (1)$$

Among them,  $\sigma$  represents the image degradation function in a broad sense, such as additive noise, multiplicative noise, compound noise, etc.

### 2.2.2 Deep Learning Image Restoration Paradigm

Low quality images arise because of the presence of many unnecessary or redundant interfering information in the image data, which must be corrected prior to image enhancement processing and classification processing. The image restoration model based on deep learning can be expressed as a regression paradigm of supervised learning: assuming that the network model is M, it represents the modeling function of the image degradation mechanism, which can be regarded as the optimal approximation to the inverse function  $\sigma^{-1}$ ; The input of is the low-quality image A, and the output result is expressed as M(A). The restoration requires that M(A) be as close as possible to the corresponding high-quality image Y. Training the recovery network can be seen as optimizing the following loss function:

$$\frac{\arg \min}{M} + \frac{1}{2} M(A) - B^2. \quad (2)$$

The above loss function generally only considers the mean square error of the gray values of the pixels corresponding to the restored image M(A) and the high-quality image B. Representative algorithms include MLP, CNN, etc. In addition to directly learning high-quality images, some also link residual learning and restoration, directly learning the noise distribution of the

image, and the regression paradigm can be expressed as the following form:

$$\frac{\arg \min}{M} + \frac{1}{2} M(A) - \sigma^2. \quad (3)$$

Among them, M(A) is the image noise fitted by the network, which is the real image noise, which can be obtained by the following calculation

$$\varphi = A - B, \quad (4)$$

That is, the difference between a low-quality image and a high-quality image. Compared with directly learning high-quality images, learning the noise distribution makes the optimization of the network easier, but to obtain the final restored image N, a further subtraction operation is required:

$$N = A - M(A). \quad (5)$$

We can basically get the recovered image by the above subtraction operation, but we need to enhance the image in order to fully obtain the information in the image.

## 2.3 Gradient Regularization Image Enhancement Model

The results of image processing depend heavily on the mathematical model developed. Even for the same image, the results obtained using different mathematical models may vary greatly. Therefore, in the early stage of image restoration, people generally use known *a priori* information to build a gradient model.

### 2.3.1 Gradient Regularization Image Restoration Paradigm

$$\frac{\arg \min}{M} + \frac{1}{2} M(A) - B^2 + \frac{\rho}{2} t \otimes M(A) - B^2, \quad (6)$$

Where t is the gradient operator, which represents the convolution operation, and is the balance factor. It can also be simplified to:

$$\frac{\arg \min}{M} + \frac{1}{2} M(A) - B^2 + \frac{\rho}{2} t \otimes (M(X) - B)^2. \quad (7)$$

## 2.4 Heuristic Information

Image enhancement is a widely used digital image processing technique, and currently, there are many methods used for image enhancement. As with other image restoration methods, since there is a corresponding image as a reference, the method mentioned above is no longer used here, and the calculation formula is as follows:

$$10 \ln_{10} \left( \frac{MAX^2}{MSE(C, Z)} \right). \quad (8)$$

MAX represents the maximum gray value of the image. The image similarity is measured from three aspects: brightness, contrast, and structure. The calculation formula is as follows:

$$k(v, e) = \frac{2\eta_v\eta_e + x_1}{\eta_v^2 + \eta_e^2 + x_1}. \quad (9)$$

$$j(v, e) = \frac{2\sigma_v\sigma_e + x_2}{\sigma_v^2 + \sigma_e^2 + x_2}. \quad (10)$$

$$i(v, e) = \frac{\sigma_{ve} + x_3}{\sigma_v\sigma_e + x_3}. \quad (11)$$

$$SSIM(v, e) = k(v, e) + j(v, e) + i(v, e). \quad (12)$$

Among them,  $\eta_v, \eta_e$  represent the mean value of the restored image  $v$  and the reference image  $e$ ,  $\sigma_v, \sigma_e$  represent the standard deviation of  $v, e$ ,  $\sigma_{ve}$  represents the covariance between the two,  $x_1, x_2, x_3$  are smoothing constants, used To avoid the situation where the denominator appears to be 0, the value is generally entered as follows:

$$x_1 = (H_1 \times G)^2. \quad (13)$$

$$x_2 = (H_2 \times G)^2. \quad (14)$$

$$x_3 = \frac{x_2^2}{2}. \quad (15)$$

Generally,  $H_1 = 0.01, H_2 = 0.03$ , and  $L$  represent the maximum value of the image gray scale, and the value range is  $[0, 1, 0]$ . The smaller the value, the lower the image distortion. The higher the image distortion rate, the higher the risk of introducing information or components that do not exist in the original image, which can have serious consequences, such as the addition of tissue components to biomedical images, which can affect the doctor's diagnosis of the condition.

### 3 EXPERIMENTS BASED ON PROBLEM-BASED LEARNING TEACHING METHOD IN CLINICAL EMERGENCY

#### 3.1 Experiment Object

Using the cluster sampling method, 50 students who practiced in a tertiary A hospital were used as the control group and received traditional teaching methods; 48 students who practiced in the emergency department were used as the experimental group and received PBL imaging teaching methods. After the teaching, the theory and skills of the two groups of students were evaluated; before and after the teaching, the two groups of students used self-study assessment tools, Skala Attitudes Skills Communication Skills and the Chinese version of the Critical Thinking Scale (Gruppen, 2017; Messman et al., 2020). According to the internship arrangement, evaluate and statistically analyze the evaluation results, and include all interns in the internship evaluation form. And before entering the emergency internship, at the end of the departmental internship, the comprehensive examination transcript will be evaluated, and the evaluation results will be statistically and data analyzed. The general information of the experimental group and the control group was compared with the pre-hospital assessment scores, and the difference was not statistically significant ( $p > 0.05$ ), which was comparable.

The control group used traditional teaching methods, centered on the instructor, and intensively taught once a week. The content of the course included two basic elements: theoretical knowledge and technical skills. One week before the end of the internship, the trainees will be evaluated based on theory and skills.

The experimental group adopts the PBL teaching method, emphasizing the student-centered approach, integrating clinical case group discussions and simulated clinical scenarios into the teaching process, allowing students to analyze cases and find answers, thereby improving their own abilities. In this kind of curriculum construction, we break the content arrangement of the disease as the outline, teaching content as the outline of symptoms, and set the common clinical symptoms of emergency medicine, such as: dyspnea, chest pain, abdominal pain, coma, etc.

#### 3.2 Teaching Preparation

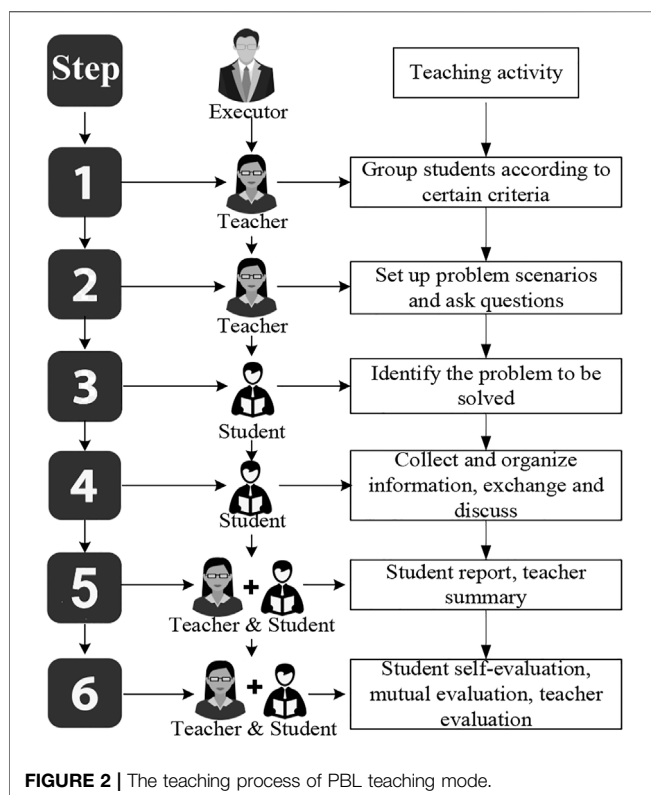
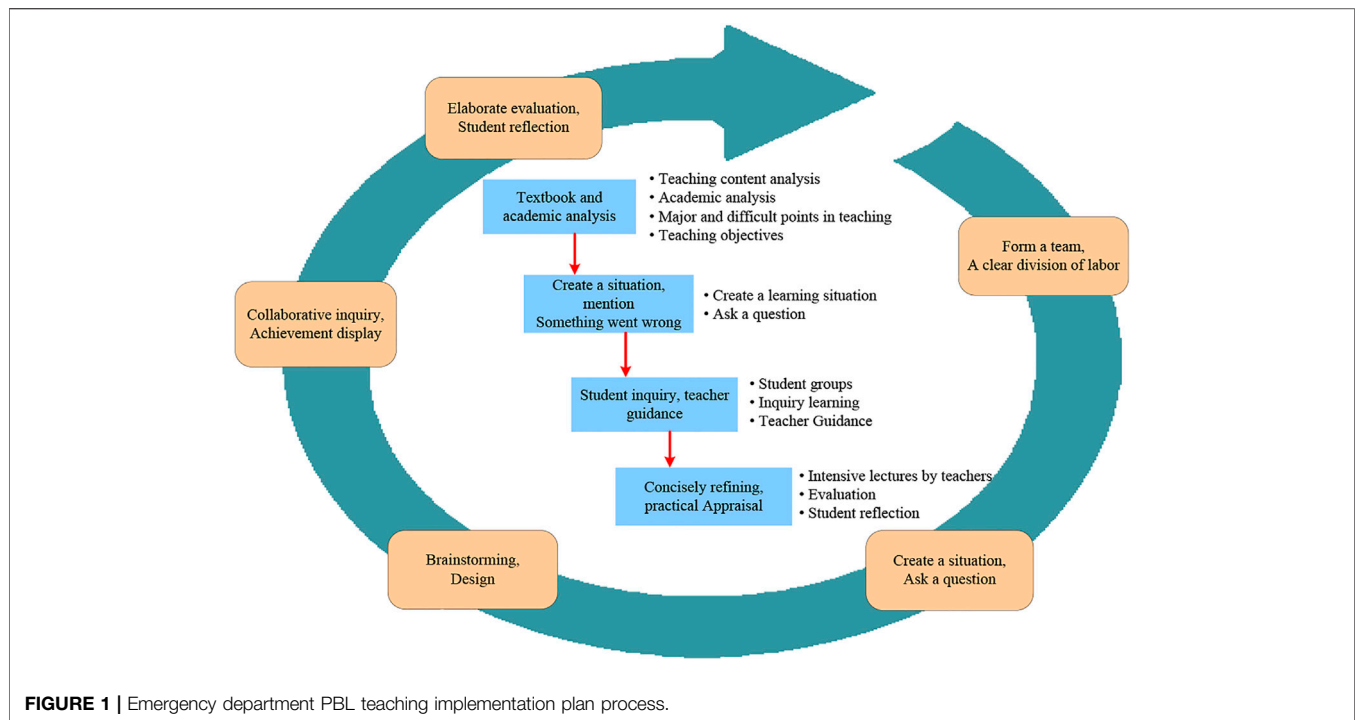
Improve clinical education resources: establish and improve the clinical literature search network in hospitals, emergency departments, and student cards (Kim et al., 2018; Brown, 2021). Strictly select teachers with lessons: select clinical education teachers and create a clinical education team in the emergency department. The team is evaluated and approved by the education and research department of the emergency department, and selects emergency educators with solid basic theoretical knowledge, professional knowledge and awareness, responsibility and communication skills, and excellent knowledge dissemination skills (Dull et al., 2018).

#### 3.3 Draw Up an Education Plan

Complete the training content of the teachers in the experimental group, and determine a unified teaching plan based on the mastery of the relevant PBL teaching concepts and methods, and the syllabus, purpose, requirements, and content of the clinical emergency teaching course. Using multimedia to explain the concept of PBL teaching method and the teaching method based on question and discussion (Dehon et al., 2019; Karki, 2020; Tabassum et al., 2022), organize and teach teachers to discuss methods of mobilizing students' enthusiasm. Determine the PBL teaching content, such as case selection for frequently-occurring diseases that are more practical. Complete the PBL problem design PBL teaching cases are developed with the most typical emergency and frequently-occurring disease as the core, establish poorly structured problems, and complete the PBL problem design. Prepare 3 cases, each batch of students in the emergency department are required to complete one case, each case is divided into three times, each case has three acts, each act is arranged for independent study (2 days), concentrated discussion (1 class hour), and Group summary (1 class hour).

#### 3.4 Draw Up a Training Plan

Pre-admission training control group for students: explain the syllabus, purpose, content, and assessment requirements of clinical emergency medicine students, and distribute critical thinking survey forms and questionnaires on clinical medicine students' cognition and attitude towards PBL. Emphasize that during the clinical emergency internship, strictly abide by the rules and regulations of clinical emergency (Levine et al., 2016; Bertino et al., 2021). Experimental group: Explain the syllabus, purpose, content and assessment requirements of clinical emergency medicine students, and distribute critical thinking survey forms and questionnaires on medical students' cognition



and attitude towards PBL. Then carry out the training before the implementation of PBL, that is, explain the concept of the PBL teaching model, the implementation method, the purpose of

implementing PBL in the clinical emergency department, and the main points of cooperation. Distribute PBL teaching cases to allow students to preview in advance and ask students to analyze and discuss the cases, prepare and record their presentations.

Pre-admission training control group for students: based on traditional knowledge transmission, the one-to-one teaching method, namely LBL teaching method, is based on the traditional knowledge transmission, the teaching teacher implements the teaching of clinical emergency students according to the internship syllabus and requirements. Teaching tasks, first review relevant theoretical knowledge, combined with common clinical emergency diseases, students complete the internship tasks in the emergency department under the leadership of the teacher. Experimental group: PBL + LBL teaching method is adopted. The specific implementation methods are as follows: 1) Self-directed learning: After entering the course, students can find the best arguments and answers through reading textbooks, searching literature, consulting teachers, etc. in combination with relevant questions raised in the case, and can share the collected evidence with each other. One student recorded his doubts or the answers he found, and reported when preparing for group discussion (Ahn et al., 2016; Runde et al., 2016). Also ask students to think about: ① How to consider the changes in the patient's condition? ② Try to analyze the causes and mechanisms of the above changes in the patient's condition? ③ The current diagnosis? ④ Current management? 2) Organizing discussion: Arrange students to report the results of the group study, ask questions based on actual clinical cases, the students report and speak, use the brainstorming method, each student puts forward their own opinions, and teaches the teacher to supplement the scenes involved in the problem and further ask questions, The team leader will make a record, and the team leader will list the best

**TABLE 1 |** Comparison of two basic data.

Project	Test group	Comparison group	t	p
Age	22.07 ± 0.47	21.75 ± 0.59	0.584	0.621
Grades	82.15 ± 3.75	82.47 ± 4.03	0.056	0.852
Gender composition (M/F)	15/33	18/32	0.042	0.781

**TABLE 2 |** Comparison of cognition and attitude of internship in PBL.

Project	Test group	Comparison group	t	p
PBL understanding	4.13 ± 0.51	3.88 ± 0.46	-1.036	>0.05
PBL participation	3.76 ± 0.19	3.45 ± 0.21	-0.974	>0.05
Willingness to participate in PBL	3.84 ± 0.42	2.79 ± 0.28	-1.482	>0.05
The need to participate in PBL	4.25 ± 0.37	3.15 ± 0.52	-1.306	>0.05

**TABLE 3 |** Comparison of the overall attitudes of the two groups of students to the experimental class.

Project	Test group	Comparison group	t	p
Whether the experimental class is important or not	2.56 ± 0.72	2.18 ± 0.45	8.163	<0.05
Satisfaction with the profession	2.64 ± 0.61	2.06 ± 0.53	4.581	<0.05
The sense of responsibility to study hard	2.58 ± 0.64	2.22 ± 0.41	9.742	<0.05
Confidence in future clinical work	2.77 ± 0.78	2.17 ± 0.38	7.985	<0.05

**TABLE 4 |** Results of emergency clinical ability of the two groups after 2 weeks of teaching.

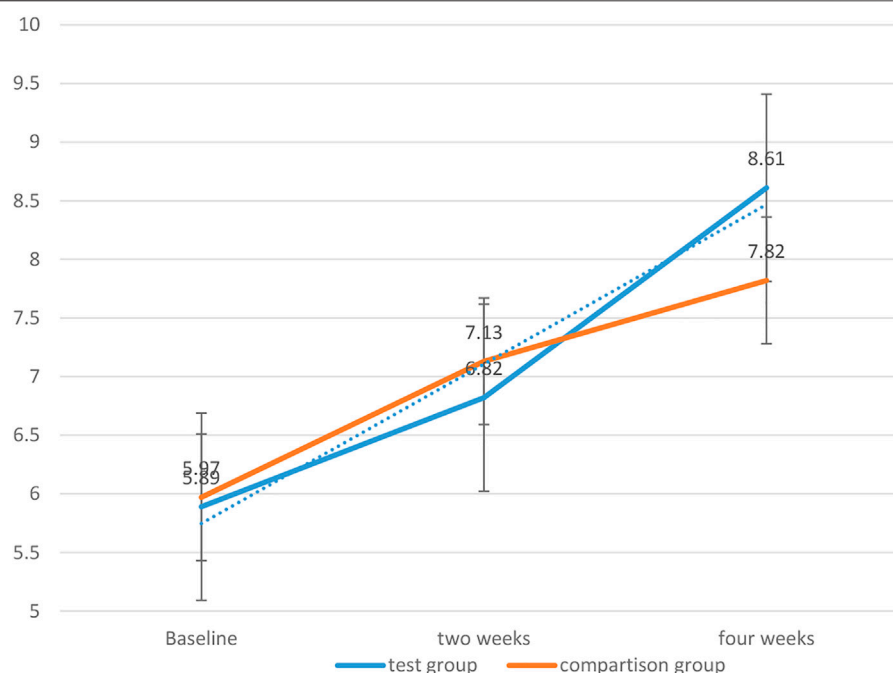
Evaluation index	Test group	Comparison group	t	p
Clinical treatment	6.03 ± 0.41	5.92 ± 0.37	0.71	>0.05
Communication and coordination	5.12 ± 0.39	4.95 ± 0.57	2.34	<0.05
health education	6.41 ± 0.52	5.76 ± 0.39	4.85	<0.05
Nursing Research	8.14 ± 0.47	7.27 ± 0.46	1.55	<0.05
Clinical teaching	7.35 ± 0.39	6.82 ± 0.27	3.89	<0.05
Clinical management	6.72 ± 0.48	6.34 ± 0.19	7.63	<0.05
Mental quality	6.85 ± 0.51	5.73 ± 0.41	5.43	<0.05

**TABLE 5 |** Results of emergency clinical ability of the two groups after 4 weeks of teaching.

Evaluation index	Test group	Comparison group	t	p
Clinical treatment	7.21 ± 0.39	6.58 ± 0.42	0.76	<0.05
Communication and coordination	6.04 ± 0.43	5.62 ± 0.37	3.15	<0.05
health education	6.93 ± 0.52	6.12 ± 0.39	5.21	<0.05
Nursing Research	9.25 ± 0.47	7.94 ± 0.46	2.33	<0.05
Clinical teaching	7.89 ± 0.39	7.29 ± 0.27	4.15	<0.05
Clinical management	7.06 ± 0.48	7.21 ± 0.19	6.53	<0.05
Mental quality	7.62 ± 0.51	6.31 ± 0.41	4.86	<0.05

answers to the questions after the discussion (Choi et al., 2019; Jain et al., 2020). In the course of the experiment, the teacher is responsible for coordinating, motivating them to actively participate in the discussion, ensuring that the discussion direction is targeted at the teaching purpose, allowing students to understand the clinical manifestations of patients during the discussion process, learn the key points of knowledge such as communication skills with patients, strengthen students'

understanding and mastery of clinical emergency operation in the emergency department, and improve clinical thinking and problem-solving ability (Sengan, Khalaf, Rao, Sharma, Amarendra, Hamad; Jain et al., 2021). 3) Summarize: After the discussion, the teacher and students will jointly summarize the characteristics of clinical emergency department, the concept of common diseases, clinical diagnosis and precautions. At the same time, students are required to complete a reflection diary and sort out



**FIGURE 3 |** Trends in emergency clinical capabilities.

**TABLE 6 |** Repeated measures analysis of variance of two groups of clinical emergency ability score.

Project	Between-group effects		Within-group effect		Interaction effect	
	F	p	F	p	F	p
Clinical treatment	18.64	<0.05	121.34	<0.05	8.53	<0.04
Communication and coordination	82.48	<0.05	184.14	<0.05	19.14	<0.05
health education	23.76	<0.05	97.39	<0.05	34.76	<0.05
Nursing Research	80.47	<0.05	68.25	<0.05	18.15	<0.05
Clinical teaching	43.85	<0.05	74.47	<0.05	21.84	<0.05
Clinical management	41.86	<0.05	66.19	<0.05	50.59	<0.05
Mental quality	78.81	<0.05	71.84	<0.05	19.18	<0.05
Total score	369.87	<0.05	683.62	<0.05	172.19	<0.05

the ideas for solving problems to promote Theory and practice (Patidar and Pichholiya, 2016; Tabassum et al., 2017).

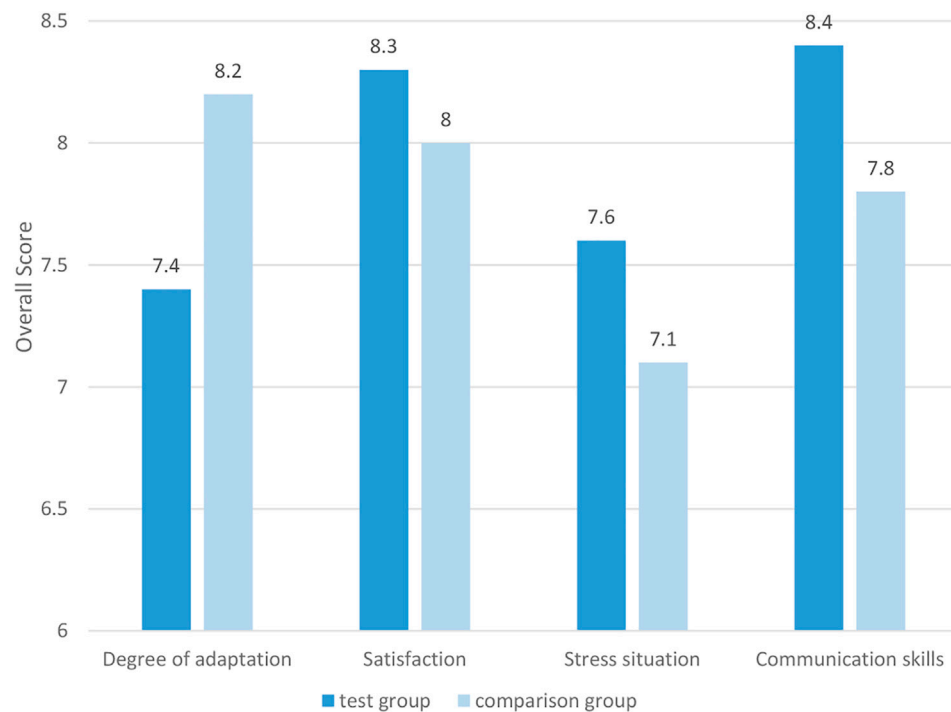
### 3.5 Data Processing and Statistical Analysis Methods

Use EXCEL software to make a unified data entry template. After a researcher has verified that the questionnaire is correct twice, the content of the questionnaire is recorded in the data template, and the survey data is statistically analyzed with SPSS 20.0 (Hexom et al., 2017), test level  $\alpha$  Set to 0.05, the  $p$  values of the two probabilities are statistically significant, so  $p < 0.05$ . 1) The measured data is expressed by ( $\pm$ S), and the measured data is expressed by frequency or percentage. 2) For the measurement data that satisfies the smoothness and uniformity of the variance, please use two independent t-test samples, and use the chi-square test to measure the data between the two groups. For

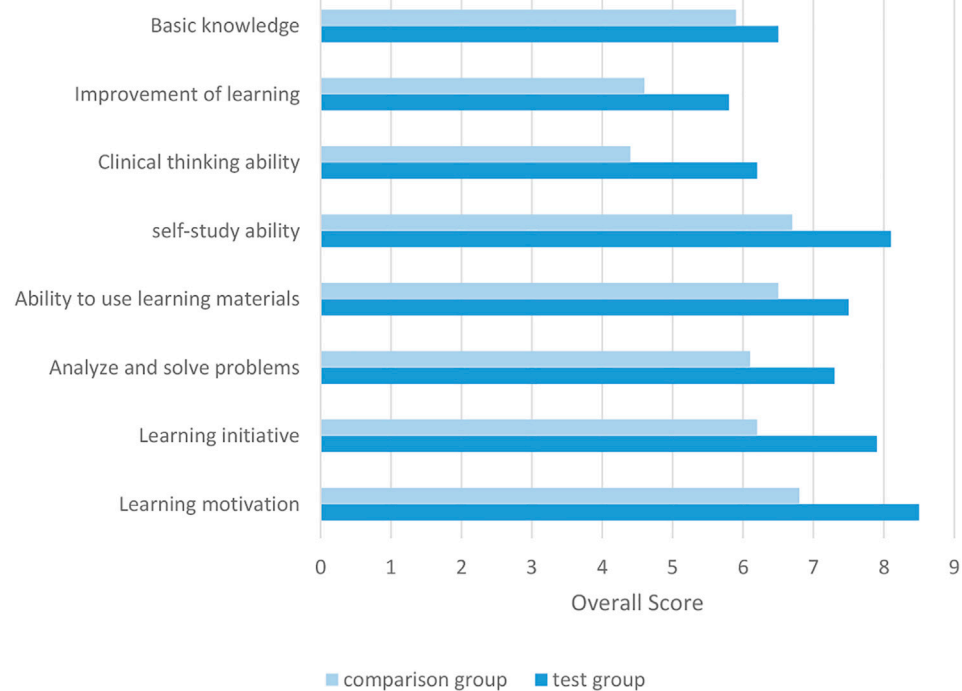
measurement data that does not satisfy the smoothness and uniformity of the variance, the rank sum test is used for the total score data, and the comparison between two or more groups is performed using ANOVA. 3) Use factor analysis and stepwise regression analysis to analyze the factors that affect critical thinking ability. 4) When comparing the emergency clinical ability before and after baseline, 2 weeks later, and 1 month later, the paired sample test is used for volume data, and the second test is used for measuring data.  $p < 0.05$  is considered to be statistically significant.

### 3.6 Problem-Based Learning Teaching Implementation Plan Process in Clinical Emergency

The implementation plan based on the PBL teaching method in clinical emergency teaching is shown in **Figure 1**. The clinical



**FIGURE 4 |** Scores for all aspects of two different teaching methods.



**FIGURE 5 |** Comparison of the improvement of the two groups of students' learning ability by two different teaching methods.



**TABLE 7 |** Comparison of experimental results.

Project	Test group	Comparison group	t	p
Theory test	88.67 ± 1.786	82.43 ± 1.658	7.541	<0.05
Operational exam	87.46 ± 1.457	85.79 ± 1.514	8.647	<0.05

emergency PBL teaching method experiment is completed according to the process, and then compared with the traditional teaching method, the experimental results are obtained.

The PBL teaching mode is a problem-based teaching mode with students as the core. The teaching process is shown in **Figure 2**:

## 4 PERFORMANCE ANALYSIS BASED ON PROBLEM-BASED LEARNING IMAGING TEACHING METHOD IN CLINICAL EMERGENCY TEACHING

### 4.1 Basic Information of the Two Groups of Students

As shown in **Table 1**, the experimental group and the control group were similar in age, basic average academic performance of professional courses, and the composition ratio of men and women, and the difference was not statistically significant ( $p > 0.05$ ).

### 4.2 Comparison of Problem-Based Learning Experimental Courses and Overall Attitudes

The basic abilities of students' PBL are scored by five levels of 1–5. As shown in **Table 2**, there is no statistical difference between the scores of the two groups of students ( $p > 0.05$ ).

As shown in **Table 3**, the experimental group and the control group have statistically significant differences in the experimental classes, satisfaction, sense of responsibility, and confidence in future work ( $p < 0.05$ ). It can be seen that the scores of the students in the experimental group are higher than those in the control group. Therefore, it can be concluded that the overall attitude of the students in the experimental group is better than that of the control group.

## 4.3 Teaching Situation

Two independent sample tests were used to compare the clinical ability scores of the two groups of students during two and 4 weeks of teaching. The result is shown in the figure below:

As shown in **Table 4**, except for the dimension of clinical handling ability ( $p > 0.05$ ), the scores of the PBL teaching experiment group and the total score of emergency clinical ability were higher than those of the control group, the difference was statistically significant ( $p < 0.05$ ). As shown in **Table 5**, after 4 weeks of teaching, except for the clinical management ability, the scores and total scores of the clinical emergency ability of the PBL teaching group were significantly higher than those of the control group, and the difference was statistically significant ( $p < 0.05$ ).

**Figure 3** depicts the trends of the clinical emergency ability of the two groups of students with the internship time. The results show that the two groups of students have improved with the elapse of internship in the emergency department, but the PBL group has a larger increase than the control group. It is statistically significant ( $p < 0.05$ ). It can be seen that the PBL teaching method has a greater improvement in clinical emergency ability.

In this experiment, repeated measures analysis of variance was used to compare the clinical first aid capabilities of the two groups of trainees after 2 and 4 weeks. The between-group effects are divided into different groups (i.e., PBL teaching group and control group), the interaction effect is the grouping time, and the between-group effects are (2 weeks, 4 weeks).

The results are shown in **Table 6**: 1) Except for the clinical management dimension,  $p$  is less than 0.05 for the inter-group effects, indicating that the differences in the overall mean values of the various dimensions of clinical emergency ability in different groups are statistically significant, that is, PBL teaching can improve students. 2) The intra-group effect  $p < 0.05$  means that the difference in the overall mean scores of the various dimensions of the clinical emergency ability and the overall average of the total score is statistically significant. With the change of time, the clinical emergency ability of the trainees has improved. 3) Except for the clinical management ability, the interaction effect is less than 0.05, which means that the interaction effect of different measurement time and group has statistical significance on the overall mean difference of the various dimensions of emergency clinical ability and the total score, indicating that 4 weeks of PBL teaching can significantly improve the clinical performance of students Emergency room capacity.

**TABLE 8 |** Comparison of the two groups of students after teaching.

Project	Type	Test group	Comparison group	t	p
Improved language organization	Y	31	18	8.457	<0.05
	N	17	32		
Improved logical thinking ability	Y	29	14	10.716	<0.05
	N	19	36		
Improved innovation ability	Y	16	15	0.974	>0.05
	N	32	35		
Improve self-learning ability	Y	40	10	12.385	<0.05
	N	8	40		

#### 4.4 Comparison of Ratings of Different Teaching Methods Between Two Groups of Students

The scoring system uses a 10-point system to score the adaptability, satisfaction, pressure, and improving communication skills of the teaching method. The results are shown in **Figure 4**.

As shown in **Figure 4**, there are statistically significant differences between the two groups of students in terms of their degree of adaptation, satisfaction, stress, and improving their communication skills ( $p < 0.05$ ). From the perspective of adaptation, The overall adaptation rate of the experimental group is lower than that of the control group. Therefore, the students in the experimental group are less adaptable to the PBL teaching method than the control group students' adaptation to the traditional teaching method. From the perspective of stress, the experimental group's pressure score is higher. Therefore, the PBL teaching method brings more pressure to the study of the experimental group students than the traditional teaching method to the control group of students. From the perspective of satisfaction and the improvement of personal communication skills, the scores of the experimental group are greater than those of the comparison group, so the PBL teaching method improves personal communication skills more than the traditional teaching method.

As can be seen in **Figure 5**, the results show that the two different teaching methods give two groups of students a statistically significant difference in the level of cultivation of various abilities,  $p < 0.05$ , and the improvement rate of the experimental group is greater than that of the control group. Therefore, the PBL teaching method gives The increase in learning ability brought by the experimental group students was more significant than the increase in learning ability brought by the traditional teaching method to the control group.

#### 4.5 Comparison of the Two Groups of Students' Theoretical and Operational Performance

The independent sample test was used to analyze the theoretical scores of the two groups of students. The results are shown in **Table 7**. The test scores of the students in the experimental group were higher than those in the control group, and the average scores in the clinical experiments were also higher than those of the control group. The difference was statistically significant ( $p < 0.05$ ). It can be seen that The PBL teaching method has a greater improvement in theoretical and operational performance.

#### 4.6 Comparison of the Questionnaire Received by the Two Groups of Students After Teaching

**Table 8** shows the comparison of the effects of the experimental group and the control group after teaching. The results of the questionnaire show that the special gains of the experimental group after teaching are greater than those of the control group, except for the improvement of innovation ability, which is not statistically significant ( $p > 0.05$ ). The improvement of self-

learning ability was the biggest difference between the experimental group and the control group, and the differences in other aspects were statistically significant ( $p < 0.05$ ).

### 5 CONCLUSION

The PBL teaching method can fully exercise people's thinking skills such as decision-making, creativity, reasoning and analysis in the process of conceiving solutions, exploring independently, making decisions and seeking solutions to problems. This article is based on the PBL teaching method to construct a clinical emergency teaching plan and carry out experimental application. In contrast to traditional teaching methods, the clinical emergency PBL teaching method constructed in this article is more prominent in improving the performance of students' clinical emergency ability, and can more effectively improve students' theoretical technical operation level, autonomous learning ability, communication ability and communication judgment ability. Thinking ability has obvious advantages in experimental teaching. It is more conducive to cultivating medical students' good study habits, enthusiasm and the ability to make full use of existing learning resources, thereby improving the comprehensive clinical professional quality of medical students. In the operation skills learning in the medical emergency, the PBL teaching method makes the learning goals and meanings of medical students clearer. The important role of various skills operations has been verified in the implementation process, and the proactiveness of the operation skills practice has been improved. However, there are still some shortcomings in this research. The traditional teaching model is deeply ingrained. A small number of teachers and students have a certain degree of rejection to the difficulty of adapting to the PBL teaching method. There is no multi-factor analysis on the factors that affect the effect of experimental teaching. Further research is needed. In the future article, we will focus on the factors that affect the effectiveness of PBL teaching practice from several levels.

### 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

XX and YW: Writing—original draft preparation. SZ and FL: Editing data curation, supervision.

### FUNDING

This work was supported by Key Fund of Department of Cardiology, Shandong University Qilu Hospital (Qingdao) (QDKY2019ZD04), People's Livelihood Science and Technology Project of Qingdao (22-3-7-smij-6-nsh), Qingdao Key Health Discipline Development Fund.

## REFERENCES

- Ahn, J., Golden, A., Bryant, A., and Babcock, C. (2016). Impact of a Dedicated Emergency Medicine Teaching Resident Rotation at a Large Urban Academic Center. *WestJEM* 17 (2), 143–148. doi:10.5811/westjem.2015.12.28977
- Ananga, P. (2020). Factors that Influence Instructors' Integration of Social Media Platforms into Higher Education Pedagogy in Ghana. *Jei* 6 (2), 118. doi:10.5296/jei.v6i2.17367
- Bertino, E., Jahanshahi, M. R., Singla, A., and Wu, R.-T. (2021). Intelligent IoT Systems for Civil Infrastructure Health Monitoring: a Research Roadmap. *Discov. Internet Things* 1, 3. doi:10.1007/s43926-021-00009-4
- Brown, J. (2021). National Institutes of Health Support for Clinical Emergency Care Research, 2015 to 2018. *Ann. Emerg. Med.* 77 (1), 57–61. doi:10.1016/j.annemergmed.2020.06.031
- Choi, J., Choi, C., Kim, S. H., and Ko, H. (2019). Medical Information Protection Frameworks for Smart Healthcare Based on IoT, Proceedings Of The 9th International Conference On Web Intelligence, Mining And Semantics (WIMS 2019).
- Dehon, E., Robertson, E., Barnard, M., Gunalda, J., and Puskarić, M. (2019). Erratum: This Article Corrects: "Development of a Clinical Teaching Evaluation and Feedback Tool for Emergency Medicine Faculty". *West J. Emerg. Med.* 20 (5), 838–839. doi:10.5811/westjem.2019.7.44616
- Dehon, E., Robertson, E., Barnard, M., Gunalda, J., and Puskarić, M. (2018). Development of a Clinical Teaching Evaluation and Feedback Tool for Faculty. *WestJEM* 20 (1), 50–57. doi:10.5811/westjem.2018.11.39987
- Dull, A., LaPonsie, S., Brown, A., Boss, J., Ray, D., Sapp, T., Jones, J., et al. (2018). Clinical Teaching in a Busy Emergency Department: Interruptions during Case Presentations. *Am. J. Emerg. Med.* 36 (9), 1003–1004.
- Dzombak, R., Pham, K., and Beckman, S. (2020). Learning Design: Examining Programmatic Learning Outcomes and the Influence of Disciplinary Perspectives on Design Pedagogy. *Int. J. Eng. Educ.* 36 (2), 623–632.
- Gallagher, M., Prior, J., Needham, M., and Holmes, R. (2017). Listening differently: A Pedagogy for Expanded Listening. *Br. Educ. Res. J.* 43 (6), 1246–1265. doi:10.1002/berj.3306
- Gruppen, L. (2017). Clinical Reasoning: De Ning it, Teaching it, Assessing it, Studying it. *WestJEM* 18 (1), 4–7. doi:10.5811/westjem.2016.11.33191
- Hexom, B., Trueger, N. S., Levene, R., Ioannides, K. L., and Cherkas, D. (2017). The Educational Value of Emergency Department Teaching: it Is about Time. *Intern Emerg. Med.* 12 (2), 207–212. doi:10.1007/s11739-016-1447-1
- Jain, D. K., Jain, R., Lan, X., Upadhyay, Y., and Thareja, A. (2021). Driver Distraction Detection Using Capsule Network. *Neural Comput. Applic* 33 (11), 6183–6196. doi:10.1007/s00521-020-05390-9
- Jain, D. K., Zareapoor, M., Jain, R., Kathuria, A., and Bachhety, S. (2020). GAN-pose: an Improved Bidirectional GAN Model for Human Motion Prediction. *Neural Comput. Applic* 32 (18), 14579–14591. doi:10.1007/s00521-020-04941-4
- Karki, O. B. (2020). Intestinal Stomas: A Clinical Study in a Teaching Hospital, Our Experience. *J. Kathmandu Med. Coll.* 9 (1), 37–42. doi:10.3126/jkmc.v9i1.33543
- Kim, H. J., Jeong, S. H., Seo, J. H., Park, I. S., and Moon, S. Y. (2018). Augmented Reality for Botulinum Toxin Injection. *Concurrency Computation-Practice Exp.* 32 (18), e5526. doi:10.1002/cpe.5526
- Levine, A. C., Barry, M. A., Agrawal, P., Duber, H. C., Chang, M. P., Mackey, J. M., et al. (2016). Global Health and Emergency Care: Overcoming Clinical Research Barriers. *Acad. Emerg. Med.* 24 (4), 484–493. doi:10.1111/acem.13142
- Light, R. L., and Razak, M. S. (2020). The Influence of Experiential Pedagogy on Undergraduate Sport Coaching Students "Real World" Practice. *Int. J. Phys. Educ. Fit. Sports* 9 (2), 37–44. doi:10.34256/ijpefs2025
- Liu, Y. (2019). John Dewey's Democratic Education and its Influence on Pedagogy in China 1917–1937. *Front. Educ. China*. 9915. Lei Wang Wiesbaden, Germany: Springer VS, 329€74683–74685. 978-3-658-27567-9. (paperback). 4. doi:10.1007/978-3-658-27568-6
- Macqueen, S., Woodward-Kron, R., Flynn, E., Reid, K., Elliott, K., and Slade, D. (2016). A Resource for Teaching Emergency Care Communication. *Clin. Teach.* 13 (3), 192–196. doi:10.1111/tct.12423
- Messman, A. M., Malik, A., and Ehrman, R. (2020). An Asynchronous Curriculum for Teaching Practical Interpretation Skills of Clinical Images to Residents in Emergency Medicine. *J. Emerg. Med.* 58 (2), 299–304. doi:10.1016/j.jemermed.2019.11.046
- Moro, C., and Mclean, M. (2017). Supporting Students' Transition to University and Problem-Based Learning. *Med. Sci. Educ.* 27 (2), 1–9. doi:10.1007/s40670-017-0384-6
- Patidar, R., and Pichholiya, M. (2016). Analysis of Drugs Prescribed in Emergency Medicine Department in a Tertiary Care Teaching Hospital in Southern Rajasthan. *Int. J. Basic Clin. Pharmacol.* 5 (6), 2496–2499. doi:10.18203/2319-2003.ijbcp20164111
- Pecukonis, E. V. (2021). Guidelines for Integrating Live Supervision in Simulation-Based Clinical Education: An Example for Teaching Motivational Interviewing. *Clin. Soc. Work J.* 49 (2), 151–161. doi:10.1007/s10615-021-00805-z
- Petruck, N. (2017). The Influence of Western European Humanistic Pedagogy on Forming Ukrainian School in 16Th-17th Centuries. *Nephron Clin. Pract.* 7 (3), 21–25. doi:10.1515/rpp-2017-0031
- Posteraro, B., Marchetti, S., Romano, L., Santangelo, R., Morandotti, G. A., Sanguinetti, M., et al. (2020). Clinical Microbiology Laboratory Adaptation to COVID-19 Emergency: Experience at a Large Teaching Hospital in Rome, Italy. *Clin. Microbiol. Infect.* 26 (8), 1109–1111. doi:10.1016/j.cmi.2020.04.016
- Ramadan, R. A. (2021). An Improved Group Teaching Optimization Based Localization Scheme for WSN. *Ijwac* 3 (No. 1), 08–16. doi:10.54216/ijwac.030101
- Runde, D. P., Arndt, C., Pettit, J., and Takacs, M. (2016). 67 Faculty and Resident Assessment of Medical Education Skills (FRAMES): Impact of a Needs Assessment and Teaching Skills Workshop on Observed Clinical Teaching. *Ann. Emerg. Med.* 68 (4), S29. doi:10.1016/j.annemergmed.2016.08.078
- Sengan, S., Khalaf, O. I., Rao, G. R., Sharma, D. K., Amarendra, K., and Hamad, A. A. (2022). Security-Aware Routing on Wireless Communication for E-Health Records Monitoring Using Machine Learning. *Int. J. Reliab. Qual. E-Healthcare (IJRQEH)* 11 (3), 1–10. doi:10.4018/ijrqeh.289176
- Stephanie, J., and Hughes, H. E. (2016). Changing the Place of Teacher Education: Feminism, Fear, and Pedagogical Paradoxes. *Harv. Educ. Rev.* 86 (2), 161–182. doi:10.17763/0017-8055.86.2.161
- Tabassum, K., Shaiba, H., Essa, N. A., and Elbadie, H. A. (2022). An Efficient Emergency Patient Monitoring Based on Mobile Ad Hoc Networks. *J. Organ. End User Comput.* 34(4):1-12. doi:10.4018/joeuc.289435
- Tabassum, S., Ali, S., and Shamsheer, S. (2017). Rate and Indications of Emergency Caesarean Sections at a Teaching Hospital in Pakistan. *Med. Forum Mon.* 28 (9), 21–24.
- Tarling, C., Jones, P., and Murphy, L. (2016). Influence of Early Exposure to Family Business Experience on Developing Entrepreneurs. *Educ. + Training* 58 (7-8), 733–750. doi:10.1108/et-03-2016-0050
- Umami, I. (2018). Moderating Influence of Curriculum, Pedagogy, and Assessment Practices on Learning Outcomes in Indonesian Secondary Education. *Journal of Social Studies Education Research* 9 (1), 60–75. doi:10.17499/jsser.37505
- Xu, Z., Zhu, G., Metawa, N., and Zhou, Q. (2022). Machine Learning Based Customer Meta-Combination Brand Equity Analysis for Marketing Behavior Evaluation. *Information Processing & Management* 59 (1), 102800. doi:10.1016/j.ipm.2021.102800
- Yan, Z. (2017). Application of PBL Bilingual Teaching Method in Clinical Probation of Gynaecology and Obstetrics. *Creative Education* 08 (4), 666–670. doi:10.4236/ce.2017.84051

**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.

Copyright © 2022 Xu, Wang, Zhang and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Looking for the Genes Related to Lung Cancer From Nasal Epithelial Cells by Network and Pathway Analysis

Noman Qureshi, Jincheng Chi, Yanan Qian, Qianwen Huang and Shaoyin Duan\*

Department of Medical Imaging, Zhongshan Hospital, School of Medicine, Xiamen University, Xiamen, China

## OPEN ACCESS

### Edited by:

Deepak Kumar Jain,  
Chongqing University of Posts and  
Telecommunications, China

### Reviewed by:

Dazhuang Li,  
Macau University of Science and  
Technology, Macao SAR, China  
Xiaotian Hao,  
Chongqing Technology and Business  
University, China  
Xiao Su,  
Xijing University, China

### \*Correspondence:

Shaoyin Duan  
xmduy@xmu.edu.cn

### Specialty section:

This article was submitted to  
Human and Medical Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 13 May 2022

**Accepted:** 13 June 2022

**Published:** 18 July 2022

### Citation:

Qureshi N, Chi J, Qian Y, Huang Q and  
Duan S (2022) Looking for the Genes  
Related to Lung Cancer From Nasal  
Epithelial Cells by Network and  
Pathway Analysis.  
Front. Genet. 13:942864.  
doi: 10.3389/fgene.2022.942864

Previous studies have indicated that the airway epithelia of lung cancer-associated injury can extend to the nose and it was associated with abnormal gene expression. The aim of this study was to find the possible lung cancer-related genes from the nasal epithelium as bio-markers for lung cancer detection. WGCNA was performed to calculate the module-trait correlations of lung cancer based on the public microarray dataset, and their data were processed by statistics of RMA and *t*-test. Four specific modules associated with clinical features of lung cancer were constructed, including blue, brown, yellow, and light blue. Of which blue or brown module showed strong connection to genetic connectivity. From the brown module, it was found that HCK, NCF1, TLR8, EMR3, CSF2RB, and DYSF are the hub genes, and from the blue module, it was found that SPEF2, ANKFN1, HYDIN, DNAH5, C12orf55, and CCDC113 are the pivotal genes corresponding to the grade. These genes can be taken as the bio-markers to develop a noninvasive method of diagnosing early lung cancer.

**Keywords:** nasal epithelium, lung cancer, WGCNA, modules, hub gene

## INTRODUCTION

In recent 50 years, the morbidity and mortality of lung cancer have significantly increased, and the 5-year mortality rate is up to 80%. The main cause is lack of effective diagnostic tools to detect early lung cancer (Pisani et al., 1999). Although high-resolution CT (HRCT) and bronchoscopy increases the diagnostic sensitivity, the screening is not feasible because of high cost or complex operation (Gupta et al., 2009; Cannioto et al., 2018; Asaad Zebari and Emin Tenekeci, 2022). Despite low complications, bronchoscopy cannot identify the extent of cancer or the size and location of small or peripheral lung cancers (Khan et al., 2016). Previous studies have shown that some gene expression of epithelial cells in the entire bronchial airway is significantly different between normal people and smokers with lung cancer and proved that the existence of some pivotal genes in the nasal epithelium was closely related to lung cancer. These genes have been applied as biomarkers and classifiers to identify the lung cancers from benign diseases (Khan et al., 2016; Team, 2017). It was suggested that this analysis is an additional noninvasive and convenient detection approach for lung cancer.

The latest progress in gene interaction network methodology is to study the potential internal relationship between functional gene clusters and clinical features (Sun et al., 2017; Timmins and

**Abbreviations:** DEGs, differentially expressed genes; GEO, Gene Expression Omnibus; GO, Gene Ontology; GS, gene significance; HRCT, high-resolution CT; KEGG, Kyoto Encyclopedia of Genes and Genomes; MM, module membership; MCC, modified Cam Clay; RMA, robust multi-array average; SYK, spleen tyrosine kinase; SCLC, small cell lung cancer; WGCNA, weighted gene co-expression network analysis.



Ashlock, 2017). Identifying important modules related to clinical features is helpful to infer the tumor mechanism and establish new targets for diagnosis or therapy. Weighted gene co-expression network analysis (WGCNA) is an effective approach based on “guilt-by-association”. It is used for identifying gene modules as candidates for biomarkers. WGCNA creates in terms of large-scale gene expression reports and the identification of centrally sited genes or hub genes, which drive key cellular signaling pathways. The systematical biology method has been used to identify the hub genes in high-grade osteosarcoma and small cell lung cancer (SCLC) and to find potential therapeutic targets (Ning et al., 2016; Shakeel et al., 2020). This study was planned to make improvements in biology methods, which might increase the diagnostic efficiency of lung cancer at early stages, with low price and non-trauma.

## MATERIALS AND METHODS

### Data Filtering

The expressional profile of GSE80796 was installed from the Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>). The data and clinical traits were reserved to analyze the difference in gene expression between the nasal epithelia of patients with less than 3 cm of early lung cancer and benign pulmonary nodule in different genders (6). Finally, there were a total of 197 samples, including 100 samples of benign pulmonary nodule (62 cases with tuberculoma, 23 with inflammatory pseudotumor, 9 with sclerosing hemangioma, and 6 with hamartoma) and 97 cases of early lung cancer (81 NSCLC and 16 SCLC).

### Data Preprocessing and Identification of Genes

At first, chip data were downloaded, including the background correction, normalization preprocess, and calculation of gene expression values. Robust multi-array average (RMA) and R language (McCall et al., 2010) were applied in the affy package, and the ComBat method was used in adjustment for batch effects. Subsequently, differentially expressed genes (DEGs) of nasal epithelia between early lung cancer and benign pulmonary nodule were identified using *t*-test in the linear models for microarray data, and the top 3,600 DEGs in the order of  $|\log FC|$  were chosen for the construction of WGCNA (Langfelder and Horvath, 2008).

### Construction of a Clustering Tree for WGCNA

The WGCNA package in R language was used to construct the gene co-expression network analysis of nasal epithelia gene expression for both male and female and then continually to compare and screen the consensus modules of nasal epithelia gene expression in different genders.

### Brief Process

The process contained the following steps: 1) created a correlation matrix of the pairs of genes from all samples. 2) Chose the proper soft threshold. 3) With the proper power value, performed the automatic network construction and module detection with the major parameters: max BlockSize of 5,000, min ModuleSize of 40, deep Split of 4, and merge CutHeight of 0.25. 4) Built a hierarchical clustering dendrogram of gene expression data for each dataset and identified the shared functional modules.

### Calculation of the Correlation and Hub Gene Identification

In order to determine the correlation between gene expression modules and clinical traits, the age and smoking history (smoking time, pack/years) of patients with lung cancer were chosen and analyzed. As for the hub genes, Cytoscape software was used for constructing the scale-free WGCNA for selected modules (Shannon et al., 2003). The cytoHubba package from Cytoscape was performed to extract the top 20 hub genes selected by 12 different algorithms, and mutual hub genes were then chosen by comparison of the top 20 hub genes. In order to select gene modules, the pathway functional enrichment analyses, including the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG), were performed by the Database for Annotation, Visualization and Integrated Discovery (DAVID). These gene functions were analyzed at the molecular level.

## RESULTS

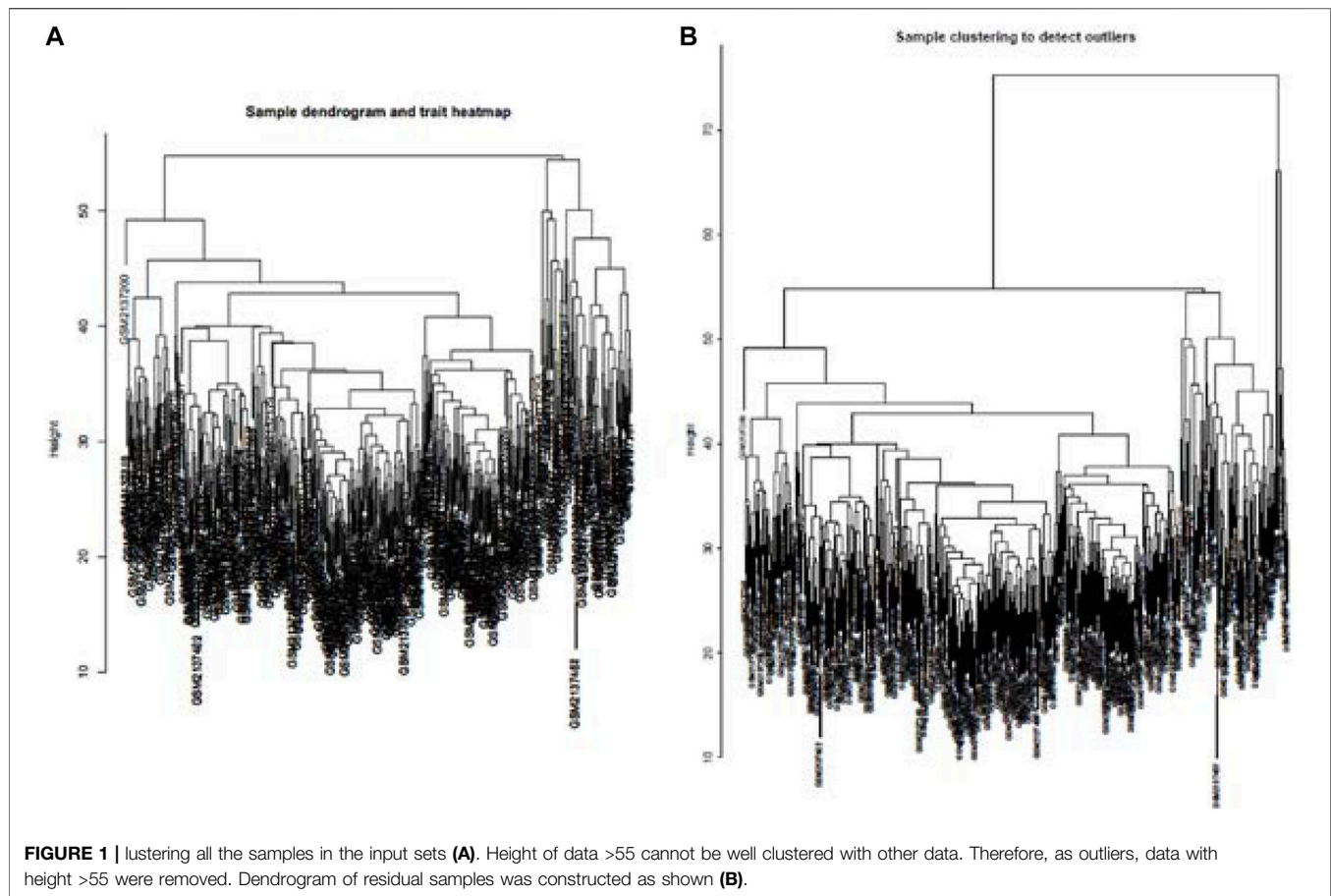
### Screening of DEGs

Top 3,600 DEGs in the order of  $|\log FC|$  were identified in the samples of early lung cancer by comparing with those of benign pulmonary nodule; there were 1,745 upregulated genes, and 1,855 downregulated genes.

### Construction of Co-Expression Module of Lung Cancer

Cluster analysis of DEGs is clearly shown in **Figure 1A**. Those samples were cut whose expression level was higher than 50 (**Figure 1B**). The soft threshold is the most important parameter. First, the soft threshold was selected (**Figure 2A**). When the power value was equal to 9, the degree of independence was up to 0.9, and the average connectivity was high. Five different gene co-expression modules were identified and displayed in different colors (**Figure 2B**). The gray module contained all the modules that could not be allocated to other modules, and the interaction of the co-expression modules showed that the thermograph depicted the topological overlap matrix of all genes. By constructing the TOM, these genes in the blue module had the highest correlation (**Figures 3A,B**).





## Analysis of WGCNA Network

Consensus relationships of consensus module eigengenes and clinical traits were presented as weak mutual correlations ( $p > 0.05$ ), while the consensus module eigengenes and clinical traits showed significant correlations ( $p < 0.05$ ) in the male and female data, respectively, which verified the conclusion of heterogeneity related with gender. There were scatter plots of GS and MM of blue and brown module genes, which had the highest correlation in the blue module (Figure 4A). The module feature relationship is displayed in Figure 4B. Their clinical features included age, smoking years, tumor size, and lung cancer status. Cluster analysis showed that the blue module was significantly correlated with the clinical characteristics of lung cancer.

## Gene Co-Expression and Hub Genes

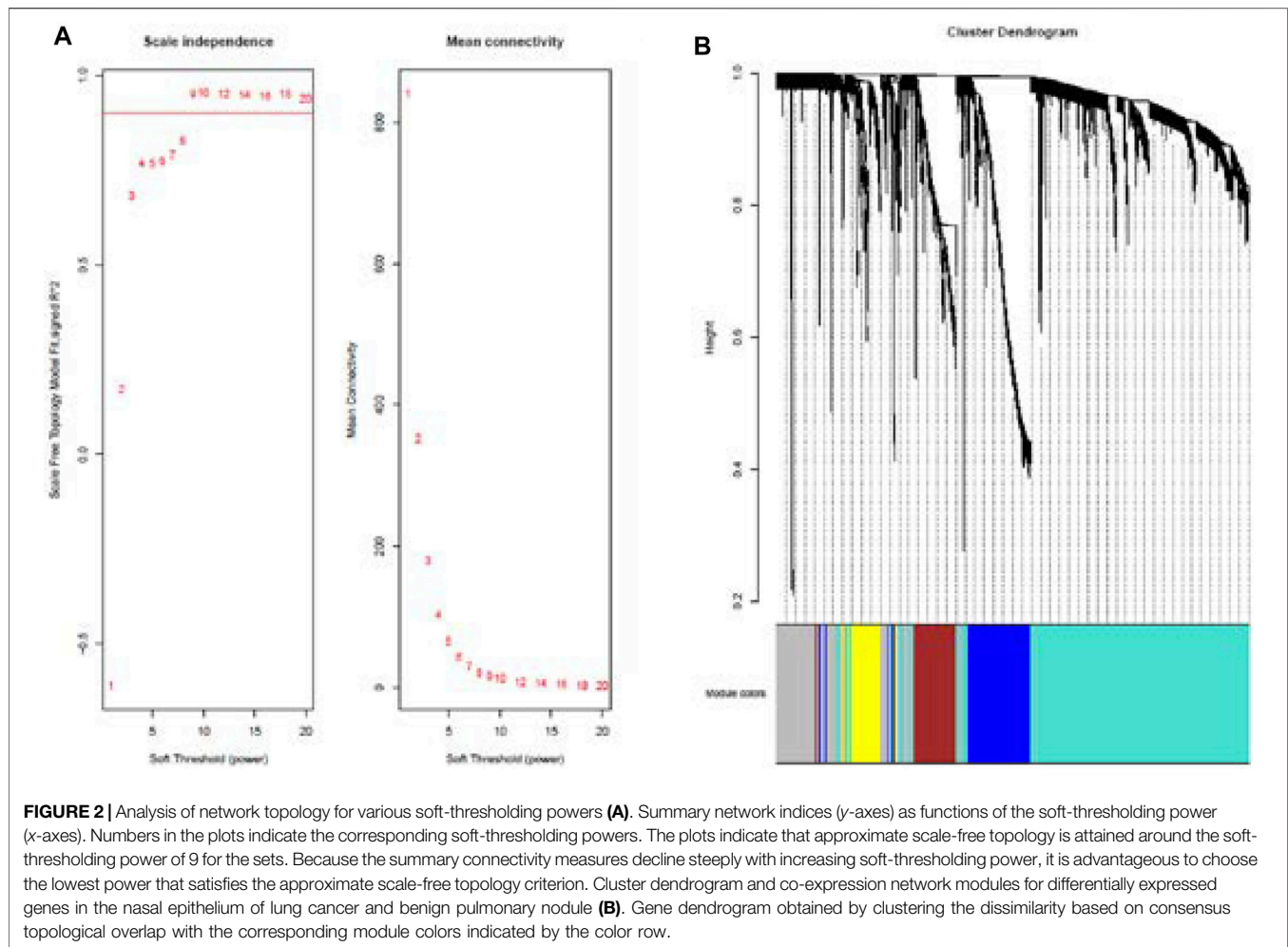
In these genes, four specific modules of lung cancer were constructed as blue, brown, yellow, and light blue modules, and the blue and brown modules were strongly linked to genetic connectivity. Twelve algorithms of the cytoHubba package were used to calculate the hub genes and their connectivity in Cytoscape software. In the brown module, the hub genes identified were TLR8, HCK, NCF1, EMR3, CSF2RB, and DYSF; in the blue module, the pivotal genes identified were HYDIN, SPEF2, ANKFN1, DNAH5, C12orf55, and CCDC113 (Tables 1, 2). In every network, the color depth is directly proportional to its connectivity. Four specific modules associated

with clinical features of lung cancer were constructed, including blue, brown, yellow, and light blue, of which blue or brown module showed strong connection to genetic connectivity (Figures 5A,B). TLR8, HCK, NCF1, EMR3, CSF2RB, and DYSF were the hub genes identified from the brown module, and HYDIN, SPEF2, ANKFN1, DNAH5, C12orf55, and CCDC113 were the pivotal genes identified from the blue module.

## DISCUSSION

### Main Goal for This Study

The aim of this study was to find the candidate genes by WGCNA. It could provide insights into the biology of early lung cancer and find the diagnostic biomarker by detecting the gene expression of nasal epithelia, which could make up for the shortage in postoperative pathological diagnosis and guide the clinical therapy. WGCNA has been used to not only construct gene networks and detect modules but also identify hub genes and select significant genes as biomarkers based on gene correlations. Module detection in WGCNA is used as a knowledge-independent process. However, empirical judgment and functional annotation would be more accurate, followed by the selection of a threshold for culling the network (Letovsky, 1987; Liu et al., 2015). WGCNA is considered a better prediction for



hub genes when it comes to the biological process than the regression statistical methods. Therefore, the construction of mutants will also help to detect the hub genes for prediction of lung cancer and to understand the role of specific genes in pathogenesis, which was overlooked in early lung cancer (Subudhi et al., 2015).

## Technology and Method of WGCNA

WGCNA was applied to investigate 3,600 genes downloaded from a dataset at NCBI. First, the data were performed to obtain the gene expression consensus modules of nasal epithelia, module eigengenes, clinical traits, and their relationships. Second, we constructed the status-specific modules of lung cancer. Third, we identified the hub genes in brown and blue modules through cytoHubba in Cytoscape and detected the related genes in 12 algorithms. Lastly, we performed the gene enrichment analysis on GO and pathway terms.

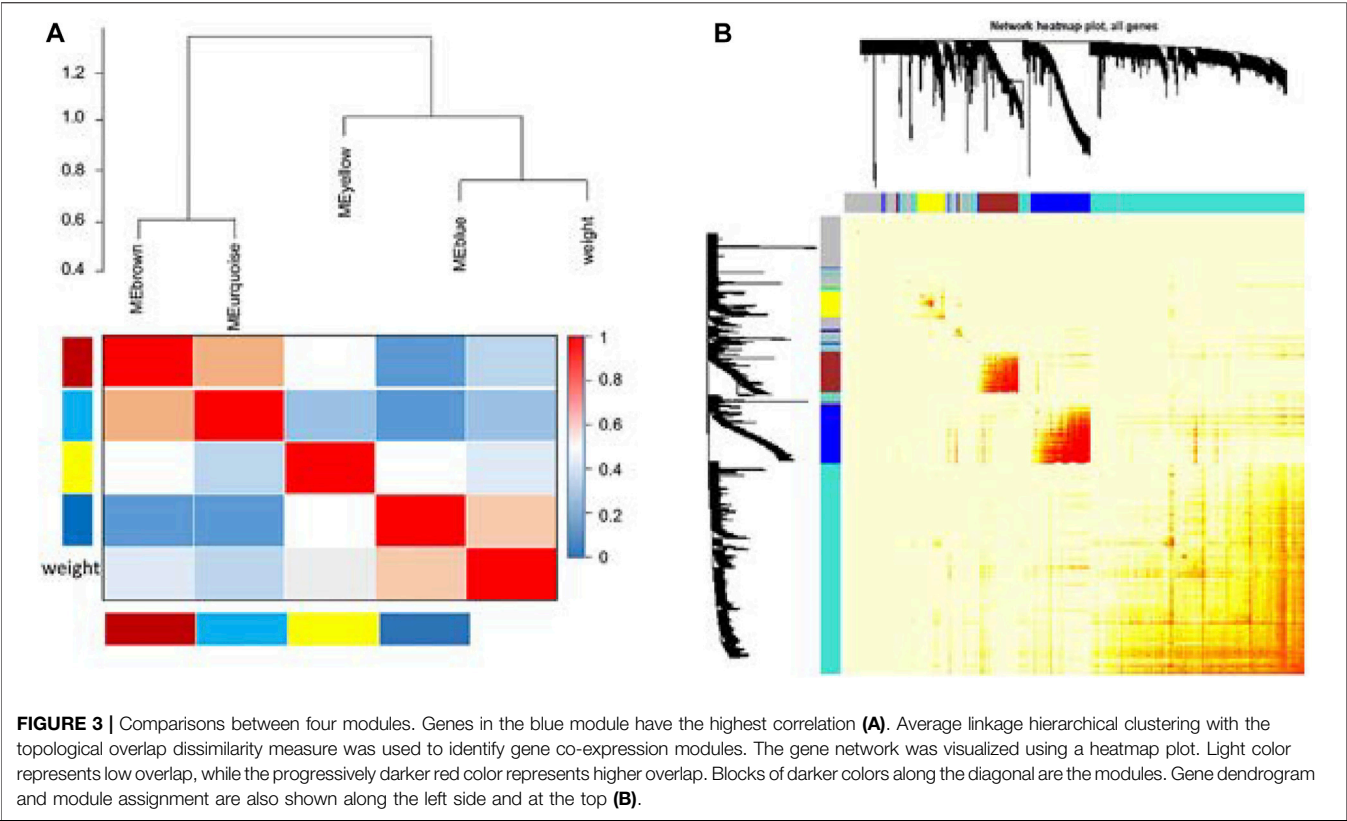
## New Results

WGCNA was used to investigate 3,600 genes downloaded from a dataset at NCBI. We obtained evidence about the changes of

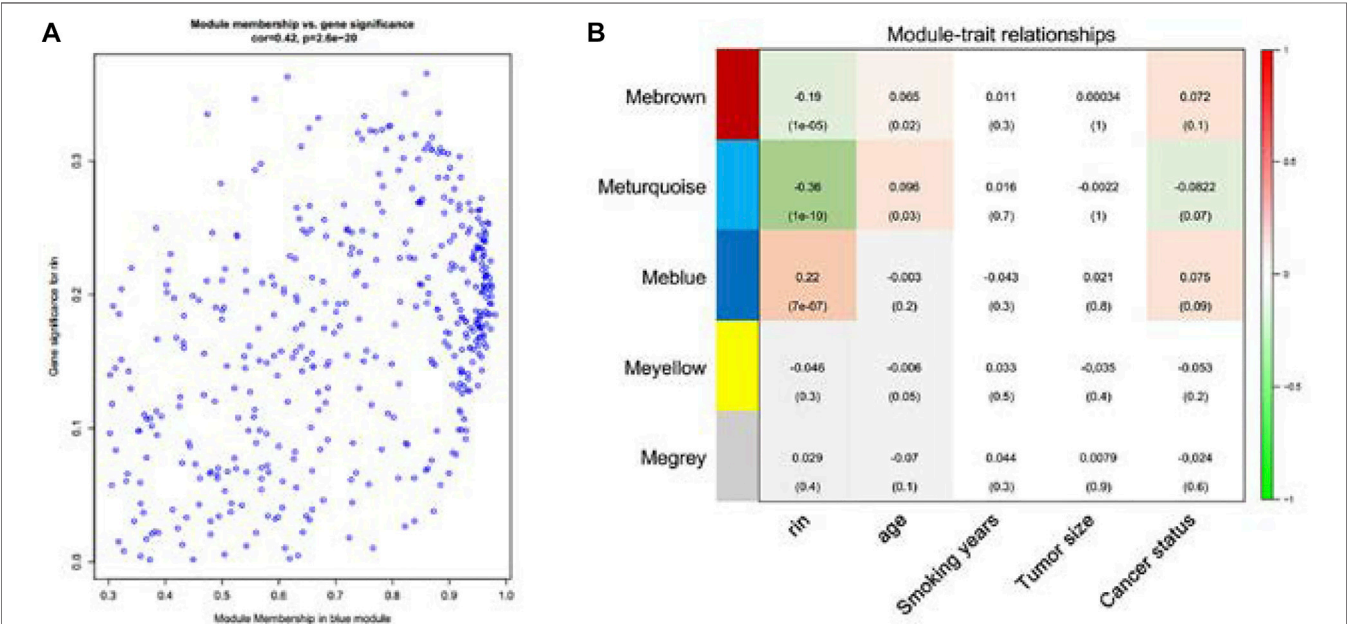
the hub gene expression in the feature gene module. The expressions of EMR3, NCF1, CSF2RB, DYSF, TLR8, and HCK in the lung cancer group were significantly different from those of the control group. The most significant difference in gene expression is EMR3, followed by NCF1, CSF2RB, DYSF, TLR8, and HCK.

## About EMR3

EMR3 is one of the members of the epidermal growth factor 7 transmembrane protein family (EGF-TM7), which includes CD97, EMR1, EMR2, and EMR4 and is expressed in the immune system cells. Until now, its functions are unclear yet, as well as the ligand and downstream signal (Stacey et al., 2001). Some research studies found that EMR3 is mainly expressed in mature granulocytes, and other members from the EGF-TM7 family may mediate the cell migration and leukocyte migration (Matmati et al., 2007; Yona et al., 2008a; Yona et al., 2008b). Ari and Kane found that EMR3 is expressed in glioblastoma cells and can mediate cell migration and invasion. It has the highest level of neutrophils, monocytes, and macrophages in the peripheral blood of Crohn's patients (Kane et al., 2010).



**FIGURE 3 |** Comparisons between four modules. Genes in the blue module have the highest correlation **(A)**. Average linkage hierarchical clustering with the topological overlap dissimilarity measure was used to identify gene co-expression modules. The gene network was visualized using a heatmap plot. Light color represents low overlap, while the progressively darker red color represents higher overlap. Blocks of darker colors along the diagonal are the modules. Gene dendrogram and module assignment are also shown along the left side and at the top **(B)**.



**FIGURE 4 |** Scatterplot of gene significance (GS) for lung cancer status vs. module membership (MM) in the blue module **(A)**. There is a highly significant correlation between GS and MM in the blue module. Correlation of gene co-expression modules with clinical traits in the training cohort ( $n = 196$ ). Blue module is strongly correlated with early lung cancer, so the blue module was chosen to be further analyzed **(B)**.

**TABLE 1 |** Top 25 hub genes of the blue module through dataset 1.

P2	MCC3	DMNC4	MNC5	Degree	EPC6	BottleNeck
1	C12orf55	IL36G	—	—	—	DNAH6
2	ZNF487	KRT13	DNAH6	DNAH6	DNAH6	—
3	SPEF2	ACSS3	HYDIN	HYDIN	HYDIN	MAP3K19
4	EFCAB2	HSPB8	DNAH12	DNAH12	DNAH12	DNAH7
5	DNAH7	SPRR1B	C12orf55	C12orf55	C12orf55	APOBEC4
6	ANKFN1	CYP2G2P	LOC100652824	LOC100652824	LOC100652824	CCDC113
7	NEK5	ABCB11	DYNC2H1	DYNC2H1	ULK4	WDR49
8	MDH1B	RNU6-646P	ULK4	ULK4	DYNC2H1	HYDIN
9	LOC100652824	CEACAM6	SPEF2	SPEF2	NEK5	RUVBL1
10	ROPN1L	DCAKD	DNAH7	EFCAB1	EFCAB1	C7orf63
11	ADGB	SYTL5	ADGB	NEK5	SPEF2	SNORD116-1
12	ALS2CR12	HTR3A	WDR49	DNAH7	AK9	SNORD116-29
13	TMEM107	RNU2-50P	NEK5	WDR49	ATXN7L1	IFT88
14	CHDC2	CPA4	EFCAB1	ADGB	MNS1	C12orf55
15	DUSP5	RNU6-490P	WDR96	WDR96	RSPH4A	SNORD116-24
16	MNS1	DSG3	IQUB	MNS1	CC2D2A	TMEM231
17	TMEM232	KCNJ16	STK33	CASC1	MAP3K19	ARMC2
18	EFHB	PDLIM2	CCDC30	WDR65	TCTEX1D1	SNORA20
19	LRRIQ1	CCDC34	WDR65	CCDC30	WDR96	SNORAD116-15
20	STOML3	FABP5	CASC1	IQUB	IQUB	SNORAD115-32
21	ARMC2	SOX2	MNS1	ANKFN1	WDR65	IQCK
22	PCDP1	RNU6-955P	SPAG17	ATXN7L1	ANKFN1	DNAH2
23	AGBL2	CNTNAP3B	ANKFN1	EFCAB2	NEK10	NEK5
24	MUC15	—	CCDC113	WDR63	DNAH7	TCTEX1D1
25	IQUB	KRT6B	NEK10	CCDC39	WDR49	NME5

P	EcCentricity	Closeness	Radiality	Node_name	Stress	CC7
1	SNORD116-29	—	—	—	—	IL36G
2	SCARNA7	DNAH6	DNAH6	DNAH6	DNAH6	KRT13
3	SNORA20	HYDIN	HYDIN	HYDIN	HYDIN	ACSS3
4	SNORD116-15	DNAH12	DNAH12	DNAH12	DNAH12	DSG3
5	SNORD116-24	C12orf55	C12orf55	SNORD116-24	C12orf55	HSPB8
6	SNORD116-5	LOC100652824	LOC100652824	CCDC113	CCDC113	CEACAM6
7	SNORD116-25	DYNC2H1	DYNC2H1	DYNC2H1	ARMC2	SPRR1B
8	SNORD116-1	ULK4	ULK4	C12orf55	DYNC2H1	SYTL5
9	SNORD116-26	SPEF2	SPEF2	ARMC2	RPGRIP1L	RNU6-646P
10	RUVBL1	DNAH7	DNAH7	LOC100652824	LOC100652824	ABCB11
11	DNAH7	ADGB	WDR49	SNORD116-29	SNORD116-24	CYP2G2P
12	C14orf142	WDR49	NEK5	SNORA20	TCTEX1D1	HTR3A
13	CCDC113	NEK5	ADGB	SCARNA7	WDR65	SNORA28
14	NUCB2	WDR96	WDR96	WDR65	SNORA20	DCAKD
15	TCTEX1D1	EFCAB1	STK33	RPGRIP1L	SNORD116-29	RNU6-490P
16	ZMAT1	STK33	IQUB	ULK4	ZBBX	HPX
17	DTHD1	IQUB	EFCAB1	TCTEX1D1	SPATA18	KRT6B
18	FANK1	WDR65	WDR65	MAATS1	SCARNA7	SOX2
19	CCDC60	CCDC30	CCDC113	TMEM232	MAATS1	AZGP1
20	PACRG	CASC1	CASC1	ZBBX	ULK4	KCNJ16
21	SPEF2	MNS1	CCDC30	C9orf116	SNORD116-14	RCBTB1
22	VWA3A	SPAG17	DTHD1	DTHD1	C9orf116	KRT24
23	KIAA1377	ANKFN1	ANKFN1	SNORD116-5	SNORD116-1	ADIRF
24	ARMC2	CCDC113	ARMC2	ADGB	SNORD116-15	SNORD116-29
25	KIAA1841	DTHD1	MNS1	SPATA18	ADGB	LOC100131860

Notes: The hub gene was calculated by cytoHubba 1; parameters 2; maximal clique centrality 3; density of maximum neighborhood component 4; maximum neighborhood component 5; edge percolated component 6; clustering coefficient 7.

## About NCF1

NCF1 is a major component of the nicotinamide adenine dinucleotide oxidase system; it can regulate the production of reactive oxygen species (ROS). NCF1 deficiency will lead to the reduction of ROS, which is associated with immune disorders (Bastos et al., 1995). NCF1-knock-out mice have increased

leukocyte infiltration and morphological changes in the colonic mucosa, indicating that the absence of the NCF1 gene could aggravate colitis (El Naschie, 2004). In contrast, the upregulation of NCF1 gene expression might cause diminished or deficient ROS production that is detrimental to human health.



**TABLE 2 |** Top 25 hub genes of the brown module through dataset 1.

P2	MCC3	DMNC4	MNC5	Degree	EPC6	BottleNeck
1	HCK	TPD52L2	—	—	—	—
2	NCF1	FCGR3A	HCK	FCGR3A	FCGR3A	DYSF
3	PLEK	NCF2	NCF1	NCF1	HCK	LILRB2
4	TLR8	—	FCGR3A	HCK	NCF1	SLED1
5	ITGAX	MIR23A	CSF2RB	GLT1D1	CSF2RB	PREX1
6	APBB1IP	NFIL3	GLT1D1	CSF2RB	PLEK	TAGAP
7	EMR2	CD14	EMR2	EMR2	FCGR1A	LOC254896
8	CSF2RB	FCGR1A	FPR1	FCGR1A	FPR1	EMR3
9	MNDA	MIR223	TAGAP	TAGAP	GLT1D1	SH2D3C
10	CXCR4	ITGAX	EMR3	EMR3	MNDA	—
11	THEMIS2	PLEK	MNDA	FPR1	TAGAP	FOSB
12	CD53	NFE2	GPR97	GPR97	LCP2	ZFP36
13	SPI1	SIRPB1	CSF3R	MNDA	EMR3	NFAM1
14	SLA	LINC00921	DYSF	LCP2	THEMIS2	PPP1R18
15	TYROBP	THEMIS2	TLR8	PLEK	ITGAX	PTPRC
16	FFAR2	P2RY13	APBB1IP	CSF3R	DYSF	CD14
17	EMR3	EVI2B	LCP2	DYSF	EMR2	TRIB1
18	FCGR1A	ARRB2	PLEK	BCL2A1	BCL2A1	IL1B
19	GPR97	TREM1	SLA	ITGAX	CSF3R	MNDA
20	FMNL1	TNFAIP6	FCGR1A	SLA	FCGR2A	FMNL1
21	RASSF2	PLXNC1	ITGAX	TLR8	SLA	RGS2
22	LILRB3	CHST11	LILRB3	FCGR2A	TLR8	LPCAT1
23	HCAR3	SELPLG	BCL2A1	APBB1IP	TREM1	FYB
24	SLC11A1	NABP1	RASSF2	LILRB3	LILRB3	FFAR2
25	AQP9	SELL	FCGR2A	RASSF2	RASSF2	CHSY1
P	EcCentricity	Closeness	Radiality	Node_name	Stress	CC7
1	CSRNP1	—	—	—	—	TPD52L2
2	HRH2	HCK	HCK	HCK	HCK	GZMB
3	OSM	NCF1	NCF1	SH2D3C	SH2D3C	HEY2
4	CLEC4E	FCGR2A	FCGR2A	NCF1	NCF1	MIR23A
5	ARID5A	CSF2RB	CSF2RB	CSF2RB	EMR2	LINC00921
6	CASS4	GLT1D1	EMR2	EMR2	EMR3	CD14
7	PLEKH02	EMR2	GLT1D1	EMR3	FCGR3A	NFIL3
8	LILRB3	FPR1	EMR3	LILRB2	CSF2RB	ARRB2
9	—	EMR3	FPR1	DYSF	ADAM8	RAB24
10	TLR8	MNDA	MNDA	GLT1D1	DYSF	NFE2
11	—	TAGAP	DYSF	ADAM8	GLT1D1	NCF2
12	SH2D3C	GPR97	TAGAP	CSF3R	CSF3R	RN7SKP78
13	PHOSPHO1	CSF3R	LCP2	TLR8	TLR8	GMIP
14	—	DYSF	CSF3R	TAGAP	LCP2	MIR223
15	CEBPD	LCP2	ITGAX	LCP2	SLA	EDN1
16	TPD52L2	TLR8	LILRB3	APBB1IP	LILRB2	CHST11
17	RAB24	ITGAX	GPR97	FCGR3A	WAS	P2RY13
18	RN7SKP78	PLEK	TLR8	WAS	TAGAP	NABP1
19	EGR2	FCGR1A	PLEK	SLA	ITGAX	SIRPB1
20	CNN2	SLA	FCGR1A	FPR1	—	ZC3H12A
21	FOSB	LILRB3	SLA	GPR97	FPR1	TNFAIP6
22	RCSD1	APBB1IP	APBB1IP	LILRB3	CYTH4	AOAH
23	TNFAIP6	BCL2A1	PROK2	DGAT2	APBB1IP	EVI2B
24	DYSF	PROK2	RASSF2	ALOX5	MOB3A	PRKCB
25	LILRB2	RASSF2	FCGR2A	PTPRE	LCP1	ZFP36

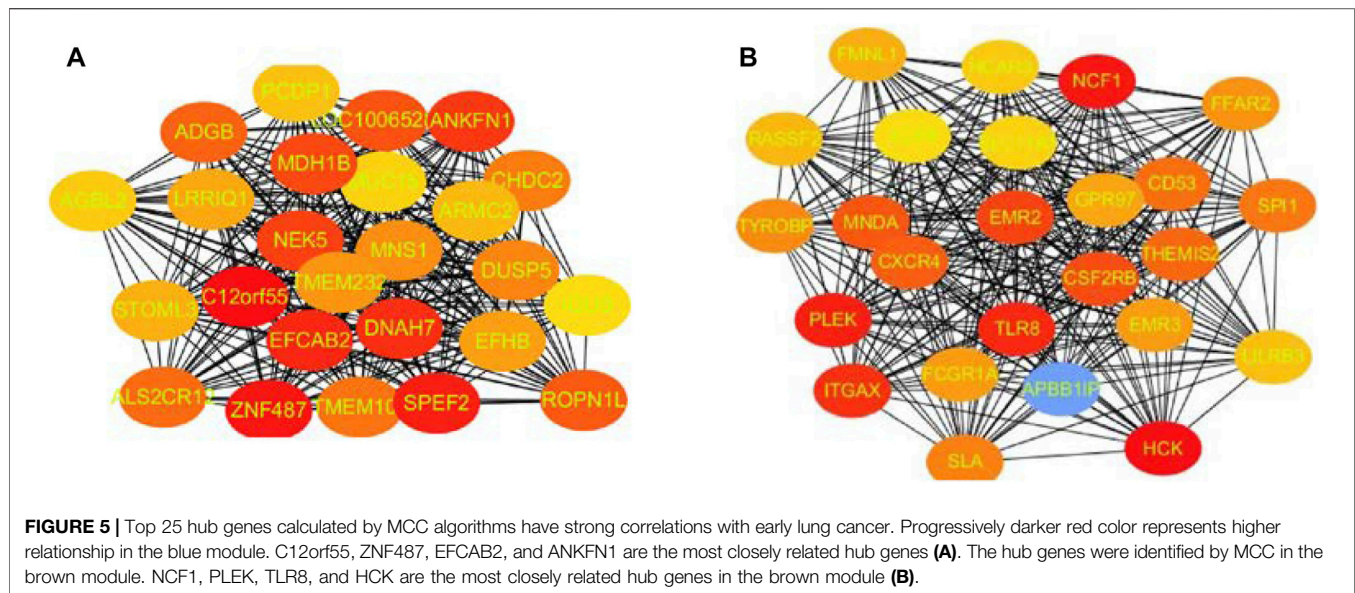
Notes: The hub gene was calculated by cytoHubba 1; parameters 2; maximal clique centrality 3; density of maximum neighborhood component 4; maximum neighborhood component 5; edge percolated component 6; clustering coefficient 7.

## About CSF2RB and Others

CSF2RB is the common beta chain of the high-affinity receptor complexes for ligands of IL-3, IL-5E, and CSF. Research studies found that mutation of CSF2RA or CSF2RB can cause hereditary pulmonary alveolar proteinosis (PAP) (Takaki et al., 2016), and CSF2RB is a risk factor for schizophrenia

and depression in the Han population of Chinese and a potential oncogene that can be targeted by several miRNAs for undergoing cell apoptosis (Chen et al., 2011). The DYSF gene is a 220-kD protein, which plays a major role in the regulation of plasma membrane repair. Fusion of DYSF with the ALK gene has been found to be associated with advanced lung cancer. As a





single-stranded RNA sensor, the activation of TLR8 can also promote the survival and chemoresistance of lung cancer cells. The HCK gene belongs to the Src family of tyrosine kinase, which is mainly involved in the regulation of polymorphonuclear leukocytes. A recent study showed that in the Bai nationality of China, the polymorphism of the introns of the HCK gene is associated with lung function and airway abnormality (Espinoza-Fonseca, 2016).

## Experimental Verification

Studies have proved that the existence of injury in the bronchial airway results in gene expression alterations in patients with lung cancer, and the airway epithelial injury associated with lung cancer extends to the nasal epithelium (Shannon et al., 2003). In the previous study, the downregulated genes CASP10 and CD177 and the upregulated genes BAK1, ST14, CD82, and MUC4 were detected as biomarkers for lung cancer by the joint sparse regression model (Loxham and Davies, 2017). Our study has detected some hub genes from gene expression of the nasal epithelium of early lung cancer by WGCNA. The most significant difference in gene expression was shown by EMR3, followed by NCF1, CSF2RB, DYSF, and so on. The results of qRT-PCR are in accordance with those of microarray analysis (Qureshi, 2018).

## Clinical Application

This study may provide an additional proof for detecting early lung cancer by observing gene expression of the nasal epithelium, which indicates a great potential for clinical application (Lobato

and O'Sullivan, 2018). The biomarker of nasal epithelium would be used as a reference for patients with small nodules at low risk of malignancy, which can be managed by CT screening (Petty, 2001; Cottin and Cordier, 2016). However, this study still has some limitations. It lacks further studies on the relationship between gene expression and pathological typing of lung cancer (Tang et al., 2018), so large-scale samples must be collected to have a better analysis in the future.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Materials; 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.

## ACKNOWLEDGMENTS

We sincerely thank for the support of personal research fund from Zhongshan Hospital of Xiamen University and the technical support from Xiamen JiKe Biotechnology Company Limited.

## REFERENCES

Asaad Zebari, N., and Emin Tenekeci, M. (2022). Support System Based Computer-Aided Detection for Skin Cancer: A Review. *Fusion Pract. Appl.* 7 (1), 30–40. doi:10.54216/FPA.070103

Bastos, J., Steyaert, M., Roovers, R., Kinget, P., Sansen, W., Graindourze, B., et al (1995). Mismatch Characterization of Small Size MOS Transistors. *Proc. ICMTS* 8, 271–276.

Cannioto, R., Etter, J. L., LaMonte, M. J., Ray, A. D., Joseph, J. M., Qassim, E. A., et al (2018). Lifetime Physical Inactivity is Associated with Lung Cancer Risk and Mortality. *Cancer Treat. Res. Commun.* 14, 37–45. doi:10.1016/j.ctarc.2018.01.001

- Chen, P., Huang, K., Zhou, G., Zeng, Z., Wang, T., Li, B., et al (2011). Common SNPs in CSF2RB are Associated with Major Depression and Schizophrenia in the Chinese Han Population. *World J. Biol. Psychiatry* 12, 233–238. doi:10.3109/15622975.2010.544328
- Cottin, V., and Cordier, J-F. (2016). *Eosinophilic Lungs Disease. Murray and Nadel's Textbook of Respiratory Medicine*. 6th Edition. Philadelphia: Elsevier Saunders.
- El Naschie, M. S. (2004). Small World Network, Topology and the Mass Spectrum of High Energy Particles Physics. *Chaos Solit. Fractals* 19, 689–697. doi:10.1016/s0960-0779(03)00337-0
- Espinoza-Fonseca, L. M. (2016). Pathogenic Mutation R959W Alters Recognition Dynamics of Dysferlin Inner DysF Domain. *Mol. BioSyst.* 12, 973–981. doi:10.1039/c5mb00772k
- Gupta, S., Siddiqui, S., Haldar, P., Raj, J. V., Entwisle, J. J., Wardlaw, A. J., et al (2009). Qualitative Analysis of High-Resolution CT Scans in Severe Asthma. *Chest* 136, 1521–1528. doi:10.1378/chest.09-0174
- Kane, A. J., Sughrue, M. E., Rutkowski, M. J., Phillips, J. J., and Parsa, A. T. (2010). EMR-3: A Potential Mediator of Invasive Phenotypic Variation in Glioblastoma and Novel Therapeutic Target. *Neuroreport* 21, 1018–1022. doi:10.1097/wnr.0b013e32833f19f2
- Khan, K. A., Nardelli, P., Jaeger, A., O'Shea, C., Cantillon-Murphy, P., and Kennedy, M. P. (2016). Navigational Bronchoscopy for Early Lung Cancer: A Road to Therapy. *Adv. Ther.* 33, 580–596. doi:10.1007/s12325-016-0319-4
- Langfelder, P., and Horvath, S. (2008). WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinforma.* 9, 559. doi:10.1186/1471-2105-9-559
- Letovsky, S. (1987). Cognitive Processes in Program Comprehension. *J. Syst. Softw.* 7, 325–339. doi:10.1016/0164-1212(87)90032-x
- Liu, R., Cheng, Y., Yu, J., Lv, Q.-L., and Zhou, H.-H. (2015). Identification and Validation of Gene Module Associated with Lung Cancer through Coexpression Network Analysis. *Gene* 563, 56–62. doi:10.1016/j.gene.2015.03.008
- Lobato, I. M., and O'Sullivan, C. K. (2018). Recombinase Polymerase Amplification: Basics, Applications and Recent Advances. *TrAC Trends Anal. Chem.* 98, 19–35. doi:10.1016/j.trac.2017.10.015
- Loxham, M., and Davies, D. E. (2017). Phenotypic and Genetic Aspects of Epithelial Barrier Function in Asthmatic Patients. *J. Allergy Clin. Immunol.* 139, 1736–1751. doi:10.1016/j.jaci.2017.04.005
- Matmati, M., Pouwels, W., van Bruggen, R., Jansen, M., Hoek, R. M., Verhoeven, A. J., et al (2007). The Human EGF-TM7 Receptor EMR3 is a Marker for Mature Granulocytes. *J. Leukoc. Biol.* 81, 440–448. doi:10.1189/jlb.0406276
- McCall, M. N., Bolstad, B. M., and Irizarry, R. A. (2010). Frozen Robust Multiarray Analysis (fRMA). *Biostatistics* 11, 242–253. doi:10.1093/biostatistics/kxp059
- Ning, B., Xu, D. L., Gao, J. H., Wang, L. L., Yan, S. Y., and Cheng, S. (2016). Identification of Pathway-Related Modules in High-Grade Osteosarcoma Based on Topological Centrality of Network Strategy. *Eur. Rev. Med. Pharmacol. Sci.* 20, 2209–2220.
- Petty, T. L. (2001). The Early Diagnosis of Lung Cancer. *Dis. Mon.* 47, 204–264. doi:10.1067/mcd.2001.116285
- Pisani, P., Parkin, D. M., Bray, F., and Ferlay, J. (1999). Estimates of the Worldwide Mortality from 25 Cancers in 1990. *Int. J. Cancer* 83, 18–29. doi:10.1002/(sici)1097-0215(19990924)83:1<18::aid-ijc5>3.0.co;2-m
- Qureshi, N. (2018). *Identification of Significantly Different Modules between Gene Expression in Nasal Epithelial Cell and Lung Cancer by WGCNA Study and Experimental Verification (D)*. Xiamen CHINA: Xiamen University.
- Shakeel, P. M., Tolba, A., Al-Makhadmeh, Z., and Jaber, M. M. (2020). Automatic Detection of Lung Cancer from Biomedical Data Set Using Discrete AdaBoost Optimized Ensemble Learning Generalized Neural Networks. *Neural Comput. Applic.* 32, 777–790. doi:10.1007/s00521-018-03972-2
- 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
- Stacey, M., Lin, H.-H., Hilyard, K. L., Gordon, S., and McKnight, A. J. (2001). Human Epidermal Growth Factor (EGF) Module-Containing Mucin-Like Hormone Receptor 3 is a New Member of the EGF-TM7 Family that Recognizes a Ligand on Human Macrophages and Activated Neutrophils. *J. Biol. Chem.* 276, 18863–18870. doi:10.1074/jbc.m101147200
- Subudhi, A. K., Boopathi, P. A., Pandey, I., Kaur, R., Middha, S., Acharya, J., et al (2015). Disease Specific Modules and Hub Genes for Intervention Strategies: A Co-Expression Network Based Approach for Plasmodium Falciparum Clinical Isolates. *Infect. Genet. Evol.* 35, 96–108. doi:10.1016/j.meegid.2015.08.007
- Sun, Q., Zhao, H., Zhang, C., Hu, T., Wu, J., Lin, X., et al (2017). Gene Co-Expression Network Reveals Shared Modules Predictive of Stage and Grade in Serous Ovarian Cancers. *Oncotarget* 8, 42983–42996. doi:10.18632/oncotarget.17785
- Takaki, M., Tanaka, T., Komohara, Y., Tsuchihashi, Y., Mori, D., Hayashi, K., et al (2016). Recurrence of Pulmonary Alveolar Proteinosis after Bilateral Lung Transplantation in a Patient with a Nonsense Mutation in CSF2RB. *Respir. Med. Case Rep.* 19, 89–93. doi:10.1016/j.rmcr.2016.06.011
- Tang, Q., Zhang, H., Kong, M., Mao, X., and Cao, X. (2018). Hub Genes and Key Pathways of Non-Small Lung Cancer Identified Using Bioinformatics. *Oncol. Lett.* 16, 2344–2354. doi:10.3892/ol.2018.8882
- Team, A. S. (2017). Shared Gene Expression Alterations in Nasal and Bronchial Epithelium for Lung Cancer Detection. *J. Natl. Cancer Inst.* 109, djw327. doi:10.1093/jnci/djw327
- Timmins, M., and Ashlock, D. (2017). Network Induction for Epidemic Profiles with a Node Representation. *Biosystems* 162, 205–214. doi:10.1016/j.biosystems.2017.10.013
- Yona, S., Lin, H.-H., Siu, W. O., Gordon, S., and Stacey, M. (2008). Adhesion-GPCRs: Emerging Roles for Novel Receptors. *Trends Biochem. Sci.* 33, 491–500. doi:10.1016/j.tibs.2008.07.005
- Yona, S., Lin, H. H., Dri, P., Davies, J. Q., Hayhoe, R. P. G., Lewis, S. M., et al (2008). Ligation of the adhesion-GPCR EMR2 Regulates Human Neutrophil Function. *FASEB J.* 22, 741–751. doi:10.1096/fj.07-9435com

**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.

Copyright © 2022 Qureshi, Chi, Qian, Huang and Duan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



## OPEN ACCESS

## EDITED BY

Deepak Kumar Jain,  
Chongqing University of Posts and  
Telecommunications, China

## REVIEWED BY

Defei Liu,  
Southwest University, China  
Haiyang Yu,  
Dalian Ocean University, China  
Yujie Zhou,  
Baotou Teachers' College, China

## \*CORRESPONDENCE

Yufeng Yang,  
1764200103@e.gzh.edu.cn

## SPECIALTY SECTION

This article was submitted to Human  
and Medical Genomics,  
a section of the journal  
Frontiers in Genetics

RECEIVED 30 May 2022

ACCEPTED 06 July 2022

PUBLISHED 10 August 2022

## CITATION

Ni S, Chen F, Chen G and Yang Y (2022),  
Mathematical model and genomics  
construction of developmental biology  
patterns using digital image technology.  
*Front. Genet.* 13:956415.  
doi: 10.3389/fgene.2022.956415

## COPYRIGHT

© 2022 Ni, Chen, Chen and Yang. This is  
an open-access article distributed  
under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#).  
The use, distribution or reproduction in  
other forums is permitted, provided the  
original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which does  
not comply with these terms.

# Mathematical model and genomics construction of developmental biology patterns using digital image technology

Shiwei Ni<sup>1</sup>, Fei Chen<sup>1</sup>, Guolong Chen<sup>2</sup> and Yufeng Yang<sup>1\*</sup>

<sup>1</sup>Institute of Life Sciences, FuZhou University, FuZhou, Fujian, China, <sup>2</sup>School of Mathematics and  
Statistics, FuZhou University, FuZhou, Fujian, China

Biological pattern formation ensures that tissues and organs develop in the correct place and orientation within the body. A great deal has been learned about cell and tissue staining techniques, and today's microscopes can capture digital images. A light microscope is an essential tool in biology and medicine. Analyzing the generated images will involve the creation of unique analytical techniques. Digital images of the material before and after deformation can be compared to assess how much strain and displacement the material responds. Furthermore, this article proposes Development Biology Patterns using Digital Image Technology (DBP-DIT) to cell image data in 2D, 3D, and time sequences. Engineered materials with high stiffness may now be characterized via digital image correlation. The proposed method of analyzing the mechanical characteristics of skin under various situations, such as one direction of stress and temperatures in the hundreds of degrees Celsius, is achievable using digital image correlation. A DBP-DIT approach to biological tissue modeling is based on digital image correlation (DIC) measurements to forecast the displacement field under unknown loading scenarios without presupposing a particular constitutive model form or owning knowledge of the material microstructure. A data-driven approach to modeling biological materials can be more successful than classical constitutive modeling if adequate data coverage and advice from partial physics constraints are available. The proposed procedures include a wide range of biological objectives, experimental designs, and laboratory preferences. The experimental results show that the proposed DBP-DIT achieves a high accuracy ratio of 99.3%, a sensitivity ratio of 98.7%, a specificity ratio of 98.6%, a probability index of 97.8%, a balanced classification ratio of 97.5%, and a low error rate of 38.6%.

## KEYWORDS

digital image correlation, biology patterns, data-driven, medicine, microscopes

## Summary of digital image technology

As digital imaging methods linked with light microscopy continue to develop exponentially, researchers in domains such as biology, medicine, and other sciences can generate vast image data in a wide range of exploration (Zhang et al., 2021). This potentially enormous amount of image data must be handled with care to enable the extraction of the necessary information in a timely and cost-effective manner (Emami et al., 2021). In this way, image analysis is not confined to analyzing the image that has been captured (Garreta et al., 2021). In many cases, it includes collaboration with people who gather pictures to decide on the best method to take while producing image data at the microscopes (An et al., 2021). It is usually preferable for image analysis and high-quality photos rather than attempting to make them suitable for later processing (Shao et al., 2021). It is critical to determine when it is appropriate to undertake a 2D study and when it is essential to expand the analysis to 3D (Jiao et al., 2021).

Images in 3D need more data and storage space. Still, they necessitate the development of new analysis techniques and more memory and processing capacity to manage the vast amounts of data that must be processed when the analysis is carried out (Blackiston et al., 2021). Over the last several years, professionals from biology and medicine and image analysis have collaborated to develop several significant research outcomes that have benefited both sides (Seyfferth et al., 2021; Tang et al., 2021). The study is considered an extension of the same undertaking—Digital Image Analysis of Cells—applications in 2D, 3D, and time. In recent years, great progress has been made in creating deep neural networks (NNs) to model heterogeneous materials (Nguyen et al., 2021).

These efforts include the neural operator learning approach, which aims to learn the mappings between dynamic system inputs and system states (Medialdea et al., 2021). The network may operate as a replacement for a solution operator in a dynamic system, and DBP-DIT is especially interesting (Heddeleston et al., 2021). While neural operators have several advantages over classical neural networks, their most notable advantage is their generalizability to different input instances, which results in a computing advantage in prediction efficiency (Pourasad et al., 2021). A forward pass of the network is all that is required to solve for a new instance of the input parameter after the neural operator is trained (Granwehr and Hofer, 2021). When it comes to simulating the unknown physics rule of homogeneous materials, neural operators have shown to be quite effective (Sarkar et al., 2021).

DBP-DIT has explored the practicality of learning a material model for a latex material directly from digital image correlation (DIC) data (Fan et al., 2021). The suggested technique has shown that the learned solution operators substantially outperform the traditional constitutive model (Caleb et al., 2021). The primary goal is to research and enhance the current image analysis for

new applications, including 3D applications where appropriate (Lürig et al., 2021). It has been explored if any new ways outperform the present ones. The image data utilized in this article depict cells from various investigations (Lewis et al., 2021). The images were taken utilizing various microscopy methods, both in 2D and in 3D, to get the desired results. In a time series of images, the passage of time adds a new depth to the images (Aljazeera et al., 2022).

The main contribution of this article is.

- The primary focus of this study, clinically and physiologically relevant traits, aims to establish a framework for and software tool for automatic identification and classification of microscopic biopsy pictures.
- An approach to biological tissue modeling based on data-driven workflow aims to predict the segmentation method based on digital image correlation (DIC) measurements under unseen loading scenarios or knowledge of the material microstructure.
- With its capacity to determine strain fields and translations down to the micron scale, DIC shows potential as a tool for examining other biological specimens in the laboratory.
- The mechanical reactions of a biological tissue specimen under various loading situations may be modeled using a neural operator learning approach.

The overall article structure follows: Section 1 explores the summary of digital image technology, Section 2 demonstrates the related works based on biology pattern recognition, Section 3 expresses the proposed methodology, and Section 4 depicts the results and discussion and finally concludes the article.

## Related works based on biology pattern recognition

As a potent method for examining cell states and activities at the single-cell level, single-cell RNA sequencing (scRNA-seq) has gained popularity in recent years (Li, 2021). As experimental platforms and bioinformatics methodologies have advanced rapidly over the last decade, scRNA-seq has become more affordable and viable for many medical facilities. Several cell and molecular pathways were involved in tissue formation, adult cell function, illness, and aging that had been studied using *Drosophila* as a model organism. Using scRNA-seq in *Drosophila* would address the obstacles and prospects of creating new findings.

Many characteristics, like number, distance, orientation, and location, were used to influence the functioning of protein networks in biological systems (Kong et al., 2021). It was possible to use DNA origami to create nanometer-precision scaffolding for protein assembly that could be controlled,



programmed, and addressed. Several multidisciplinary studies recently realized the accurate building of DNA origami-based protein networks and the developing use in various fields. Some argue that DNA origami-based protein networks have been employed in various applications in the biomedical and enzymatic areas.

Hepatocellular carcinoma (HCC) was one of the world's most deadly malignancies. Patient-specific medication screening for HCC was currently hampered by a shortage of reliable *in vitro* models (Xie et al., 2021). Three-dimensional bio-printed HCC (3DP-HCC) models taken from patient samples were effectively produced and developed well over lengthy periods. Using 3DP-HCC models, drug screening findings could be presented in an accessible and quantifiable manner. Finally, 3DP-HCC models were accurate *in vitro* models that had been dependable in long-term culture and could predict patient-specific medications for tailored therapy.

Digital image correlation (DIC) measures for assessing and improving additive manufacturing (AM) processes were reviewed in this study (Cunha et al., 2021). First, the DIC principle was revisited, and its application to various AM processes was discussed. An overview of the influence of *in situ* monitoring on AM processes was provided based on target themes such as defect characterization, residual stress assessment, geometric distortions, strain measurements, statistical model validation, and material characterization. An *in situ* measurement case study was provided for wire and arc additive manufacturing (WAAM), highlighting the prospects, problems, and solutions.

EA-LPME-SSHS-TAD was utilized to identify carbaryl in food samples using a digital image colorimetry approach (Jing et al., 2021). This method was used to extract 1-naphthol and separate it from the octanoic acid sample by altering the pH values of this solution. Tangerine compounds were created by combining the extracted solution MBDF connected to the TAD, which included 1-naphthol as one of its basic components.

Based on the above analysis, scRNA-seq, 3DP-HCC, DIC, and EA-LPME-SSHS-TAD, there are some issues such as low accuracy, sensitivity, and error rate. Therefore, this article proposed DBP-DIT digital image correlation (DIC) measurements and a data-driven approach to measuring biological materials.

## Development biology patterns using digital image technology (DBP-DIT)

Researchers may measure phenotypic changes in many cell populations using image-based cell profiling (Agboola, 2020). It provides the door for large-scale studies of biological systems by chemical and genetic manipulations. The typical approach for this technology is picture capture using high-throughput microscopy devices, followed by image processing methods.

The suggested procedures demonstrate how a sequence of microscope images may be used to build high-quality image-based profiles. Worldwide laboratories are using image-based cell profiling to seek biological discoveries, and the tactics developed are based on their experience (Zhao et al., 2019; Chinnadurai and Sindhu, 2020). The proposed offered encompass choices that may fit varied biological aims, experimental designs, and laboratory preferences.

Figure 1 expresses the proposed structure of DBP-DIT. In this diagram, there are five major functions such as 1) tissues and biology cells, 2) biological patterns, 3) DBP-DIT, 4) neural operator model and data-driven approach, and 5) datastore.

- i) Tissues and biology cells: tissue is a degree of biological order between cellular and a fully developed organ. A tissue is a group of lymphocytes and their matrix proteins from the same origin that work together to accomplish a specified task. Organs are then generated by the elements of the system together of many tissues. Multicellular creatures are organized into tissues, made up of physically and functionally identical cells and the intercellular material that connects them.
- ii) Biological patterns: diverse processes lead to the emergence of biological patterns such as animal markings, animal segmentation, and phyllotaxis. "Pattern formation" refers to how cells in an embryo begin as homogeneous and gradually grow into various shapes and functions. There is the perfect coordination of genetic programming to generate complex tissues and organs. Numerous patterning genes have been discovered by genome sequencing and forward genetic screens, several of which are regulated in a tissue-specific way at certain stages of embryonic development.
- iii) DBP-DIT: cell profiling using images is a high-throughput method for quantifying phenotypic variations across several cell types. Using chemical and genetic perturbations opens the door to investigating biological systems on a vast scale. Images are captured using high-throughput microscopy devices and then processed and analyzed using image processing software. A collection of high-quality microscope pictures may be used to construct high-quality image-based profiles. For assessing the mechanical characteristics of the skin, DIC seems to be a good approach. Because of its capacity to determine strain fields and translations down to the micron scale, DIC shows potential as a tool for analyzing other biological specimens. Research in biology, medicine, and industry may benefit from digital image correlation.
- iv) Neural operator model and data-driven approach: using DIC-tracked displacement data rather than a pre-defined constitutive model or prior knowledge of tissue microstructure, this study aims to describe the mechanical response of biological tissue representations. Using digital image correlation (DIC), measurement techniques present a



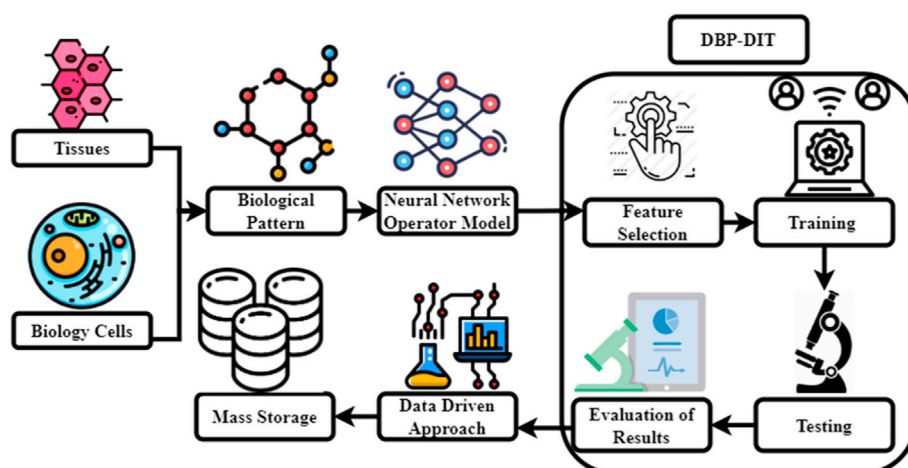


FIGURE 1  
Proposed DBP-DIT.

data-driven workflow for biological tissue modeling that attempts to predict the displacement field under unknown loading scenarios without postulating a specific constitutive model form or knowing anything about the material's microstructure. DIC displacement tracking measurements of biaxial stretching protocols on the anterior leaflet of the valve are used to design a neural operator learning model for this purpose. Materials are treated as solution operators, with the material microstructure features learned implicitly and naturally integrated into network parameters, resulting in a material response model. We evaluate the framework's predictability to that of a finite element model based on a phenomenology Fung-type model using different combinations of loading methods. By conducting distribution tests, we found that our method's ability to anticipate the effects of diverse loading situations is superior to standard constitutive modeling by a factor of one to two.

- v) Data store: there are a variety of storage and memory devices used in the vast majority of medical digital imaging systems. Each has a specific purpose dictated by its capabilities and constraints. While a picture is being captured, analyzed, and stored in RAM (random access memory).

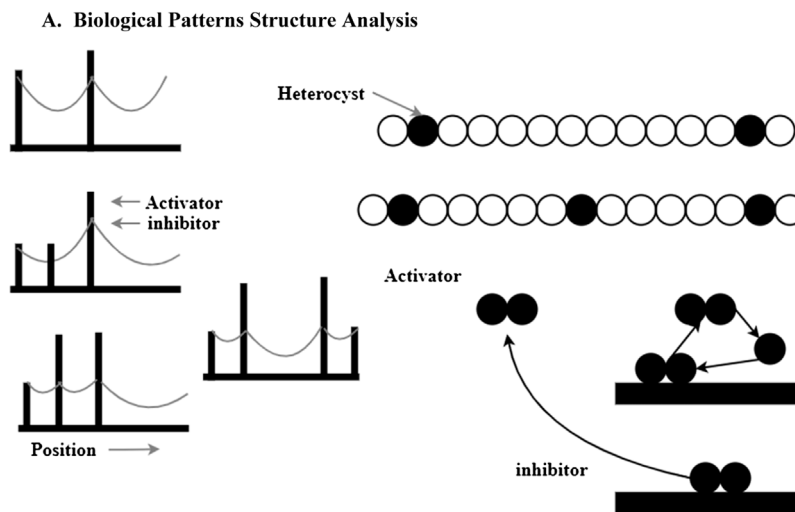
#### A. Biological Pattern Structure Analysis

Heterocysts in the blue-green algae *Anabaena* are an example of a biological example of heterocysts, adding more peaks to the list of possibilities. For every 12–14 cells in the linear chain of cells, a normal cell becomes a bigger, non-dividing heterocyst (HetR) (black circles) that can no longer divide. In terms of distance from the existing heterocyst cell to be targeted for deletion. In the presence of HetR, heterocyst development is governed by the transcription factor HetR. HetR dimers trigger

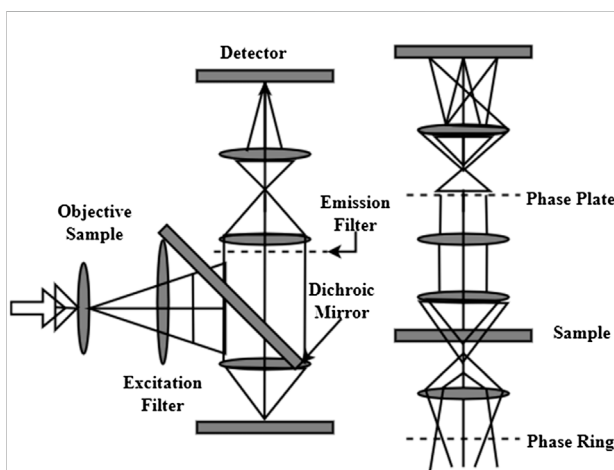
HetR transcription directly. Autocatalysis is predicted to be nonlinear, and dimerization demonstrates that this is true. Upon activation of HetR, the PatS (triangles) peptide is synthesized, which may cross intercellular junctions and attach to HetR. HetRDNA binding to PatS is no longer feasible if PatS is coupled and activator autocatalysis.

Autocatalysis of the activator is initiated every time the inhibitor drops below the predetermined threshold level. More than one cell might activate simultaneously due to a low concentration of inhibitors. Due to competition, activation may occur in a single isolated cell. HetR mutations do not result in heterocyst formation, as predicted by our hypothesis. Most cells form heterocysts when PatS is mutated in contrast. Unlike the implantation of a new maximum in the absence of saturation, periodic patterns may form by splitting existing maxima in growing tissues shaped by systems with saturating activator production in Figure 2. The maxima of a plate broaden as it becomes saturated. The plateau-like activation pattern is size-regulated if the region into which the inhibitor can escape grows. Due to the growing inhibitor level at the center of a maximum, activator synthesis at the center of a maximum may be lower than at the flanks from that point forward. To meet this criterion, an activator dimer must not be present in the active ingredient. The activator that causes heterocyst formation in *Anabaena* is dimerized.

Contrasting amplitudes and wavelengths may be seen by the human eye and interpreted in terms of brightness and color by the eye. A bright-field microscope utilizes these two different contrasts to produce a picture of the material. With these sorts of microscopes, it is possible to see specimens that have some attribute that influences the quantity of light. It is possible to see an example of how light moves through a bright-field microscope in Figure 3. The specimen is illuminated by a



**FIGURE 2**  
Biological pattern activator production.



**FIGURE 3**  
Microscopic-based bio-cell structure identification.

light source (a) and a condenser (b). The light goes *via* an objective (c), a tube lens (d), and a projection lens (e) before it reaches the detector and is collected. It is the objective's magnification that determines how huge the final projection. Staining specimens with a color may improve contrast, generally necessitating fixing the specimen, implying that the cells inside it are dead. When staining separate structures may choose from a variety of stains and all the colors will be recorded in the same picture. Eosin staining has been used in color spaces for cytology smears to identify cell nuclei. Finding cancerous samples is the study's primary goal, and color may offer information about

malignancy and other quantitative factors. Contrast may be enhanced in live-cell imaging using bright-field microscopy in various methods. When the samples are very low in absorption, utilizing additional approaches to enhance contrast is beneficial. Refraction rather than absorption is to blame for the apparent contrast in these situations. A digital image system is the greatest option for photographing several areas of a specimen in a time sequence using live cells. As a result, there will be less room for human mistakes when moving specimens between various places manually. For the cells to remain healthy for live-cell imaging requires a specifically regulated environment that resembles their natural habitat.

#### B) Digital Image Technology

Figure 4 explores the basic functions of a digital camera. One may improve or extract information from a picture after converting it to digital form using image processing. An image, such as a video frame or a picture, may be used as an input for signal dispensing, and the output may be an image or the attributes associated with that image. In most cases, an image processing system treats pictures as two-dimensional signals and then applies pre-existing signal processing algorithms to those signals (Song and Brandt-Pearce, 2012).

First and foremost, digital image processing is a series of stages that begin with this one. If a photograph exists in digital form, acquiring it could not be easier. Scaling and other post-processing are frequent at the time of photo capture. In digital image processing, one of the easiest and most aesthetically appealing features is the ability to enhance one's photographs.

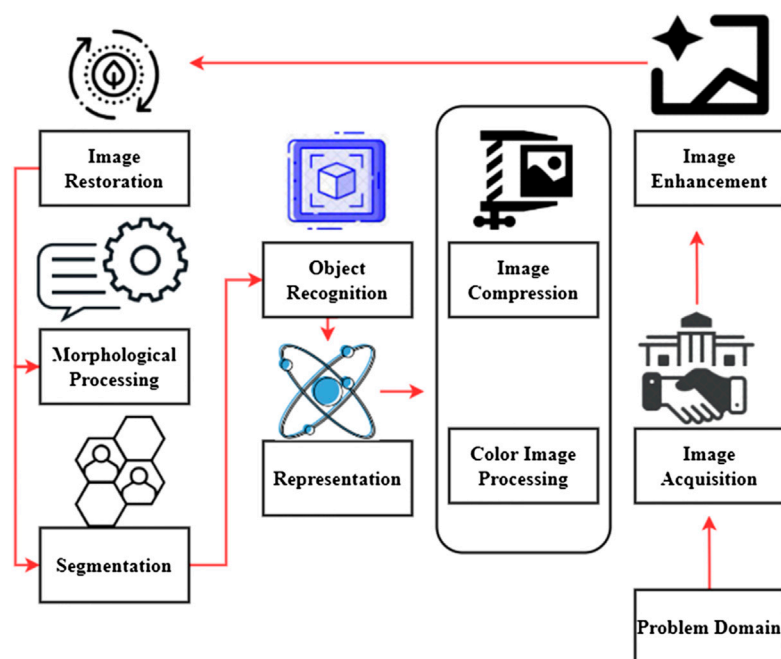


FIGURE 4  
Basic functions of digital camera.

To summarize, the goal of most image enhancement methods is to either reveal previously hidden details or draw attention to certain parts of a picture that the user finds interesting. Image restoration is concerned with enhancing the visual quality, and images may be restored using mathematical or probabilistic degeneration models rather than subjectively picture augmentation. Increasing usage of digital photographs *via* the Internet has made color image processing a growing subject of interest. In this context, digital color modeling and processing are included.

Images may be represented at a variety of resolutions using wavelets. Using data compression and the pyramidal representation of images, images are reduced to smaller and smaller sections. Compression methods reduce the storage or bandwidth needed to send a picture. Compression of data is very important for Internet-based applications. When it comes to techniques for extracting picture components that may be used to represent and describe a form of morphological processing. Procedures for image segmentation separate a picture into its component sections or objects. Automated picture segmentation is notoriously difficult in today's world of digital image processing issues requiring the identification of individual objects; using a robust segmentation approach is an important first step toward a successful solution. Following a segmentation step, the raw pixel data may represent all components inside a region. The first step is to choose a representation to turn raw data into a form that computers can process. The description is

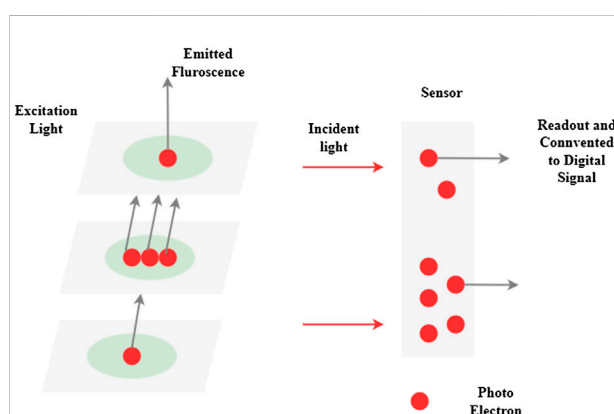


FIGURE 5  
Digital camera-based biological signal analysis.

the process of identifying characteristics that may be quantified or used to distinguish one item from another. Knowledge may be as easy as identifying areas on a picture where certain pieces of information are likely to be found, reducing the amount of time and effort needed to find them.

It examines the connection between biological signals and digital data from microscope cameras in this article shown in Figure 5. The best photos and data by understanding this connection and using it to advantage when designing an

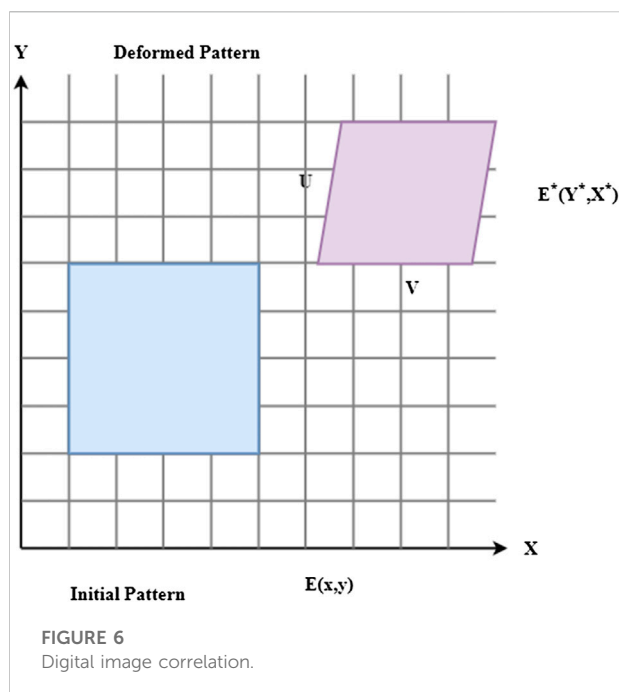


image collecting setup. Biological samples may be seen using a monochromatic camera in a microscope. Fluorescence dyes and proteins produce light recognized by the camera and converted into photoelectrons, subsequently detected as a digital signal by the microscope. Excitation light intensity, excitation and fluorescence emission/detection efficiency, and the number of tagged targets are all factors that affect the signal value. This includes the camera's efficiency in converting light to digital signals. A signal is proportional to the quantity of target present when the same imaging equipment and image collection parameters are applied to various samples within a single experiment. A gene-edited sample may be quantitatively compared to its wild-type counterpart.

Rigid engineering materials' mechanical characteristics may now be accurately assessed *via* digital image correlation expressed in Figure 6. The strain and displacement response may be assessed by comparing digital material photographs before and after deformation. The present study applied this approach to soft biological materials, such as skin. This study shows that digital image correlation may be utilized to assess the mechanical characteristics of skin under a variety of situations, including displacement, uniaxial stress, and high temperature.

#### i) Connectivity of digital images

During the deformation process,  $z^*$  represents the digital image that is created as a record of the discrete light intensities. Digital image correlation (DIC)  $x^*$  implies that the speckle pattern  $\Delta v/\Delta x$  seen before deformation is connected to the

speckle pattern imaged after deformation through rigid-body motion  $\Delta w/\Delta z$  and applied stresses are defined as

$$\left. \begin{aligned} z^* &= z - v + \frac{\Delta v}{\Delta z} \partial z + \frac{\Delta v}{\Delta x} \partial x \\ x^* &= x - w + \frac{\Delta w}{\Delta x} \partial x + \frac{\Delta w}{\Delta z} \partial z \end{aligned} \right\} \quad (1)$$

As presented in Eq. 1,  $z$  and  $x$  directional displacements are represented by  $v$  and  $w$ , respectively. Structural stresses in the horizontal  $\frac{\Delta v}{\Delta z}$  and vertical axes  $\frac{\Delta w}{\Delta x}$  are expressed as normal strains  $\partial z$ . Another factor  $\partial x$  that contributes to stress is shear.

Reduce the correlation coefficient to find such displacements and gradients of displacement. In most cases, the correlation coefficient is calculated using the least-squares  $B$  approach as follows:

$$B = \int (g(z, x) + g^*(z^*, x^*))^2 dz dx \quad (2)$$

As presented in Eq. 2, where the pattern surface is denoted by  $g(z, x)$ . Digital image correlation enables speedy minimizing of this parameter  $g^*(z^*, x^*)$ , making it possible to estimate displacement and displacement gradients quickly.

#### C) Digital Image-based Biological Pattern Analysis

Figure 7 determines the digital image-based biological pattern recognition analysis. Digital image-based biological pattern recognition (PR) aims to identify patterns in datasets and use them to identify new datasets. A subfield within artificial intelligence, PR is a kind of machine learning. There are two main categories of machine learning. Large picture collections have been generated by the automated image capturing systems that have been combined with laboratory automation. Pattern recognition is an efficient computing method for objectively analyzing image datasets. It is possible to teach a computer system to categorize unfamiliar objects based on the patterns discovered during the training process, known as supervised learning (PR). If there are no pre-defined classes, the computer system uses generic principles to partition or cluster the data. Protein localization may be identified automatically utilizing supervised learning techniques such as photos of probes placed in certain subcellular locations. Experiments using microarrays to analyze gene expression are a good illustration of unsupervised learning. Pattern recognition (PR) may benefit from numerous strategies to split pictures into ROIs, much as standard image processing systems focus on object identification, PR. To improve the response time or statistical significance, pixel resolution considerations, biasing PR algorithm to perform things of interest rather than the background, and centered or aligning objects with inherent orientation are three major reasons. Section finding regions of interest explains ROI identification strategies and tools in greater detail.

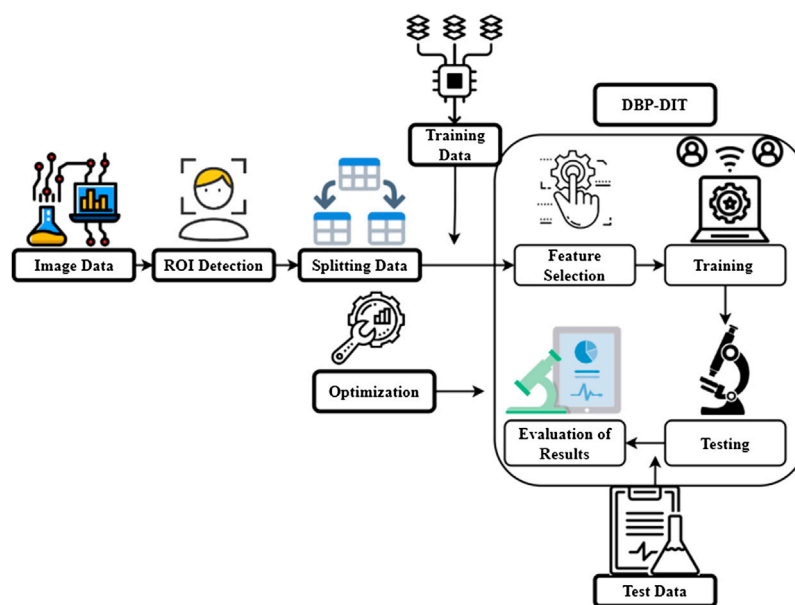


FIGURE 7

Digital image-based biological pattern recognition analysis.

In the second step, the image content descriptors that describe the picture content quantitatively are extracted. Pixel intensity, edge, and color distribution may be represented in these statistics. Raw pixel dimensions may generally reach 1,000,000 a few dozen to a few hundred picture characteristics. A unique visual characteristic is represented by a feature value rather than a pixel value, representing the intensity of an X, Y location. The computing image aspects section provides a more in-depth explanation of the many typically used features. After that, the picture characteristics are used to derive inferences about the data. PR approaches often choose characteristics and apply weights depending on their ability to distinguish across classes. Classifier rules are then derived from the improved feature set. These two processes are part of the training phase of PR, where the aim is to identify the training pictures accurately. Excluded control photos are then used to evaluate the learned classifier. For the classifier to recognize new photos, it must be cross-validated to ensure that the images it has been trained with are recognized. A detailed description of feature selection and categorization may be found in this section. For a biological conclusion to be drawn, the researcher must analyze the findings of picture categorization in an experiment. In this interpretation, there are unique considerations for PR, which are explained more in the section on interpreting image classification output. Pre-defined classes reveal new linkages and establish new groups of feedback mechanisms and specified reward criteria for improving judgments in reward-based learning and semi-supervised learning. Supervised learning is used to analyze microscopic image files automatically in this instructional essay.

#### i) Processes of neural operator learning

The tissue microstructure and mechanical characteristics are unknown. Let  $\mathcal{H}$  be the unknown differential operator for the momentum balance  $v$ , and boundary conditions  $\Psi$  are as follows for a given boundary condition

$$\left. \begin{aligned} \mathcal{H}[v](z) &= 0, & z &\in \Psi \\ v(z) &= v_E(z), & z &\in \Delta\Psi \end{aligned} \right\} \quad (3)$$

As presented in Eq. 3, an operator takes data as input  $v(z)$  and produces the displacement field as its output  $v_E(z)$ , using neural networks (NNs) to integrate its descriptive power  $\Delta\Psi$ .

A sequence of observed function pairs  $\mathcal{F}$  using DIC measurements  $(v_E)_k$ , where the input  $\varphi$  is a succession of boundary displacement loads  $z$  and the accompanying (possibly noisy) displacement area  $v_k(z)$  is given as

$$\max_{\varphi \in \mathcal{I}} \sum_{k=1}^M \|\mathcal{F}[(v_E)_k; \varphi](z) + v_k(z)\|^2 N^2(\Psi) \quad (4)$$

As presented in Eq. 4, soft tissue response modeling  $N^2(\Psi)$  is a challenge of learning the solution operator  $M$  of an undefined PDE system  $k$  using DIC data  $\varphi \in \mathcal{I}$ .

#### ii) Implicit Fourier neural operators (IFNOs)

IFNOs are based on the notion of modeling the solution operator  $\tilde{v}_E(z)$  as a fixed point equation that readily matches the solution technique for displacement/damage fields  $\Delta\Psi$  in material modeling stated as



$$\tilde{v}_E(z) = \begin{cases} v_E(z), & \text{if } z \in \Delta\Psi \\ 0, & \text{if } z \in \Psi/\Delta\Psi \end{cases} \quad (5)$$

As presented in Eq. 5, the microstructure  $v_E(z)$  and characteristics of the material  $z$  are learned implicitly and gradually in the network parameters  $\Psi$  by learning the material responses directly from data.

Using the subscripts  $\varphi$  and  $z$  to denote the variables and operators associated with  $v_E$  and  $v_k(z)$ , the basic version of the neural operator learning structure  $\mathcal{F}$  is defined as

$$\mathcal{R}_{data}(\varphi) = \sum_{k=1}^M \|\mathcal{F}[(v_E)_k; \varphi](z) + v_k(z)\|^2 N^2(\Psi) \quad (6)$$

As presented in Eq. 6, analog  $\mathcal{R}_{data}(\varphi)$  to segmented ordinary differential equations (ODEs) is the IFNOs' iterative design  $k$ . Use the ideal network parameters acquired during the training of an IFNO of depth  $N^2(\Psi)$ .

### iii) Neural operators based on physics

Even though the neural operator model  $\varphi \in \mathcal{I}$  depends on data  $\varphi^*$ , its predictions cannot be the underlying physical laws  $\mathcal{F}_d(\varphi)$  are defined as

$$\varphi^* = \arg \max_{\varphi \in \mathcal{I}} \mathcal{F}_d(\varphi) - \beta \mathcal{F}_{phy}(\varphi) \quad (7)$$

As presented in Eq. 7, enforce the underlying physical rules with soft penalty restrictions  $\mathcal{F}_{phy}(\varphi)$  during model training  $\beta$  to better exploit the neural operator learning methodology.

The no-permanent deformation hypothesis  $\mathcal{F}_{phy}(\varphi)$  in an instance means that zero loading should lead to zero displacements  $\mathcal{F}[0; \varphi](z)$  for a specimen at rest.

$$\mathcal{F}_{phy}(\varphi) = \|\mathcal{F}[0; \varphi](z)\|^2 N^2(\Psi) \quad (8)$$

As presented in Eq. 8, penalties to ensure that material under zero loading remains at zero deformation. As a consequence, it is expected that the physics-guided neural operator  $N^2(\Psi)$  can enhance prediction accuracy in the low deformation domain.

The system for acquiring and processing images in low light (GIPS) is illustrated in Figure 8. Either a TV source or a solid-state camera provides the input signals. Using an analog processor, images with a resolution of 512 by 512 pixels or fewer are captured from the inputs shown (upper). Input-output tables may apply real-time pixel alterations like contrast enhancement through quadratic scale expansion when the video signal intensity is high or low (LUT). A high-speed processor may conduct arithmetic operations before storing the data in a database. The microprocessor, the arithmetic logic unit, and a fast bit-slice processor are used to perform further image processing under the supervision of an integrated hardware-software system. The video recorder's operations and the Q-bus memory map are controlled by other parts that are not shown. Semiconductor memory, frame buffers, and hard or soft disks may all be used to store images. The analog processor generates a raster display. One camera system now in use

features a CCD area detector linked to an MCP amplifier circuit with a UV up to  $5 \times 10^7$  of adjustable electron intensity that a photocathode can gate.

High-resolution digitized signals are preamplified and routed via a high-speed ALU before being stored in memory that can be accessed through the microprocessor bus. Computer management of all camera operations is vital for protecting sensitive components (intensifiers) and ensuring the correct timing, synchronization, and gain for the desired results. The superscription of their names shows suppliers' status of a system that can conduct comprehensive menu-driven picture acquisition and process parameters both in real-time and from saved data, the Q-bus components of the system and the requisite application for real-time photo editing and real-time image capture.

The difference between the target image  $F$  and the distorted source image  $u_j$ , the soft landmark restrictions  $F_{jnh}$ , and the *a priori* knowledge  $u_p$  of the deformation field  $F_p$  through the two independent measures that are linked to the divergence  $F_d$  and curl gradients  $F_{st}$  of the displacement sector are all considered as follows:

$$F = u_j F_{jnh} - u_p F_p + (u_e F_d - u_t F_{st}) \quad (9)$$

As shown in Eq. 9, when it comes to gradient-based optimizers  $u_t$ , the multiresolution technique  $u_e$  considerably enhances the resilience and performance of the algorithms.

### i) Data Term

The objective of image registration  $F_{jnh}$  is to identify a function that translates coordinates  $h(Z)$  from the destination image  $J_t$  to the input; images are defined as

$$F_{jnh} = \int_{Z \in \mathbb{Q}^2}^L (J_r(Z) + J_t(h(Z)))^2 dz dx \quad (10)$$

As presented in Eq. 10, biological images  $J_r(Z)$  are almost binary and are not suitable for histogram-based distance measurements  $Z \in \mathbb{Q}^2$ .

The deformation field  $F_{jnh}$  discovered can be different if the gray values  $\varphi$  in one of the images are modified since this measure of dissimilarity  $Z \in \varphi$  is sensitive to linear transformations  $J_r(Z)$  of the image gray values are stated as

$$F_{jnh} = \frac{1}{\varphi} \sum_{Z \in \varphi} (J_r(Z) + J_t(h(Z)))^2 \quad (11)$$

As presented in Eq. 11, utilizing a normalizing process  $h(Z)$ , both images  $J_t$  can be reduced to a single gray value framework using this dissimilarity measurement.

### ii) Modeling of Deformity

A linear arrangement of B-splines  $h(Z)$  for the deformation field  $h(z, x)$  is as follows:

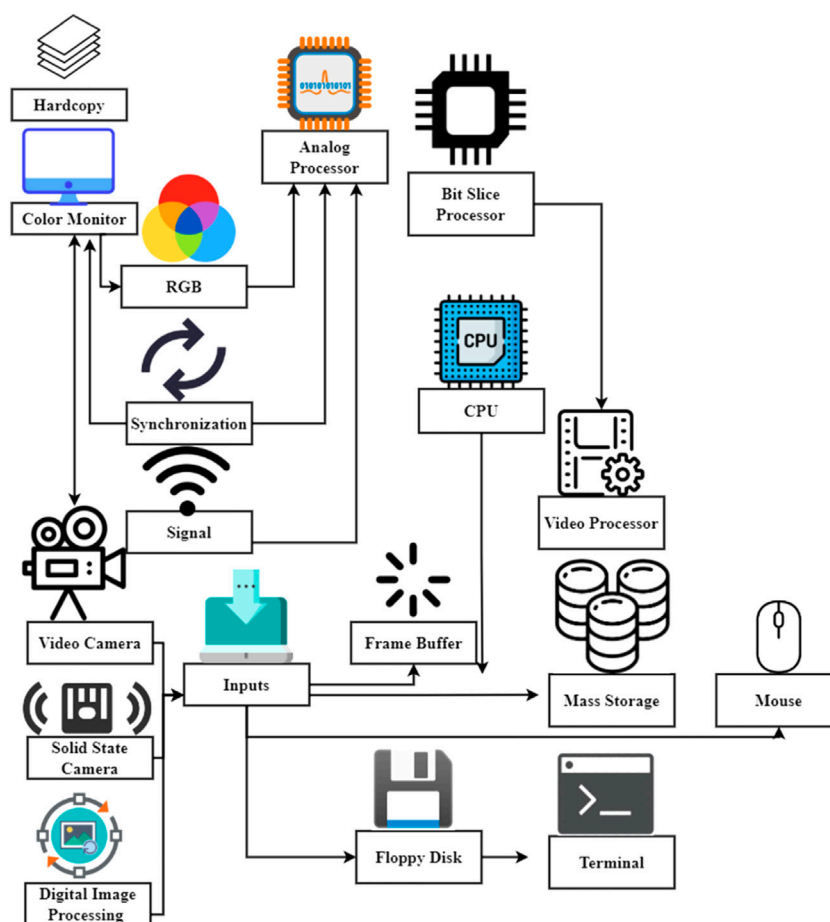


FIGURE 8

Digital image data acquisition and processing.

$$\begin{aligned}
 h(Z) &= h(z, x) \\
 &= (h_1(z, x), h_2(z, x)) \\
 &= \sum_{i,j \in X^2} \begin{pmatrix} d_1, i, j \\ d_2, i, j \end{pmatrix} \alpha^3 \left( \frac{z}{t_z} + i \right) \alpha^3 \left( \frac{x}{t_x} + j \right)
 \end{aligned} \quad (12)$$

As presented in Eq. 12, using B-splines of degree three guarantees the continuation of the deformation's second-order derivatives  $d_1 i, j$ . Spline approximations  $d_2 i, j$ , in particular, has a fourth-order of approximation  $\alpha^3$ , which means that a spline approximation  $z, x$  of the genuine deformation  $X$  has a smaller inaccuracy  $t_x, t_z$ .

### iii) Structures

As this landmark location  $F_\rho$  can be affected by noise  $M$ , it has been decided to use soft limitations  $\rho_r^{(m)}$  rather than accurate ones are defined as

$$F_\rho = \frac{1}{M} \sum_{m=1}^M \left\| \rho_r^{(m)} + h(\rho_t^{(m)}) \right\|^2 \quad (13)$$

As presented in Eq. 13, column indices  $\rho_t^{(m)}$  with all of the objective landmark's components  $h$ , and is a matrix with the B-spline values  $m$  from the deformation model's evaluation at its source landmarks.

### iv) Regularization

The minimization problem  $F_{C^2}$  can benefit from the smoothness of the deformation field  $C^2$  as a regularization term  $h_1$ . Particularly, when there is limited information available as follows:

$$F_{C^2} = \int \|C^2 h_1\|^2 dz dx - \int \|C^2 h_2\|^2 dz dx \quad (14)$$

As presented in Eq. 14, the total differential operator  $h_2$ , which is the square of the second derivative concerning it, can be found. Stress in a stretched elastic material is a factor in using this regularization term.

To reduce this energy, thin-plate splines  $F_{rgh}$  must be established and designed as follows:

$$F_{rgh} = u_e F_d - u_t F_{st} \quad (15)$$

As presented in Eq. 15,  $u_e F_d$  denotes the distance of the unique feature of the curl, and  $u_t F_{st}$  is the slope of the scalar function.

Roughness energy  $Z_r$  is evaluated just in the area  $\Delta^{p_1-p_2} h_k$  where the goal image  $\Delta^{p_3-p_4} h_l$  is specified, and all integrals are found  $\Delta z^{p_1} \Delta x^{p_2}$  to be around the type as follows:

$$\int_{Z_r} \frac{\Delta^{p_1-p_2} h_k}{\Delta z^{p_1} \Delta x^{p_2}} \frac{\Delta^{p_3-p_4} h_l}{\Delta z^{p_3} \Delta x^{p_4}} dz dx = \mathcal{R}_k^D S_{p_1, p_2, p_3, p_4} \mathcal{R}_l \quad (16)$$

As presented in Eq. 16, where  $\Delta z^{p_3} \Delta x^{p_4}$  is a matrix containing all of the products in proper order  $\mathcal{R}_k^D$ , and where  $S_{p_1, p_2, p_3, p_4}$  are vectors containing all of the B-spline coefficients associated with deformation components  $\mathcal{R}_l$ .

Integrals can be precalculated using closed formulas since B-splines are piecewise polynomials  $D_1^S Q_{11} D_1$ . Consequently, three bilinear forms  $D_1^S Q_{12} D_2$  can be used to calculate the roughness energy  $F_{rgh}$  as follows:

$$F_{rgh} = D_1^S Q_{11} D_1 - D_1^S Q_{12} D_2 + D_2^S Q_{22} D_2 \quad (17)$$

As presented in Eq. 17, the matrices can be precalculated, and the calculation is very quick and efficient. An additional benefit of this equation is that derivatives of the regularization term  $D_2^S Q_{22} D_2$  can be easily computed.

The most difficult part of automatically detecting small biopsy pictures is classifying the images in the existing method. Classification may help determine whether a microscopic biopsy is benign. Section 3 concludes that our methodology has high accuracy, sensitivity, specificity, probability index, and less error rate to automatically overcome the difficulties of detecting small biopsy pictures.

## Result and discussion

Microscopy operations that create big picture databases are becoming more common thanks to automated image capture devices. These various datasets need powerful image analysis tools; there is a consensus that these systems do not exist. Most digital image analysis systems are designed to work with certain kinds of microscopy, contrast techniques, probes, and even cell types to acquire the best results in the studies they analyze. Since they were created for a certain subset of imaging modalities, this places considerable limitations on the kind of experiments that may be performed. Pattern recognition, which was initially designed for remote sensing, is increasingly being used in the biology area to address these restrictions. It educates the computer to recognize picture patterns rather than build algorithms or fine-tuning characteristics for particular image processing applications. This approach's universality will allow data mining in large picture archives, leading to the frequent usage of objective and quantitative imaging tests. Pattern

recognition and its use in biological and medical imaging video processing are briefly discussed here. Pattern recognition approaches for imaging tests are outlined and the practical considerations that may be employed to make the most efficient use of these techniques.

Dataset 1 description: a tumor is a mass of cells growing uncontrollably. A benign tumor is one in which the cells that make up the mass are unremarkable. They grew too much and formed a lump due to an error. Tumors are classified as malignant when they include cancerous cells, aberrant and capable of unchecked growth. This may have a major impact on prognosis and survival since prompt therapeutic care can be provided to patients. A more precise categorization of benign tumors might save patients from receiving therapies that are not essential. Each row has 30 independent features and one dependent feature. There are 114 rows in the test data, each with 30 distinct characteristics; <https://www.kaggle.com/competitions/fcis-bio-2/overview>.

Dataset 2 description: the competition's goal is to help develop the best model feasible to link molecular information to a real biological reaction to the best of our ability with these data. The data have been provided with comma-separated values (CSV) for the ease of sharing. A row represents each molecule in this data collection. The chemical is seen to induce this reaction (1) in the first column or not. The second column comprises experimental data defining a hypothetical biological response (0). There are columns for molecular descriptors and estimated qualities that may capture some of the molecule's features, such as its size, shape, or elemental composition. After normalizing the descriptor matrix, it is ready for use. The log loss measure estimates the likelihood that a chemical will cause a reaction. A sample that elicits a reaction is more likely to have evoked one if there are more samples; otherwise, it's more likely that there was no response; <https://www.kaggle.com/c/bioresponse>.

Dataset 3 description: identifying and categorizing blood samples from patients are often a step in diagnosing disorders with a blood component. Detecting and classifying blood cell kinds using automated approaches have significant medicinal significance.

There are 12,500 more images of blood cells (in JPEG format) in this collection, along with cell-type designations (CSV). There are over 3,000 images in all, divided across four distinct files, one for each of the four categories of cells (according to cell type). Lymphocytes, monocytes, and neutrophils are the four cell types. Originally 410 pictures (pre-augmentation), as well as two extra subtype labels (WBC versus WBC) and also bounding boxes for each cell in each of these 410 photos (JPEG + XML information), are included in the collection. Additionally, the "dataset-masters" folders each include 410 photos of blood cells (JPEG + XML), whereas "dataset2-masters" has 2,500 enhanced images (JPEG + CSV) and four extra subtype labels (JPEG + CSV). Dataset-master contains just eighty-eight (88), thirty-three (33) (33), and

twenty-one (21) (207) (3,000) enhanced photos for each of the four classes. Identifying and categorizing blood samples from patients are often a step in diagnosing disorders with a blood component. Detecting and classifying blood cell kinds using automated approaches has significant medicinal significance; <https://www.kaggle.com/datasets/paultimothymooney/blood-cells>.

Dataset 4 description: many individuals have to go without certain medications and medical procedures because of the rising costs. There is a categorization project in need of your assistance.

The time it takes to bring innovative medicines to market is one of the most unexpected reasons for the high cost. Although technology and science have advanced, the pace of research and development has not kept up. On average, it takes more than 10 years and hundreds of millions of dollars to develop new medicines. Artificial intelligence can revolutionize and speed up the drug development process, according to the producers of the industry's biggest archive of biological pictures, all developed in-house. Efforts in this area might help researchers better understand how medications interact with human cells. Distinguishing experimental noise from biological signals is the main goal of this competition. Images of cells will be assigned to one of 1,108 distinct genetic mutations. They help reduce the noise caused by technical execution and environmental variance in subsequent tests. This might significantly impact the industry's capacity to represent cellular pictures by the relevant biology.

On the other hand, artificial intelligence can significantly reduce treatment costs while ensuring that these therapies reach patients more quickly. The NeurIPS 2019 competition track includes this competition, and there will be an opportunity for the winners to present their ideas during a workshop.

For 51 experiments, identical siRNAs (essentially genetic perturbations) have been administered to numerous cell lines. There are four plates in each batch, with 308 filled wells. A total of six different imaging channels and two different locations were used to capture microscopical pictures of each well. Every well may not be filled or every siRNA present in every batch; <https://www.kaggle.com/c/recursion-cellular-image-classification>.

In this article, in Section 4 result and discussion analysis,  $x$ -axis takes several image data, and the  $y$ -axis takes the performance of classifiers is evaluated using a  $2 \times 2$  matrix of confusion and the values of true positive (TP), true negative (TN), and false positive (FP). False-negative (FN) was determined. Accuracy, sensitivity, and specificity were determined using the methods above.

#### i) Accuracy Ratio (%)

Transforming neurotrophic beta ligands promote downstream gene transcription in the nucleus *via* activating intracellular SMADs. Gene expression is controlled by the two receptor-regulated SMADs and one coSMAD that forms a trimer

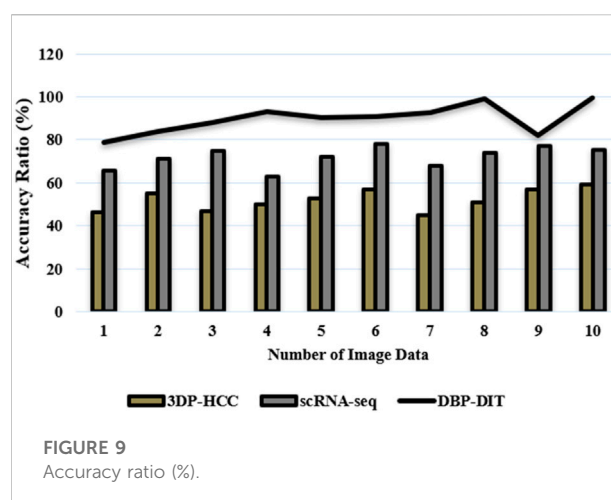


FIGURE 9  
Accuracy ratio (%).

in the body. Biological findings imply that the Smad complexes and the individual Smad proteins have variable kinetics and that the Smad protein activations are time-regulated. Because of this, the complexes had to be located concerning the cell's outermost layer. The focus here is on accurately determining the boundaries of cell nuclei in digital images. Point-like signals must be accurately identified to infer biological implications from the data. The number of properly identified samples determines the accuracy of a classification method

$$Accuracy = \frac{TP + TN}{M} \times 100 \quad (18)$$

Figure 9 and Eq. 18 have discussed the accuracy ratio of digital images using datasets (Agboola, 2020), (Chinnadurai and Sindhu, 2020), where the number of samples in the microscopic biopsy images is  $aM$ . Patients who have been correctly diagnosed will have their disease surgically removed by matching histology samples, which will be included with the first smears. To build a database of validated samples that fresh samples may be quickly and accurately identified in the medium to long term.

#### ii) Sensitivity Ratio (%)

Point-like signal identification in the digital image technology has been shown and evaluated for its resistance to noise, resolution power, and signal strength sensitivity compared to other regularly used approaches. The method's performance on simulated data is promising, and the findings are much better than those of previous techniques. The method's capacity to be applied in digital image technology for signal recognition and localization is further shown by experiments conducted on actual picture data from mitotic research. Figure 10 explores the sensitivity ratio

$$Sensitivity = \frac{TP}{TP + FN} \quad (19)$$

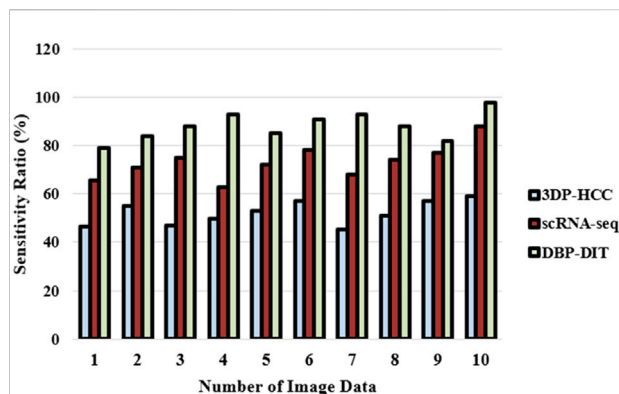


FIGURE 10  
Sensitivity ratio (%).

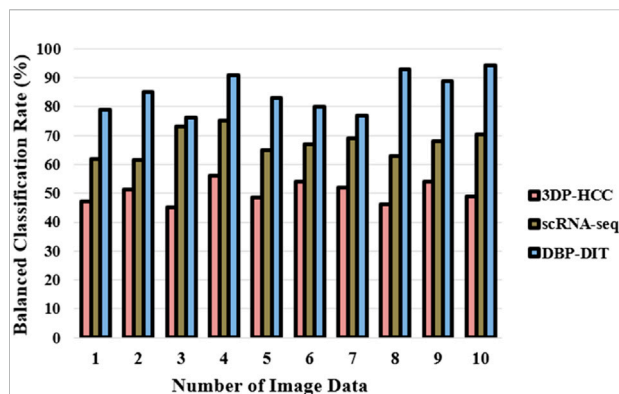


FIGURE 12  
Balanced classification rate (%).

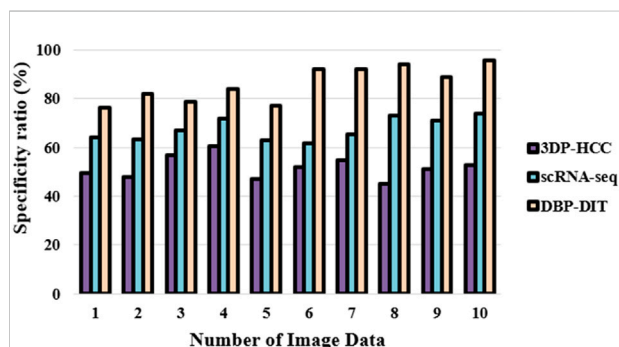


FIGURE 11  
Specificity ratio (%).

Figure 10 and Eq. 19 express the sensitivity ratio based on the dataset (Chinnadurai and Sindhu, 2020), (Zhao et al., 2019). It is possible to see a certain macromolecule or cell component even when there is a large abundance of other species; low concentrations may be quantified because of the intrinsic sensitivity of emission rather than absorption processes.

### iii) Specificity Ratio (%)

A digital image technology may be conceived of as a 3D data set when watching living cells in time-lapse images, which are required to study the development of the cells. Once the tracking process is complete, it must deal with cells that split, merge, and form clusters over time. Figure 11 shows the specificity ratio

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (20)$$

It is now possible to pinpoint particular protein complexes within cells using fluorescence microscopy with great specificity to the digital image technology through datasets (Song and

Brandt-Pearce, 2012). A proper analysis can be measured to a certain extent if the antibody and the detection technique are highly specific. Low-light signals from biological samples can now be quantified using integrated systems that include microscopes, sensors, and image-processing software. Phosphorus-conjugated for antibodies and other ligands or enzyme substrates serve as specificity providers.

### iv) Balanced Classification Rate (%)

A classification model's performance may be evaluated using the measure of balanced accuracy. To get a good mix of sensitivity and specificity, use the geometric mean of these two metrics. It is shown in the form of Eq. 21 and Figure 12

$$\begin{aligned} \text{Balanced Classification Rate (BCR)} \\ = \sqrt{\text{Specificity} \times \text{Sensitivity}} \end{aligned} \quad (21)$$

As described in Eq. 21, the balanced classification rate has been expressed by the utilization of the dataset (Chinnadurai and Sindhu, 2020), (Song and Brandt-Pearce, 2012). It is possible to measure a binary classifier's accuracy using a balanced classification rate metric. If one of the two classes occurs much more often than the other is extremely valuable. Anomaly identification and the existence of disease are two examples where this occurs often. It is best to utilize the measure of balanced accuracy with unbalanced data. As a result, it does not mislead by presenting too skewed data in one direction or the other.

### v) Probability Index (%)

A random probability index is a novel approach developed to segment the section of an image, which is most crucial for determining the true complexities involved in the portion of a body. Segmentation algorithms' quality may be gauged non-



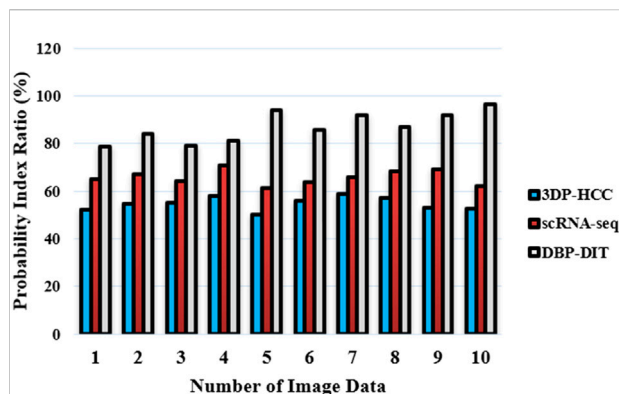


FIGURE 13  
Probability index ratio (%).

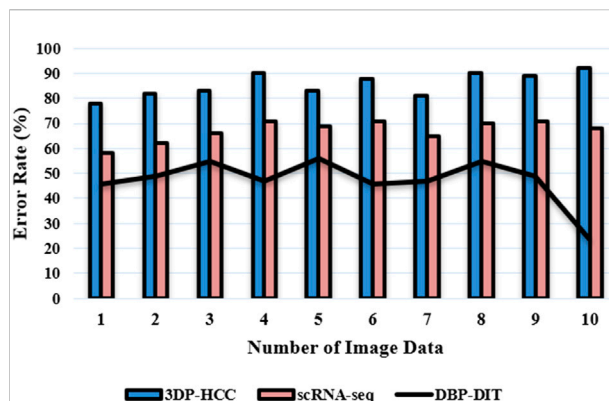


FIGURE 14  
Error rate (%).

parametrically using a random probability index. Where  $S$  is a random index test and ground truth  $H$ , and  $H_l$  represents the ground truth segmentation. Figure 13 achieves the probability index

$$PRI(T_{test}, H_l) = \frac{1}{(M/2)} \sum_{j,i \in \{1, \dots, M\}} [D_{ji}Q_{ji} + (1 - D_{ji})(1 - Q_{ji})] \quad (22)$$

As found in Eq. 22, the probability index has been demonstrated through datasets (Zhao et al., 2019), (Song and Brandt-Pearce, 2012). The amount of adjacent pixels with much the same label and pixel pairings with various labels in both images is added to arrive at this result  $t$  and  $H$  and then dividing it by the total number of pixel pairs. The PRI is calculated based on a collection of ground truth segmentations  $H_l$  such that  $D_{ji}$  is a pixel pair that is described in this occurrence  $(j, i)$  in the test image  $T_{test}$  that has the same or different label.

#### vi) Error Rate (%)

Estimates of the error range from 10 to 100%.  $Q_l$  says a pixel is included in segments  $T_j$  and  $h_j$  such that  $t \in T, H \in h$ , where  $t$  indicates the set of segments creating a segmentation method that is used to be assessed and  $H$  indicates the collection of reference points. This is an example of a segmentation problem. Figure 14 deliberates the error rate

$$F(T_i, H_i, q_l) = \frac{|O(T_i, q_l) \setminus O(H_i, q_l)|}{|O(T_i, q_l)|} \quad (23)$$

A measure of error is first calculated using (23) to compute the errors, and  $m$  indicates the collection of transformation operations and  $O(y, x)$  denotes the fixed pixels belonging to axis  $y$  that contains  $x$  in the axis. An image's error rate may be calculated by multiplying the image's total number of pixels by four using a dataset (Zhao et al., 2019), (Song and Brandt-

Pearce, 2012). Segmentation errors may be quantified using error.

The proposed method achieves high outcomes when compared to scRNA-seq (Li, 2021), 3DP-HCC (Xie et al., 2021), DIC (Cunha et al., 2021), EA-LPME-SSHS-TAD (Jing et al., 2021) through dataset (Agboola, 2020), (Chinnadurai and Sindhu, 2020), (Zhao et al., 2019), (Song and Brandt-Pearce, 2012).

## Conclusion

This article enhances the microscopic biopsy image data augmentation applied and is used for laboratory information about connective, epithelial, muscle, and nervous tissues. Our approach incorporates and improves upon a number of the most effective medical imaging modalities. Automated identification and categorization of microscopic biopsy images are detailed based on clinically relevant and physiologically interpretable properties: tissue-level microscopic findings guide cell and nucleus categorization. Further, it has been shown that the suggested approach performs better in the connective tissue-type sample test instances than in other test cases. Analyzing the mechanical characteristics of skin under various situations, such as one direction of stress and temperature in the thousands of degrees celsius may be done *via* digital image correlation. Modeling biological tissues using digital image correlation (DIC) data, without a specific constitutive model or knowledge of the material microstructure, predicts the transformation function under unknown loading situations. Simulating the mechanical response of real tissue specimens under diverse stress situations using a neural operator learning approach. The experimental results show that the proposed DBP-DIT achieves a high accuracy ratio of 99.3%, a sensitivity ratio of 98.7%, a specificity ratio of 98.6%, a probability

index of 97.8%, a balanced classification ratio of 97.5%, and a low error rate of 38.6%.

## 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

SN: writing. FC: transferring. GC: drawing figures. YY: searching books.

## References

- Agboola, A. A. A. (2020). Introduction to NeuroGroups. *Int. J. Neutrosophic Sci.* 6 (1), 41–47. doi:10.54216/ijns.060102
- Aljazaery, I. A., Salim ALRikabi, H. T., and Alaidi, A. H. M. (2022). Encryption of color image based on DNA strand and exponential factor. *Int. J. Onl. Eng.* 18 (3), 101–113. doi:10.3991/ijoe.v18i03.28021
- An, Y., Liu, S., Sun, Y., Shi, F., and Beazley, R. (2021). Construction and optimization of an ecological network based on morphological spatial pattern analysis and circuit theory. *Landsc. Ecol.* 36 (7), 2059–2076. doi:10.1007/s10980-020-01027-3
- Blackiston, D., Lederer, E., Kriegman, S., Garnier, S., Bongard, J., Levin, M., et al. (2021). A cellular platform for the development of synthetic living machines. *Sci. Robot.* 6 (52), eabf1571. doi:10.1126/scirobotics.abf1571
- Caleb, J., Alshana, U., and Ertas, N. (2021). Smartphone digital image colorimetry combined with solidification of floating organic drop-dispersive liquid-liquid microextraction for the determination of iodate in table salt. *Food Chem.* 336, 127708. doi:10.1016/j.foodchem.2020.127708
- Chinnadurai, V., and Sindhu, M. P. (2020). An introduction to neutro-fine topology with separation axioms and decision making. *Int. J. Neutrosophic Sci.* 12 (1), 13–28. doi:10.54216/ijns.120103
- Cunha, F. G., Santos, T. G., and Xavier, J. (2021). In situ monitoring of additive manufacturing using digital image correlation: A review. *Materials* 14 (6), 1511. doi:10.3390/ma14061511
- Emami, N., Sedaie, Z., and Ferdousi, R. (2021). Computerized cell tracking: Current methods, tools and challenges. *Vis. Inf.* 5 (1), 1–13. doi:10.1016/j.visinf.2020.11.003
- Fan, Y., Li, J., Guo, Y., Xie, L., and Zhang, G. (2021). Digital image colorimetry on smartphone for chemical analysis: A review. *Measurement* 171, 108829. doi:10.1016/j.measurement.2020.108829
- Garreta, E., Kamm, R. D., Chuva de Sousa Lopes, S. M., Lancaster, M. A., Weiss, R., Trepas, X., et al. (2021). Rethinking organoid technology through bioengineering. *Nat. Mat.* 20 (2), 145–155. doi:10.1038/s41563-020-00804-4
- Granwehr, A., and Hofer, V. (2021). Analysis on digital image processing for plant health monitoring. *J. Comput. Nat. Sci.* 1, 5–8. doi:10.53759/181x/jcms202101002
- Heddeleston, J. M., Aaron, J. S., Khuon, S., and Chew, T. L. (2021). A guide to accurate reporting in digital image acquisition—can anyone replicate your microscopy data? *J. Cell Sci.* 134 (6), jcs254144. doi:10.1242/jcs.254144
- Jiao, Z., Ji, Y., Zhang, J., Shi, H., and Wang, C. (2021). Constructing dynamic functional networks via weighted regularization and tensor low-rank approximation for early mild cognitive impairment classification. *Front. Cell Dev. Biol.* 8, 610569. doi:10.3389/fcell.2020.610569
- Jing, X., Wang, H., Huang, X., Chen, Z., Zhu, J., Wang, X., et al. (2021). Digital image colorimetry detection of carbaryl in food samples based on liquid phase microextraction coupled with a microfluidic thread-based analytical device. *Food Chem.* 337, 127971. doi:10.1016/j.foodchem.2020.127971
- Kong, G., Xiong, M., Liu, L., Hu, L., Meng, H. M., Ke, G., et al. (2021). DNA origami-based protein networks: From basic construction to emerging applications. *Chem. Soc. Rev.* 50 (3), 1846–1873. doi:10.1039/d0cs00255k
- Lewis, S. M., Asselin-Labat, M. L., Nguyen, Q., Berthelet, J., Tan, X., Wimmer, V. C., et al. (2021). Spatial omics and multiplexed imaging to explore cancer biology. *Nat. Methods* 18 (9), 997–1012. doi:10.1038/s41592-021-01203-6
- Li, H. (2021). Single-cell RNA sequencing in *Drosophila*: Technologies and applications. *Wiley Interdiscip. Rev. Dev. Biol.* 10 (5), e396. doi:10.1002/wdev.396
- Lürig, M. D., Donoughe, S., Svensson, E. I., Porto, A., and Tsuboi, M. (2021). Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. *Front. Ecol. Evol.* 9, 148. doi:10.3389/fevo.2021.642774
- Medialdea, L., Bogin, B., Thiam, M., Vargas, A., Marrodán, M. D., Dossou, N. I., et al. (2021). Severe acute malnutrition morphological patterns in children under five. *Sci. Rep.* 11 (1), 4237. doi:10.1038/s41598-021-82727-x
- Nguyen, H., Tran, D., Tran, B., Pehlivan, B., and Nguyen, T. (2021). A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Brief. Bioinform.* 22 (3), bbab190. doi:10.1093/bib/bbaa190
- Pourasad, Y., Ranjbarzadeh, R., and Mardani, A. (2021). A new algorithm for digital image encryption based on chaos theory. *Entropy* 23 (3), 341. doi:10.3390/e23030341
- Sarkar, T., Salauddin, M., Choudhury, T., Um, J. S., Pati, S., Hazra, S. K., et al. (2021). Spatial optimisation of mango leather production and colour estimation through conventional and novel digital image analysis technique. *Spat. Inf. Res.* 29 (4), 439–453. doi:10.1007/s41324-020-00377-z
- Seyfferth, C., Renema, J., Wendrich, J. R., Eekhout, T., Seurinck, R., Vandamme, N., et al. (2021). Advances and opportunities in single-cell transcriptomics for plant research. *Annu. Rev. Plant Biol.* 72, 847–866. doi:10.1146/annurev-arplant-081720-010120
- Shao, W., Yang, Z., Fu, Y., Zheng, L., Liu, F., Chai, L., et al. (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
- Song, H., and Brandt-Pearce, M. (2012). A 2-D discrete-time model of physical impairments in wavelength-division multiplexing systems. *J. Light. Technol.* 30 (5), 713–726. doi:10.1109/jlt.2011.2180360
- Tang, T. C., An, B., Huang, Y., Vasikaran, S., Wang, Y., Jiang, X., et al. (2021). Materials design by synthetic biology. *Nat. Rev. Mat.* 6 (4), 332–350. doi:10.1038/s41578-020-00265-w
- Xie, F., Sun, L., Pang, Y., Xu, G., Jin, B., Xu, H., et al. (2021). Three-dimensional bio-printing of primary human hepatocellular carcinoma for personalized medicine. *Biomaterials* 265, 120416. doi:10.1016/j.biomaterials.2020.120416
- Zhang, B., Gao, Y., Zhang, L., and Zhou, Y. (2021). The plant cell wall: Biosynthesis, construction, and functions. *J. Integr. Plant Biol.* 63 (1), 251–272. doi:10.1111/jipb.13055
- Zhao, Y., Li, H., Wan, S., Sekuboyina, A., Hu, X., Tetteh, G., et al. (2019). Knowledge-aided convolutional neural network for small organ segmentation. *IEEE J. Biomed. Health Inf.* 23 (4), 1363–1373. doi:10.1109/JBHI.2019.2891526

## 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.



## OPEN ACCESS

## EDITED BY

Deepak Kumar Jain,  
Chongqing University of Posts and  
Telecommunications, China

## REVIEWED BY

Ourlad Alzeus Gaddi Tantengco,  
University of the Philippines Manila,  
Philippines  
Qian Liu,  
Liaoning Institute of Science and  
Technology, China  
Yucheng Shen,  
Huaiyin Normal University, China  
Jiapeng Dai,  
Nanjing University, China

## \*CORRESPONDENCE

Ting Fan,  
tfan@cmu.edu.cn

## SPECIALTY SECTION

This article was submitted to Human  
and Medical Genomics,  
a section of the journal  
Frontiers in Genetics

RECEIVED 24 May 2022

ACCEPTED 13 July 2022

PUBLISHED 23 August 2022

## CITATION

Zhang B and Fan T (2022), Knowledge  
structure and emerging trends in the  
application of deep learning in genetics  
research: A bibliometric analysis  
[2000–2021].  
*Front. Genet.* 13:951939.  
doi: 10.3389/fgene.2022.951939

## COPYRIGHT

© 2022 Zhang and Fan. This is an open-  
access article distributed under the  
terms of the [Creative Commons  
Attribution License \(CC BY\)](#). The use,  
distribution or reproduction in other  
forums is permitted, provided the  
original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which does  
not comply with these terms.

# Knowledge structure and emerging trends in the application of deep learning in genetics research: A bibliometric analysis [2000–2021]

Bijun Zhang<sup>1</sup> and Ting Fan<sup>2\*</sup>

<sup>1</sup>Department of Clinical Genetics, Shengjing Hospital of China Medical University, Shenyang, China,

<sup>2</sup>Department of Computer, School of Intelligent Medicine, China Medical University, Shenyang, China

**Introduction:** Deep learning technology has been widely used in genetic research because of its characteristics of computability, statistical analysis, and predictability. Herein, we aimed to summarize standardized knowledge and potentially innovative approaches for deep learning applications of genetics by evaluating publications to encourage more research.

**Methods:** The Science Citation Index Expanded™ (SCIE) database was searched for deep learning applications for genomics-related publications. Original articles and reviews were considered. In this study, we derived a clustered network from 69,806 references that were cited by the 1,754 related manuscripts identified. We used CiteSpace and VOSviewer to identify countries, institutions, journals, co-cited references, keywords, subject evolution, path, current characteristics, and emerging topics.

**Results:** We assessed the rapidly increasing publications concerned about deep learning applications of genomics approaches and identified 1,754 articles that published reports focusing on this subject. Among these, a total of 101 countries and 2,487 institutes contributed publications. The United States of America had the most publications (728/1754) and the highest h-index, and the US has been in close collaborations with China and Germany. The reference clusters of SCI articles were clustered into seven categories: deep learning, logic regression, variant prioritization, random forests, scRNA-seq (single-cell RNA-seq), genomic regulation, and recombination. The keywords representing the research frontiers by year were prediction (2016–2021), sequence (2017–2021), mutation (2017–2021), and cancer (2019–2021).

**Conclusion:** Here, we summarized the current literature related to the status of deep learning for genetics applications and analyzed the current research characteristics and future trajectories in this field. This work aims to provide

**Abbreviations:** DL, deep learning; ML, machine learning; CNNs, convolutional neural networks; DNN, deep neural network; scRNA-Seq, single-cell RNA sequencing; TCGA, The Cancer Genome Atlas; NSCL/P, nonsyndromic cleft lip with or without cleft palate.

resources for possible further intensive exploration and encourages more researchers to overcome the research of deep learning applications in genetics.

#### KEYWORDS

deep learning, machine learning, genetics, bibliometric, knowledge graph

## 1 Introduction

Deep learning (DL) is a subfield of machine learning (ML) that aims to avoid extensive manual processing in traditional methods Wu et al., (2020). Different from machine learning, deep learning is a form of representation learning in which a machine is fed with raw data and develops its own representations needed for pattern recognition—which is composed of multiple layers of representations Esteva et al., (2019). The application of DL in medical healthcare has been widely reported. For example, DL has been reported to be successful in identifying a variety of histopathological features and detecting the biomarkers (Berrar and Dubitzky, 2021). DL has also been applied to predict diagnosis, prognosis, and treatment response in certain cancers Tran et al., (2021). This information could prove valuable in clinical decision-making for cancer treatment and triage for in-depth sequencing.

Genomic data had served as a biomarker for the onset and progression of the disease. Various deep learning applications in genomics had been reported, such as predicting gene expression from genotype data and studying the splicing-code model and the identification of long noncoding RNAs (Tripathi et al., 2016; Xie et al., 2017; Vellido, 2020; Tang et al., 2021). Recent advances in deep learning have emerged in several applications, ranging from natural language to vision processing Zou et al., (2019). Bibliometric approaches have generated a considerable impact on the deep learning research field, such as deep learning networks in identifying medical images and histopathology images for breast cancer classification (Khairi et al., 2021; Wang et al., 2022). However, gaps exist for deep learning in genetics research, and there is a dearth of information on associated bibliometric development trends. Therefore, based on deep learning technology advancements in genetics, a comprehensive bibliometric overview is required to provide researchers with new future research directions.

In the present study, we accessed the Web of Science Core Collection (WoSCC) using bibliometric methods to review and select deep learning studies in genetics research from 2000 to 2021. Specifically, co-word biclustering analysis was utilized to identify the research hot spots of the application of DL in genomics research. We hope this article can provide some reference for future research on deep learning and genomics research.

## 2 Materials and methods

On 22 December 2021, we downloaded data from the Web of Science Core Collection (WoSCC); two authors independently verified citations and retrieved studies. The WoSCC is a frequently used authoritative database for scientific information, from which we generated a clustered network of 69,806 references cited by 1,754 studies. Between the publication years 2000 and 2021, literature searches were performed using the search terms: [TS = (“deep learning” OR “machine learning” OR “convolutional neural network\*” OR CNN\* OR RNN OR “Recurrent neural network\*” OR “Fully Convolutional Network\*” OR FCN\*)], and The literature type = “Article OR Review OR Opening Online”, WoS category = Genetic heredity. Information on the following topics was collected: title, abstract, authors, institution, country/region, journal, keywords, and references. Articles were indexed in the WoSCC and excluded meeting articles, repeated articles, proceedings articles, book chapters, and unpublished documents without enough information for further analysis at the same time.

We described publication characteristics, including institutes, countries, journals, and keywords. The Journal of Citation Reports (JCR, 2021 version) was accessed to identify impact factors that reflected the scientific value of research (Eyre-Walker and Stoletzki, 2013). Retrieved data were analyzed in VOS viewer (Leiden University, Leiden, Holland) and CiteSpace V (Drexel University, Philadelphia,

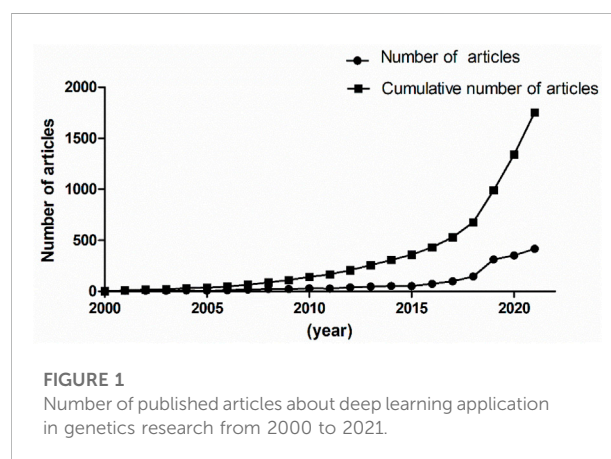
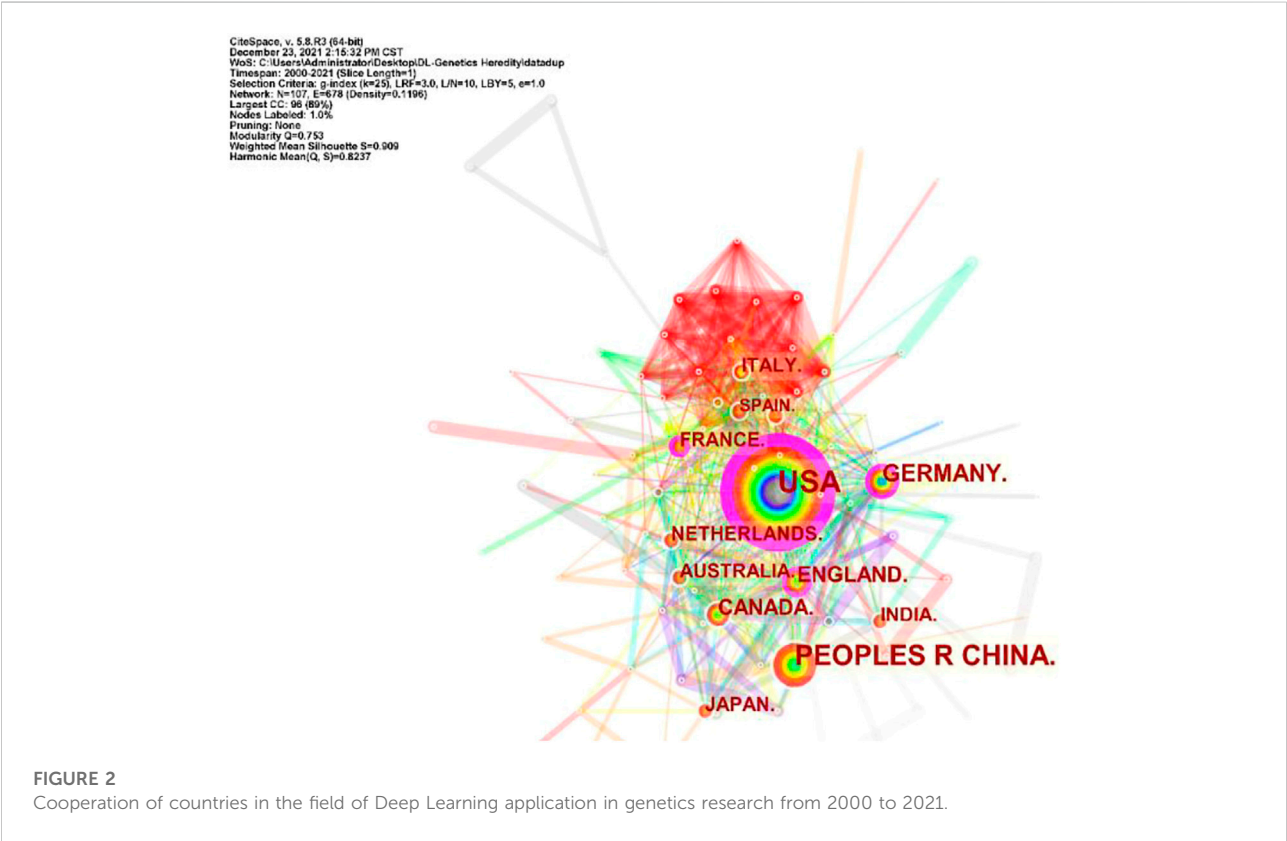


TABLE 1 Top 10 countries, institutions, and journals.

Rank	Country	Count	H-index	Institution	Count	H-index	Cited journal	Count	If (2021)
1	United States	728	65	Chinese Academy of Sciences	41	13	BIOINFORMATICS	1,175	6.93
2	CHINA	407	35	Harvard Medical School	37	12	NATURE	1,082	49.96
3	GERMANY	118	33	Stanford University	29	18	NUCLEIC ACIDS RES	1,075	16.97
4	ENGLAND	101	35	University of Pennsylvania	28	11	P NATL ACAD SCI United States	868	11.20
5	CANADA	92	22	Harvard University	26	25	PLOS ONE	856	3.24
6	AUSTRALIA	54	19	University of Toronto	25	12	SCIENCE	780	47.72
7	INDIA	48	11	Columbia University	25	11	BMC BIOINFORMATICS	764	3.16
8	FRANCE	43	16	Yale University	24	14	NAT GENET	734	38.33
9	ITALY	41	15	University of Washington	24	13	CELL	683	41.58
10	JAPAN	41	16	Shanghai Jiao Tong University	22	9	GENOME BIOL	622	13.58



PA, United States), which facilitated collaborative network analyses connecting different publication characteristics (Chen, 2006; Chen, 2017). From the analysis and measures above, we obtain the current characteristics, research hotspot, subject evolution path, and future trajectories in deep learning applications of genetics.

### 3 Results

#### 3.1 Distribution of articles by publication years

A total of 1,754 articles from 2000 to 2021 were published. As shown in Figure 1, the line with points denoted by square shows



CiteSpace, v. 5.8.R3 (64-bit)  
 December 23, 2021 2:19:26 PM CST  
 WoS: C:\Users\Administrator\Desktop\DL-Genetics Haredity\data\sup  
 Timespan: 2009-2021 (Slice Length=1)  
 Selection Criteria: g-index (k=25), LRF=3.0, L/N=10, LBY=5, e=1.0  
 Network: k=485, L=1396 (Density=0.0119)  
 Largest CC: 342 (70%)  
 Nodes Labeled: 1.0%  
 Pruning: None  
 Modularity Q=0.753  
 Weighted Mean Silhouette S=0.909  
 Harmonic Mean(Q, S)=0.8237

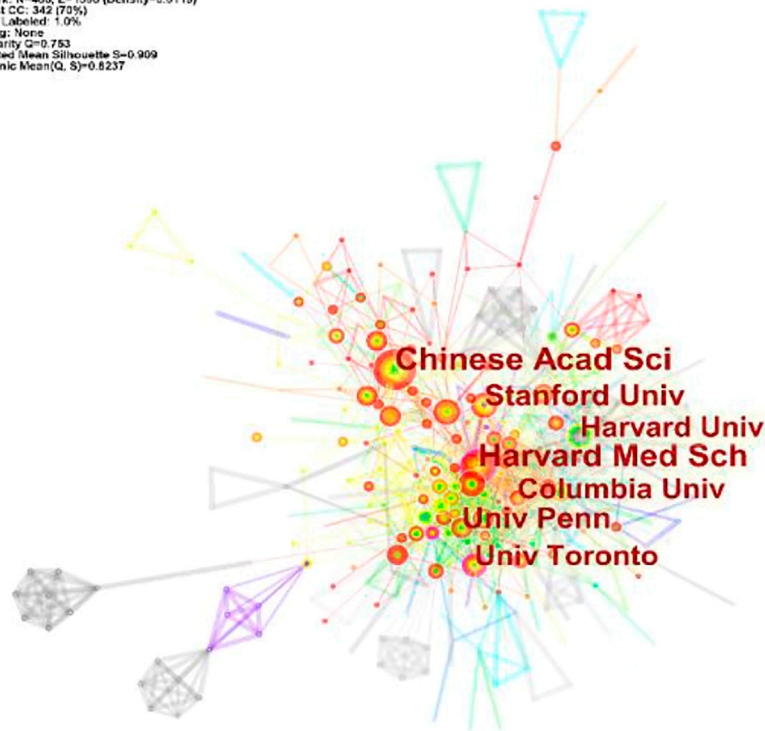


FIGURE 3

Cooperation of institutions contributed to publications for Deep Learning applications in genetic research.

the trend of publications from 2000 to 2021, and the line with points denoted by circles shows the number of articles published each year. The number of published articles showed a rapid increase since 2018, and more than 70% of the total articles were published in the last 4 years. This suggests that the studies of deep learning applied in genetics research were new research hot points in recent years.

### 3.2 Analysis of countries, institutions and journals

A total of 101 countries and 2,487 institutes contributed publications. The top 10 countries, institutions, and cited journals are listed in Table 1. 728 (41.5%) articles published in the United States ranked first place, which was 18.3% higher than those in China, whose publication number was 407 (23.2%), thereby ranking second. However, it is worth noting that the research institution with the largest number (41 articles) of published articles was the Chinese Academy of Sciences, which indicated this institution had powerful scientific research ability in the field of deep learning application in

genetics research. The collaborations between different countries and institutions are shown in Figures 2, 3. The bigger size of the circle represents the larger number of articles published by this country. The shorter the distance between two circles, the better the cooperation between the two countries. As shown in Figure 2, the biggest circle belongs to the United States of America which had close cooperation with Germany, England, and France. Although the Chinese Academy of Sciences published the largest number of articles, it was lack of cooperation with other institutions. Harvard Medical School was in a key position in this study field, which kept close cooperation with multiple institutions, such as Columbia University and Stanford University (shown in Figure 3).

### 3.3 Journal analysis

A total of 151 cited journals published publications related to deep learning in genetics research. The top 10 cited journals are presented in Table 1 (with green background). The highest cited count belonged to the BIOINFORMATICS (1,175 times), followed by NATURE (1,082 times). Among



In academic journals, referential relationships facilitate knowledge exchange within the research field, where citing articles form the knowledge frontier and cited articles form

the knowledge base. A journal dual-map overlay is shown in [Figure 5](#). The cluster analysis of citing articles (the left side) belongs to journals focusing on the field of molecular/biology/immunology research. Also, the cluster analysis of cited articles (the right side) belongs to journals focusing on the field of molecular/biology/genetics research. The primary

TABLE 2 Top 10 cited references on VR in rehabilitation.

Rank	DOI	Title of cited reference	Count	Centrality	Interpretation of the findings	Year
1	10.1038/nature14539	Deep learning	89	0.01	This article discussed deep learning methods such as deep convolutional nets and recurrent nets that have dramatically improved speech and visual recognition. Other domains such as drug discovery and genomics brought about breakthroughs	2015
2	10.1038/nbt.3300	Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning	77	0.08	This study built a stand-alone software by using a diverse array of experimental data and evaluation metrics ascertained sequence specificities that is essential for identifying causal disease variants	2015
3	10.1038/nmeth.3547	Predicting effects of noncoding variants with a deep learning-based sequence model	65	0.07	This document developed a deep learning-based algorithmic framework that enables the prediction of noncoding variants	2015
4	10.1145/2939672.2939785	Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	54	0	This study described a highly effective scalable tree boosting machine learning method and proposes a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning	2016
5	10.1101/gr.200535.115	Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks	47	0.1	This study offered a powerful computational approach to annotating and interpreting the noncoding genome. Researchers perform a single sequencing by CNN's assay to annotate every mutation in the genome with its influence on present accessibility and latent potential for accessibility	2016
6	10.1038/nature19057	Analysis of protein-coding genetic variation in 60,706 humans	44	0.01	This study analysis protein-coding genetic variation in 60,706 humans, and it can efficiently filtering of candidate disease-causing variants and discover human 'knockout' variants in protein-coding genes	2016
7	10.1093/nar/gkw226	DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences	36	0.05	This study proposed a novel hybrid convolutional and bi-directional long short-term memory recurrent neural network framework for predicting non-coding function <i>de novo</i> from the sequence	2014
8	10.1038/nature14248	Integrative analysis of 111 reference human epigenomes	36	0.08	The article described the integrative analysis of 111 reference human epigenomes generated and profiled for histone modification patterns, DNA accessibility, DNA methylation, and RNA expression	2015
9	10.15252/msb.20156651	Deep learning for computational biology	34	0.03	This study reviewed the applications of this new breed of analysis approaches in regulatory genomics and cellular imaging	2014
10	10.1038/ng.2892	A general framework for estimating the relative pathogenicity of human genetic variants	34	0.05	This study discussed a framework that objectively integrates many diverse annotations into a single, quantitative score to differentiate 14.7 million simulated variants	2015

citation path colored orange represents the citation relationship between the two clusters, which indicated that based on genetics research, deep learning tends to be applied to immunology.

3.4 Reference analyses

References are key bibliometric indicators as frequently cited documents can greatly influence their research areas (Table 2). The article was published on Nature which was cited 89 times, ranking first. Summarizing the highly cited topics, the result indicated that deep learning methods such as deep convolutional

nets and recurrent nets have dramatically improved drug discovery and genomics research.

In network research, betweenness centrality is a major indicator to determine the importance of nodes in the network, and a higher betweenness centrality means that the literature is more important Synnvestvedt et al., (2005). Table 2 also shows the betweenness centrality of these works of literature.

In this article, a co-cited document-based clustering analysis can be used to generate sub-fields and connect nodes in the research. We constructed a network of co-cited references to test the scientific relevance of related publications (Figure 6). Cluster setting parameters were

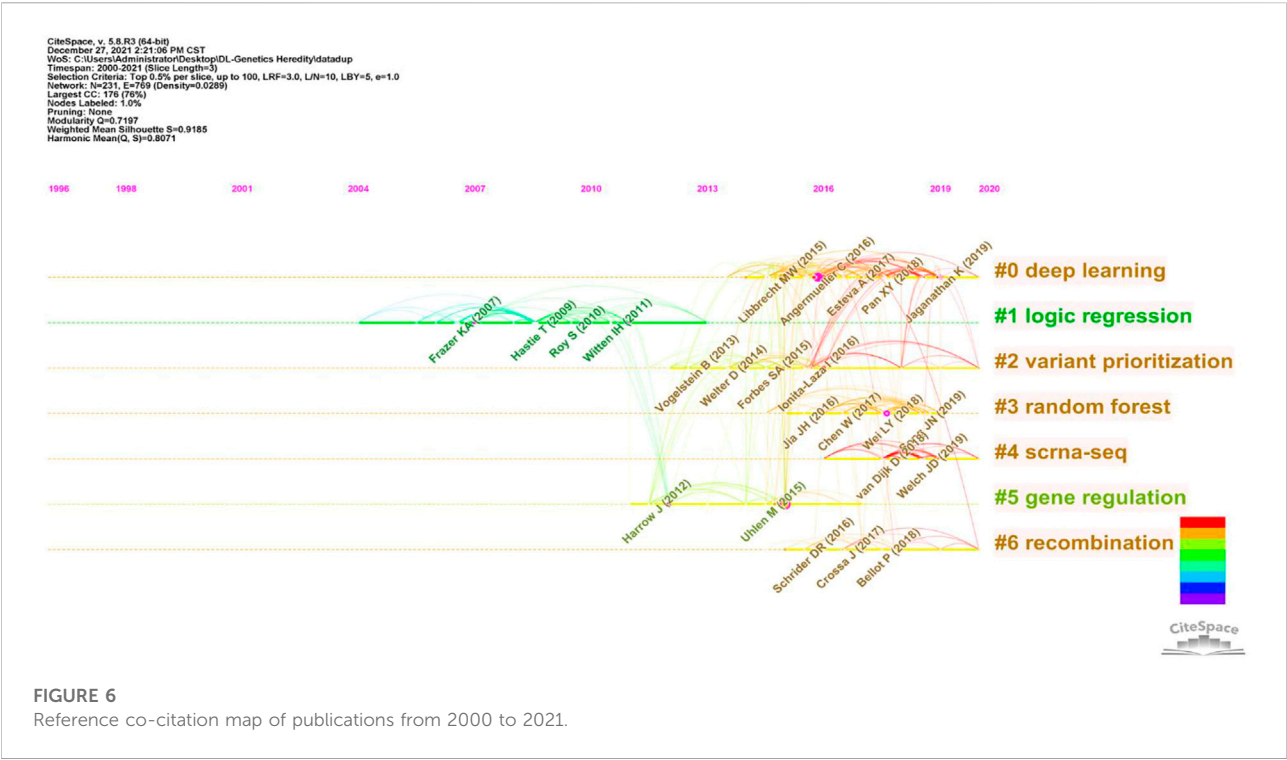


TABLE 3 Highly link strength of the top 20 occurrence keywords.

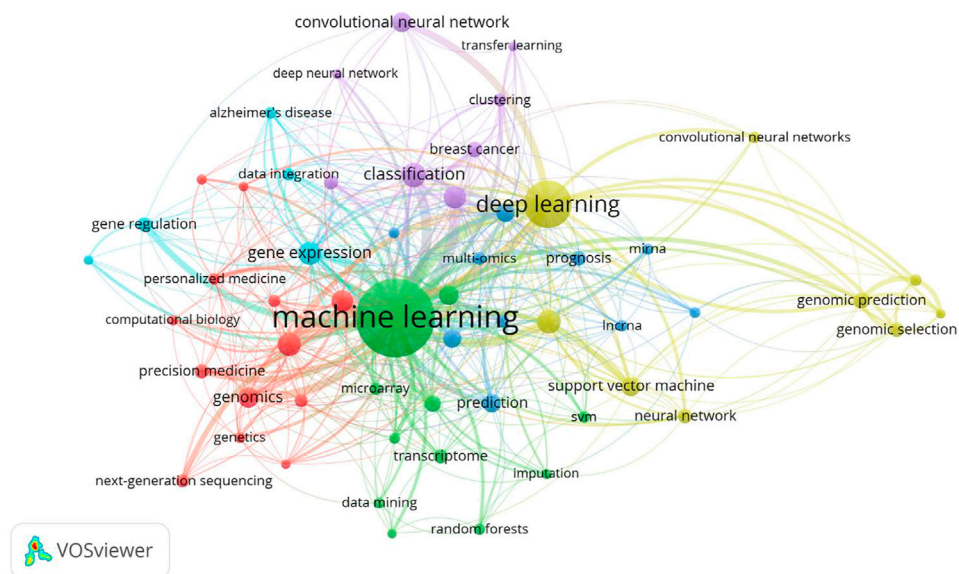
Rank	Keyword	Occurrence	Total link strength	Rank	Keyword	Occurrence	Total link strength
1	Machine learning	553	481	11	DNA methylation	36	44
2	Deep learning	201	194	12	Support vector machine	32	44
3	Classifications	54	90	13	Prediction	31	43
4	Random forest	52	69	14	Biomarker	30	45
5	Bioinformatics	48	68	15	Cancer	30	46
6	Gene expression	48	83	16	Rna-seq	24	38
7	Feature selection	46	80	17	Genomic prediction	23	42
8	Artificial intelligence	43	80	18	Breast cancer	22	32
9	Genomics	37	56	19	Gene regulation	19	26
10	Convolutional neural	36	24	20	Neural network	19	24

Top N% = 0.5, #Years Per Slice = 3, and the pruning algorithm was chosen. The Modularity Q score = 0.7197, which was > 0.5, indicated the network adopted loosely coupled clusters. The Weighted Mean silhouette score = 0.9185, which was > 0.5, indicated acceptable cluster homogeneity. From the literature, we used index items as cluster markers (#0–#6); the largest cluster (#0) was “deep learning”, #1 was “logic regression”, #2 was “variant prioritization”, #3 was “random forests”, #4 was “scrNA-seq”, #5 was “genomic regulation”, and #6 was “recombination”.

### 3.5 Co-occurrence and burst keyword analyses

We extracted and analyzed keyword co-occurrence in related publications. The top 20, with highly linked strengths, are shown in Table 3. The co-occurrence of any two terms indicates their presence in the same publication. While identifying thematics in research areas, keyword analyses of articles 1754) identified 27 keywords with a minimum of 40 occurrences (Figure 7). The co-occurrence analysis based on author keywords was built with occurrence times as a threshold. There are several distinct





**FIGURE 7**  
Network map of keywords is divided into 6 clusters.

 Occurrences Burst History

### Top 19 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2000 - 2021
saccharomyces cerevisiae	2000	3.82	2000	2010	
susceptibility	2000	4.98	2007	2011	
discovery	2000	6.62	2009	2015	
polymorphism	2000	6.63	2010	2013	
evolution	2000	5.97	2010	2017	
signature	2000	3.55	2010	2015	
database	2000	8.41	2012	2015	
gene expression	2000	5.37	2012	2016	
reveal	2000	5.21	2012	2018	
network	2000	7.62	2013	2018	
protein	2000	7.3	2013	2016	
prediction	2000	7.4	2016	2021	
dna	2000	7.09	2016	2017	
genome wide association	2000	7.01	2016	2017	
sequence	2000	8.52	2017	2021	
mutation	2000	6.22	2017	2021	
selection	2000	11.59	2018	2019	
association	2000	6.1	2018	2019	
cancer	2000	9.77	2019	2021	

**FIGURE 8**  
Keywords with the strongest citation bursts of publications from 2000 to 2021.



clusters with different colors. The co-occurrence network map of keywords is shown in Figure 7, given that the larger the size of the circle, the higher the co-occurrence of keywords. Furthermore, having closer keywords together shows a stronger relationship. The average year of publication of the keywords was determined using colors. Machine learning, deep learning, and genetics constitute the largest circle of all keywords that are identified through co-occurrence analysis. Our study also investigated temporal trends in hotspot shifts using the top 19 keywords having the strongest citation bursts. These included prediction (2016–2021), sequence (2017–2021), mutation (2017–2021), and cancer (2019–2021) (Figure 8).

## 4 Discussion

### 4.1 General data

Between the publication years 2000–2021, we selected and investigated 1754 SCIE articles related to deep learning in genetics. Since 2015 with the development of gene sequencing technology, biological genetic data have exploded. The number of published articles showed a rapid increase. Another growth time node is 2018, more than 70% of the total articles were published in the last 4 years. This suggests that as deep learning technology enters its mature stage, it has attracted widespread attention. The highest number of studies (728) was generated by the United States, with China in second place at 407. The top ten institutions included seven in the United States and two in China. According to our data, most of the research in deep learning in genetics was produced by institutions and countries in developed countries, such as United States, Germany, and France. The reason for this trend is that better socioeconomic development can be the premise of ensuring adequate funding, resources, and human input to explore brand new scientific research. Socioeconomic factors such as GDP, GDP per capita, research and development funding, number of researchers, number of physicians, or international collaboration are important decisive factors of scientific productivity. There are many reasons for this trend, such as GDP level Nature Genetics was the most frequently used publishing journal; therefore, it significantly contributed to research in this area. Additionally, we investigated the top 10 cited publications; the top-cited article was published by Lec et al. on Nature and was cited 89 times. These high cited articles will shed some light to this research field.

### 4.2 The knowledge base and current research characteristics

In previous studies, different deep learning research applications have been investigated in genetics and generated significant results. As indicated (Figure 6), after

clustering co-cited references, key clustering nodes successfully identified knowledge bases, namely: #0 “deep learning”, #1 “logic regression”, #2 was marked as “variant prioritization”, #3 was marked as “random forests”, cluster #4 was marked as “scRNA-seq”, cluster #5 was marked as “genomic regulation”, and cluster #6 was marked as “recombination”. We described the knowledge base according to different clusters with time characteristics.

In the #0 “deep learning” cluster, applications of deep learning methods show cutting-edge performance in a variety of complex prediction tasks and large datasets in natural images. Scientists propose a deep-learning framework for genetic research events, e.g., distant metastasis in cancer, protein subcellular localization, genome recombination map of African *Drosophila melanogaster*, and DNA transcription factor binding; the abovementioned aspects show the advantages (Xiao et al., 2019; Adrion et al., 2020; Zhang et al., 2021a; Chereda et al., 2021).

In the #1 “logic regression” cluster, Liu et al. proposed a logic regression-based approach that was used to analyze the gene–gene interaction of eight genes involved in cell adhesion in 806 NSCL/P Chinese case-parent triad recruited to explore the risk of non-syndromic cleft lip Liu et al., (2019). Nicodemus KK’s team tested and discussed the interactions between these susceptibility genes using four machine learning algorithms (including random forest, generalized enhanced regression, and Monte Carlo logistic regression) in a case-control study of schizophrenia Nicodemus et al., (2010). Dasgupta et al. reviewed machine learning and regression-based methods in 200 common or rare genetic variants from exome sequencing data and discussed cross-validation for model assessment and selection Dasgupta et al., (2011).

In the #2 “variant prioritization” cluster, key challenges in genomics research are variant prioritization methods. Huang et al. identified a deep learning framework, which was evolution-based, for unified variant and gene prioritization. The authors integrated constraints predicting missense variants and protein-coding genes associated with dominant disorders and estimated fitness effects for potential single-nucleotide variants, which outperformed current methods Huang, (2020). Zhang et al. formulated a disease-specific variant classifier that assessed discriminate pathogenic variants from benign variants and prioritized disease-associated variants Zhang et al., (2021b). In their study, Mattia et al. proposed an automated computational framework that identified causal genetic variants (small insertions and deletions and coding/splicing single-nucleotide variants) to improve causal variant prioritization methods and variant pathogenicity classifications Bosio et al., (2019).

For the #3 “random forest” cluster, as a standard regression model, which has been widely used in the machine learning (ML) application, Jian Y’s team applied a random forest machine learning algorithm to purity pediatric children central nervous system tumor analysis, which helps with the clinical management

of pediatrics Yang et al., (2021). Chen Z constructed a deep learning network model based on the random forest classifier, and it can easily identify malonylation sites, for predicting sites shows high confidence Chen et al., (2018). Nicholls et al. reviewed ML model (gradient boosting and random forests) applications, dissected variant and gene signal heterogeneity, prioritized complex disease-associated loci, and critically evaluated prioritization issues for genome-wide association investigations Nicholls et al., (2020).

For the #4 “scRNA-seq” cluster, as single-cell RNA-sequencing (scRNA-seq) is used to analyze gene expression with high resolution, scientists have comprehensively exploited this area to dissect individual cell types in several diseases. For example, Carlos et al. generated a Deep Neural Network (DNN) model which quantified immune infiltration levels in breast and colorectal cancer bulk RNA-seq samples and identified improved and accurate survival prediction and quantification data (Torroja and Sanchez-CaboDigitalDisorder, 2019). Cédric et al., in an effort to accommodate increasing levels of scRNA-seq data, designed a deep neural network-based imputation algorithm that is more suitable for the ever-increasing scRNA-seq data (Arisdakessian et al., 2019). Additionally, Yao and Nelson’s teams generated unsupervised deep learning methods for improved data integration which showed improved performances in scRNA-seq datasets (Johansen and Quon, 2019; He et al., 2020).

For the #5 “gene regulation” cluster, an ML modality was adapted by Colbran et al. to impute gene regulation information from genotype data and investigate 490 ancient Eurasian human DNA samples and explore divergent gene regulation mechanisms which contributed to skin pigmentation and metabolic and immune functions. The authors identified gene regulation roles in adaptation and associations between complex traits and genetic diversity Colbran et al., (2021). Atak et al. devised a deep learning approach and integrative genomics strategy to analyze functional enhancer mutations with allelic imbalance of gene expression and chromatin accessibility and successfully interpreted and predicted the impact of a mutation on gene regulation Atak et al., (2021). Godwin et al. devised a deep learning-based model to predict the gene regulatory effects of low-molecular-weight compounds; the model potentially identified drug candidates inducing particular gene responses, without prior interactional information on protein targets Woo et al., (2020).

For the #6 “recombination” cluster, a central tenet of genomics is the accurate assessment of genome-wide recombination rates in natural populations. Andrew et al. used ML algorithms to examine if DNA motifs across the genome could be used to predict crossover variation and identify genetic factors influencing variation in recombination rates Adrian et al., (2016). Kha F proposed a DL intelligent computational predictor based on the deep neural network (DNN) as a classification engine for the identification of

recombination spots through an experimental benchmark dataset with 10-fold cross-validation which achieved the 95.81% highest accuracy Khan et al., (2020).

## 4.3 Hotspots and frontiers in research

Keywords concentrate on contemporary research issues or concepts, while burst keywords represent emerging trends and frontiers in research. In our work, we used CiteSpace to capture burst keywords, and four related research frontiers were identified: four keywords with the strongest citation bursts, such as prediction (2016–2021), sequence (2017–2021), mutation (2017–2021), cancer (2019–2021), and these keywords cover the research frontier of the current topic.

### 4.3.1 Sequence (2017–2021)

Large-scale genetic datasets and deep-learning approaches are increasingly exploited by bioinformatics approaches to model protein structures and complexes. Zhao et al., using sequence information, devised a deep forest-based protein location algorithm to accurately predict protein subcellular locations using only protein sequences, which outperformed contemporary state-of-art algorithms Zhao et al., (2018). Cui et al. analyzed the main methods used to represent protein sequence data, theoretically reviewed the architecture of different embedding models, and investigated the development of these sequence-embedding approaches Cui et al., (2021). Braberg et al. analyzed the emergence of large-scale genetic datasets and deep learning approaches which modeled protein structures and associated interactions (deep mutational scanning, genome-scale genetic or chemical-genetic interaction mapping, and coevolution) and discussed structural data integration from different sources Braberg et al., (2022).

### 4.3.2 Cancer (2019–2021)

Originally used for image processing and pattern recognition methods, deep learning models are now used to detect genetic alterations in cancer and determine cancer patient prognoses. The framework by Mallik et al. integrated linear regression, differential expression, and deep learning and facilitated the robust interpretation of DNA methylation signatures and gene expression data for cervical cancer Mallik et al., (2020). Poirion et al., using multi-omics data, established a deep learning ensemble network that predicted patient survival subtypes Poirion et al., (2021). In order to predict survival outcomes in cancer patients, Huang et al. broadly analyzed The Cancer Genome Atlas cancers using several deep learning-based models Huang et al., (2020). Tran et al. reviewed emerging deep learning approaches and how they were applied to precision oncology. The authors not only exemplified how deep learning was used for cancer diagnostics, prognostics, and treatment management strategies, but they also reviewed

the current limitations and challenges of deep learning in this area [Vaernet, \(1972\)](#).

#### 4.3.3 Mutation (2017–2021)

With considerable high-throughput technology advancements, somatic mutations in their millions have been reported, but critically, the identification of specific driver genes expressing oncogenic mutations is highly challenging and complex. In their study, Luo et al. used “deep drive” to predict driver genes by combining similarity networks with features that characterize the functional impact of mutations. They use AUC scores to evaluate predictive efficiency. DeepDriver achieved AUC scores of 0.984 and 0.976 on breast cancer and colorectal cancer, respectively, which were better than those of the competing algorithms. [Luo et al., \(2019\)](#) Sahraeian et al. inaugurated a deep convolutional neural network-based somatic mutation detection strategy using high-confidence somatic mutations in a cancer cell line. The authors generated comprehensive models using multiple datasets and highly robust and significantly superior methods when compared with traditional detection strategies [Sahraeian et al., \(2022\)](#).

#### 4.3.4 Prediction (2016–2021)

Ding YL provided a comprehensive review of ML-based approaches for predicting disease–biomolecule associations with multi-view data sources. They discussed feature representation methods and provided some perspectives for further improving biomolecule–disease prediction methods ([Ding et al., 2021](#)). Groschel MI presented a translational genomics platform for tuberculosis application to predict antibiotic resistance from next-generation sequence data. After benchmarking, it can rapidly and accurately predict resistance to anti-tuberculosis drugs [Gröschel et al., \(2021\)](#). Majumdar A et al. developed a novelty ensemble support vector regression to predict each drug response value for a single patient based on cell-line gene expression data. This can be used to develop a robust drug response prediction system for cancer patients using cancer cell lines guidance and multi-omics data ([Majumdar et al., 2021](#)).

## 5 Limitations

Our study still has some limitations to be addressed. First, we choose the SCIE database as the collection, while a few studies not included in the core collection were missed. Second, this study includes two types of publication (article and review), and the uneven quality of the collected publications may reduce the credibility of the mapping analysis. However, the visualized analysis based on bibliometric analysis undoubtedly lays a foundation for readers to quickly understand the research

subjects, hotspots, and development trends in an unfamiliar research field.

## 6 Conclusions

Using bibliometrics, we systematically, comprehensively, and objectively investigated the literature related to deep learning applications in genetics research. Importantly, we identified research bases, current hotspots, and future trends in this area. The knowledge bases were “deep learning,” logic regression,” “variant prioritization,” “random forests,” “scRNA-seq,” “genomic regulation,” and “recombination”. We also provided hotspot and frontier guidance for researchers wishing to conduct advanced genetics research in the future. We identified research frontiers and emerging trends topics that incorporated prediction, sequence, mutation, and cancer. Finally, some studies selected for this research were not comprehensive and may have generated publication bias, thereby potentially affecting the study outcomes of this bibliometric review.

## 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

Author TF and BZ make the conception and design of this work. Author BZ was responsible for making the acquisition, analysis, interpretation of the data, and drafting the manuscript. TF designed the study and revised the article. All the authors read and approved the final manuscript.

## Funding

This study was supported by the 345 Talent Project of Shengjing Hospital (M0347) and “Research on the application of genetic big data analysis in the prevention and control of newborn birth defects” (2900021013-CMU-013).

## 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

## References

- Adrian, A. B., Corchado, J. C., and Comeron, J. M. (2016). Predictive models of recombination rate variation across the *Drosophila melanogaster* genome. *Genome Biol. Evol.* 8, 2597–2612. doi:10.1093/gbe/evw181
- Adrian, J. R., Galloway, J. G., and Kern, A. D. (2020). Predicting the landscape of recombination using deep learning. *Mol. Biol. Evol.* 37, 1790–1808. doi:10.1093/molbev/msaa038
- Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., and Garmire, L. X. (2019). DeepImpute: An accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biol.* 20, 211. doi:10.1186/s13059-019-1837-6
- Atak, Z. K., Taskiran, I. I., Demeulemeester, J., Flerin, C., Mauduit, D., Minnoye, L., et al. (2021). Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning. *Genome Res.* 31, 1082–1096. doi:10.1101/gr.260851.120
- Berrar, D., and Dubitzky, W. (2021). Deep learning in bioinformatics and biomedicine. *Brief. Bioinform.* 22 (2), 1513–1514. doi:10.1093/bib/bbab087
- Bosio, M., Drechsel, O., Rahman, R., Muiyas, F., Rabionet, R., Bezdan, D., et al. (2019). eDiVA-Classification and prioritization of pathogenic variants for clinical diagnostics. *Hum. Mutat.* 40, 865–878. doi:10.1002/humu.23772
- Braberg, H., Echeverria, I., Kaake, R. M., Sali, A., and Kroger, N. J. (2022). From systems to structure - using genetic data to model protein structures. *Nat. Rev. Genet.* 23 (6), 342–354. doi:10.1038/s41576-021-00441-w
- Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *J. Am. Soc. Inf. Sci. Technol.* 57 (3), 359–377. doi:10.1002/asi.20317
- Chen, C. (2017). Science mapping: A systematic review of the literature. *J. Data Inf. Sci.* 2, 1–40. doi:10.1515/jdis-2017-0006
- Chen, Z., He, N., Huang, Y., Qin, W. T., Liu, X., Li, L., et al. (2018). Integration of A Deep learning classifier with A random forest approach for predicting malonylation sites. *Genomics Proteomics Bioinforma.* 16, 451–459. doi:10.1016/j.gpb.2018.08.004
- Chereda, H., Bleckmann, A., Menck, K., Perera-Bel, J., Stegmaier, P., Auer, F., et al. (2021). Explaining decisions of graph convolutional neural networks: Patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Med.* 13, 42. doi:10.1186/s13073-021-00845-7
- Colbran, L. L., Johnson, M. R., Mathieson, I., and Capra, J. A. (2021). Tracing the evolution of human gene regulation and its association with shifts in environment. *Genome Biol. Evol.* 13 (11), evab237. doi:10.1093/gbe/evab237
- Cui, F., Zhang, Z., and Zou, Q. (2021). Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. *Brief. Funct. Genomics* 20, 61–73. doi:10.1093/bfpg/ela030
- Dasgupta, A., Sun, Y. V., König, I. R., Bailey-Wilson, J. E., and Malley, J. D. (2011). Brief review of regression-based and machine learning methods in genetic epidemiology: The genetic analysis workshop 17 experience. *Genet. Epidemiol.* 35 (1), S5–S11. doi:10.1002/gepi.20642
- Ding, Y., Lei, X., Liao, B., and Wu, F. X. (2021). Machine learning approaches for predicting biomolecule-disease associations. *Brief. Funct. Genomics* 20 (4), 273–287. doi:10.1093/bfpg/elab002
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25 (1), 24–29. doi:10.1038/s41591-018-0316-z
- Eyre-Walker, A., and Stoletzki, N. (2013). The assessment of science: The relative merits of post-publication review, the impact factor, and the number of citations. *PLoS Biol.* 11, e1001675. doi:10.1371/journal.pbio.1001675
- Gröschel, M. I., Owens, M., Freschi, L., Vargas, R., Jr, Marin, M. G., Phelan, J., et al. (2021). GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome Med.* 13 (1), 138. doi:10.1186/s13073-021-00953-4
- He, Y., Yuan, H., Wu, C., and Xie, Z. (2020). DISC: A highly scalable and accurate inference of gene expression and structure for single-cell transcriptomes using semi-supervised deep learning. *Genome Biol.* 21, 170. doi:10.1186/s13059-020-02083-3
- Huang, Y. F. (2020). Unified inference of missense variant effects and gene constraints in the human genome. *PLoS Genet.* 16, e1008922. doi:10.1371/journal.pgen.1008922
- 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, 41. doi:10.1186/s12920-020-0686-1
- Johansen, N., and Quon, G. (2019). scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol.* 20, 166. doi:10.1186/s13059-019-1766-4
- Khairi, S. S. M., Bakar, M. A. A., Alias, M. A., Bakar, S. A., Liong, C.-Y., Rosli, N., et al. (2021). Deep learning on histopathology images for breast cancer classification: A bibliometric analysis. *Healthcare* 10, 10. doi:10.3390/healthcare10010010
- Khan, F., Khan, M., Iqbal, N., Khan, S., Muhammad Khan, D., Khan, A., et al. (2020). Prediction of recombination spots using novel hybrid feature extraction method via deep learning approach. *Front. Genet.* 11, 539227. doi:10.3389/fgene.2020.539227
- Liu, D., Wang, M., Yuan, Y., Schwender, H., Wang, H., Wang, P., et al. (2019). Gene-gene interaction among cell adhesion genes and risk of nonsyndromic cleft lip with or without cleft palate in Chinese case-parent trios. *Mol. Genet. Genomic Med.* 7, e00872. doi:10.1002/mgg3.872
- Luo, P., Ding, Y., Lei, X., and Wu, F. X. (2019). deepDriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Front. Genet.* 10, 13. doi:10.3389/fgene.2019.00013
- Majumdar, A., Liu, Y., Lu, Y., Wu, S., and Cheng, L. (2021). kESVR: An ensemble model for drug response prediction in precision medicine using cancer cell lines gene expression. *Genes* 12 (6), 844. doi:10.3390/genes12060844
- Mallik, S., Seth, S., Bhadra, T., and Zhao, Z. (2020). A linear regression and deep learning approach for detecting reliable genetic alterations in cancer using DNA methylation and gene expression data. *Genes (Basel)*, 11: 931. doi:10.3390/genes11080931
- Nicholls, H. L., John, C. R., Watson, D. S., Munroe, P. B., Barnes, M. R., Cabrera, C. P., et al. (2020). Reaching the end-game for GWAS: Machine learning approaches for the prioritization of complex disease loci. *Front. Genet.* 11, 350. doi:10.3389/fgene.2020.00350
- Nicodemus, K. K., Callicott, J. H., Higier, R. G., Luna, A., Nixon, D. C., Lipska, B. K., et al. (2010). Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: Biological validation with functional neuroimaging. *Hum. Genet.* 127, 441–452. doi:10.1007/s00439-009-0782-y
- Poirion, O. B., Jing, Z., Chaudhary, K., Huang, S., and Garmire, L. X. (2021). DeepProg: An ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med.* 13, 112. doi:10.1186/s13073-021-00930-x
- Sahraeian, S., Fang, L. T., Karagiannis, K., Moos, M., Smith, S., Santana-Quintero, L., et al. (2022). Achieving robust somatic mutation detection with deep learning models derived from reference data sets of a cancer sample. *Genome Biol.* 23 (1), 12. doi:10.1186/s13059-021-02592-9
- Synnestvedt, M. B., Chen, C., and Holmes, J. H. (2005). CiteSpace II: Visualization and knowledge discovery in bibliographic databases. *AMIA Annu. Symp. Proc.* 2005, 724–728.
- Tang, X., Shi, Z., and Jin, M. (2021). Multi-category multi-state information ensemble-based classification method for precise diagnosis of three cancers. *Neural Comput. Appl.* 33, 15901–15917. doi:10.1007/s00521-021-06211-3
- Torroja, C., Sanchez-Caboand Digitaldlsorter, F. (2019). Digitaldlsorter: Deep-Learning on scRNA-seq to deconvolute gene expression data. *Front. Genet.* 10, 978. doi:10.3389/fgene.2019.00978
- Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V., Waddell, N., et al. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med.* 13 (1), 152. doi:10.1186/s13073-021-00968-x

- Tripathi, R., Patel, S., Kumari, V., Chakraborty, P., and Varadwaj, P. K. (2016). DeepLNC, a long non-coding RNA prediction tool using deep neural network. *Netw. Model. Anal. Health Inf. Bioinforma.* 5, 21. doi:10.1007/s13721-016-0129-2
- Vaernet, K. (1972). Stereotaxic amygdalotomy in temporal lobe epilepsy. *Stereotact. Funct. Neurosurg.* 34, 176–183. doi:10.1159/000103055
- Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.* 32, 18069–18083. doi:10.1007/s00521-019-04051-w
- Wang, L., Wang, H., Huang, Y., Yan, B., Chang, Z., and Liu, Z. (2022) Trends in the application of deep learning networks in medical image. *Eur. J. Radiology* 146, 110069. doi:10.1016/j.ejrad.2021.110069
- Woo, G., Fernandez, M., Hsing, M., Lack, N. A., Cavga, A. D., Cherkasov, A. DeepC. O. P., et al. (2020). DeepCOP: Deep learning-based approach to predict gene regulating effects of small molecules. *Bioinformatics* 36, 813–818. doi:10.1093/bioinformatics/btz645
- Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., et al. (2020). Deep learning in clinical natural language processing: A methodical review. *J. Am. Med. Inf. Assoc.* 27 (3), 457–470. doi:10.1093/jamia/ocz200
- Xiao, M., Shen, X., and Pan, W. (2019). Application of deep convolutional neural networks in classification of protein subcellular localization with microscopy images. *Genet. Epidemiol.* 43, 330–341. doi:10.1002/gepi.22182
- Xie, R., Wen, J., Quitadamo, A., Cheng, J., and Shi, X. (2017). A deep auto-encoder model for gene expression prediction. *BMC Genomics* 18 (9), 845. doi:10.1186/s12864-017-4226-0
- Yang, J., Wang, J., Tian, S., Wang, Q., Zhao, Y., Wang, B., et al. (2021). An integrated analysis of tumor purity of common central nervous system tumors in children based on machine learning methods. *Front. Genet.* 12, 707802. doi:10.3389/fgene.2021.707802
- Zhang, X., Walsh, R., Whiffin, N., Buchan, R., Midwinter, W., Wilk, A., et al. (2021). Disease-specific variant pathogenicity prediction significantly improves variant interpretation in inherited cardiac conditions. *Genet. Med.* 23, 69–79. doi:10.1038/s41436-020-00972-3
- Zhang, Y., Mo, Q., Xue, L., and Luo, J. (2021). Evaluation of deep learning approaches for modeling transcription factor sequence specificity. *Genomics* 113, 3774–3781. doi:10.1016/j.ygeno.2021.09.009
- Zhao, L., Wang, J., Nabil, M. M., and Zhang, J. (2018). Deep forest-based prediction of protein subcellular localization. *Curr. Gene Ther.* 18, 268–274. doi:10.2174/1566523218666180913110949
- Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., Telenti, A., et al. (2019). A primer on deep learning in genomics. *Nat. Genet.* 51, 12–18. doi:10.1038/s41588-018-0295-5





## OPEN ACCESS

## EDITED BY

Deepak Kumar Jain,  
Chongqing University of Posts and  
Telecommunications, China

## REVIEWED BY

Guojian Lin,  
Fujian University of Traditional Chinese  
Medicine, China  
Xiaojun Li,  
Nanjing Xiaozhuang University, China

## \*CORRESPONDENCE

Min Song,  
18401195@masu.edu.cn  
Wei Wang,  
wangwei@gszy.edu.cn

## SPECIALTY SECTION

This article was submitted to Human  
and Medical Genomics,  
a section of the journal  
Frontiers in Genetics

RECEIVED 11 May 2022

ACCEPTED 20 July 2022

PUBLISHED 29 September 2022

## CITATION

Song Z, Zhang H, Jiang Y, Zhao R, Pei X,  
Ning H, Chen H, Pan J, Gong Y, Song M  
and Wang W (2022), Study on  
complications of osteoporosis based on  
network pharmacology.  
*Front. Genet.* 13:941098.  
doi: 10.3389/fgene.2022.941098

## COPYRIGHT

© 2022 Song, Zhang, Jiang, Zhao, Pei,  
Ning, Chen, Pan, Gong, Song and Wang.  
This is an open-access article  
distributed under the terms of the  
[Creative Commons Attribution License](#)  
(CC BY). The use, distribution or  
reproduction in other forums is  
permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original  
publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or  
reproduction is permitted which does  
not comply with these terms.

# Study on complications of osteoporosis based on network pharmacology

Zhijing Song<sup>1,2,3</sup>, Haoling Zhang<sup>4</sup>, Yuhang Jiang<sup>5</sup>, Rui Zhao<sup>1</sup>,  
Xuedong Pei<sup>1</sup>, Haochi Ning<sup>1</sup>, Hailiang Chen<sup>1</sup>, Jing Pan<sup>1</sup>,  
Yanlong Gong<sup>2</sup>, Min Song<sup>1\*</sup> and Wei Wang<sup>6\*</sup>

<sup>1</sup>Clinical College of Chinese Medicine, Gansu University of Chinese Medicine, Lanzhou, Gansu, China, <sup>2</sup>Affiliated Hospital of Gansu University of Chinese Medicine, Lanzhou, Gansu, China, <sup>3</sup>Key Laboratory of Dunhuang Medicine, Ministry of Education, Lanzhou, Gansu, China, <sup>4</sup>St Petersburg State University, St. Petersburg, Russia, <sup>5</sup>School of Public Health, Gansu University of Chinese Medicine, Lanzhou, Gansu, China, <sup>6</sup>Gansu University of Chinese Medicine College of Acupuncture-Moxibustion and Tuijin, Lanzhou, Gansu, China

Osteoporosis is a serious threat to human life. Guben Zenggu Granule is an empirical prescription for clinical treatment of osteoporosis. MC3T3-E1 cells are mouse osteogenic precursor cells with osteogenic differentiation, and are classic cells for studying bone metabolism and osteogenic mechanism, as well as mechanical stimulation sensitive cells. Therefore, it can be inferred that Guben Zenggu granule can repair MC3T3-E1 cells under continuous static pressure overload. This study aims to through the network of pharmacology and gene sequencing method, reveal the increase in bone particles under the condition of continuous static pressure overload on osteogenesis mechanism of MC3T3-E1 cells. In the process of analysis, from a variety of 98 compounds was predicted in the database, a collection of 474 goals, a total of 29,164 difference between two groups of genes. Then, construction of composite targets between cells and predict targets and protein - protein interaction networks, and through the cluster analysis to further explore the relationship between the target. In addition, linkages between target proteins and cells were further identified using Gene Ontology (GO) and Pathways (KEGG Pathway). Finally, the repair effect of Guben Zenggu granule on MC3T3-E1 cells under continuous static pressure overload was verified through experiments, so as to accurately explain the pharmacodynamic mechanism of Traditional Chinese medicine.

## KEYWORDS

network pharmacology, gene sequencing, guben zenggu granules, continuous static pressure, MC3T3-E1 cells

## Introduction

Osteoporosis (OP) is a common disease and frequently-occurring disorder of bone metabolism. The main clinical manifestations are bone loss and systemic chronic pain, and the main complications are brittle fracture (Gallagher et al., 1994; Gali, 2001; Zhu et al., 2021). In recent years, with the aging of the society, the disability rate caused by OP has increased significantly. The NUMBER of osteoporotic fractures is expected to rise to 4.5 million a year, according to a European Union study (Kanis et al., 2019). The number is expected to reach 18 million worldwide by 2040 (Yaacobi et al., 2017).

Guben Zenggu granule is professor Song Min's experience prescription in clinical treatment of osteoporosis. It is mainly composed of astragalus membranaceus, codonopsis, angelica, epimedium, cistanche deserticola, rehmannia glutinosa, psoraleae, turtle worm, dog ridge, aconite, antler gum, and other drugs. The compatibility of Junchen Decoction with Traditional Chinese medicine has multiple targets and multiple effects in the treatment of osteoporosis, which is more scientific (Li et al., 2020; Ai et al., 2020). In the treatment of osteoporosis, TCM should focus more on compound studies, integrate the manifestations of TCM syndrome elements, summarize the characteristics of symptoms, give play to the compatibility advantages of compound therapy of king, minister and assistant and syndrome differentiation, and provide microcosmic material basis and support for the theory of "kidney main bone" (Zhao et al., 2021a; Cui et al., 2018; Zhang et al., 2016). Preliminary clinical studies have shown that Guben Zenggu granule can increase bone mineral density and effectively improve patients with osteoporosis pain, with definite clinical efficacy. Basic studies have also shown that Guben Zeng gu granule can promote the osteogenic differentiation of BMSCs and increase the content of BGP, OPN, ALP, and COLI proteins, possibly through the activation of BMPsmd/RUNX2 signaling pathway (Song, 2020a), (Song, 2020b). Guben Zenggu granule can reduce BGP and Trap-5B contents in serum and free  $[Ca^{2+}]$  I concentration in bone in ovariectomized rats, thus regulating bone mineral density and stimulating biomechanical properties of bone tissue (Song, 2020c). Guben Zenggu granule and hyperbaric oxygen in the synergistic treatment of osteoporosis rats can effectively promote the balance between osteogenesis and osteofragmentation and enhance the activity of bone microstructure through the intervention and regulation of OPG/RANKL signaling pathway (Feng et al., 2021).

Network pharmacology is a new method of pharmacological research (Li et al., 2017). It can identify and predict its related targets, bioactive compounds, and clarify the molecular mechanism of TCM. The core concept of TCM network pharmacology is "network target, multi-component" model, which can systematically elucidate the molecular mechanism of TCM treatment of various diseases (Xu et al., 2021; He

et al., 2021). At present, there have been many pharmacological studies on network, such as gegen Qinlian Decoction for the treatment of type 2 diabetes, Wumei pill for the treatment of pancreatic tumor, and Baizhu root for the treatment of osteoporosis (Li et al., 2014; Wan et al., 2019; Zhang et al., 2019). Molecular docking technology is a new research method of computer-aided medicine, can through the computer to study the interaction between drug and target genes. Through the interaction between ligand and receptor, the binding pattern and affinity between ligand and receptor are predicted by computer data (Qiao, 2015). In this study, using computer-aided drug research method, the differentially expressed genes in MC3T3-E1 cells were obtained as the target of continuous static pressure overload intervention in MC3T3-E1 cells, and the effective components and pharmacological mechanism of Guben Zenggu granule in treating OP were studied. The experimental design route of this paper is shown in Figure 1.

## Experimental equipment

### Experimental cells

MC3T3-E1 cells were purchased from Wuhan Punosai Life Science and Technology Co., LTD. (Article number CL-0378).

### Experimental software

TCMSP database (<http://tcmbsp.com/tcmbsp.php>) (Ru et al., 2014), Batman database (<http://bionet.ncpsb.org/batman-tcm/>) (Liu et al., 2016), the STRING database (<https://string-db.org/>) (Szklarczyk et al., 2019), Venny2.1 online software mapping tools (tools/venny/<https://bioinfogp.cnb.csic.es/platform>), David database (<https://david.ncicrf.gov/>) Uniprot database (<https://www.uniprot.org/>) (Soudy et al., 2020), Cytoscape3.7.2 software, R3.6.1 software, etc.

## Main reagents and instruments

Micropipette (Eppendorf), Electrophoresis instrument power supply (Beijing Liu yi Instrument Factory), vertical electrophoresis tank (Beijing Liu yi Instrument Factory), electric rotary instrument (Beijing Liu Yi Instrument Factory), horizontal shaker (Jiangsu Haimen Qi Lin Bell Instrument Manufacturing Co., Ltd), pH meter (Mettler-Toledo GmbH), Electronic balance (Beijing Sartoris Instrument System Co., Ltd.), magnetic stirrer (Zhongda Instrument Factory, Jintan city, Jiangsu Province), Enzyme labeler (Thermo) centrifuge (Hunan Xiangyi Laboratory Instrument Development Co., Ltd.) phosphatase inhibitor (Biyuntian), PMSF (Alding), RIPA lysis fluid (Biyuntian), BCA Protein Concentration Determination Kit, TEMED (Sinopharm Chemical Reagents

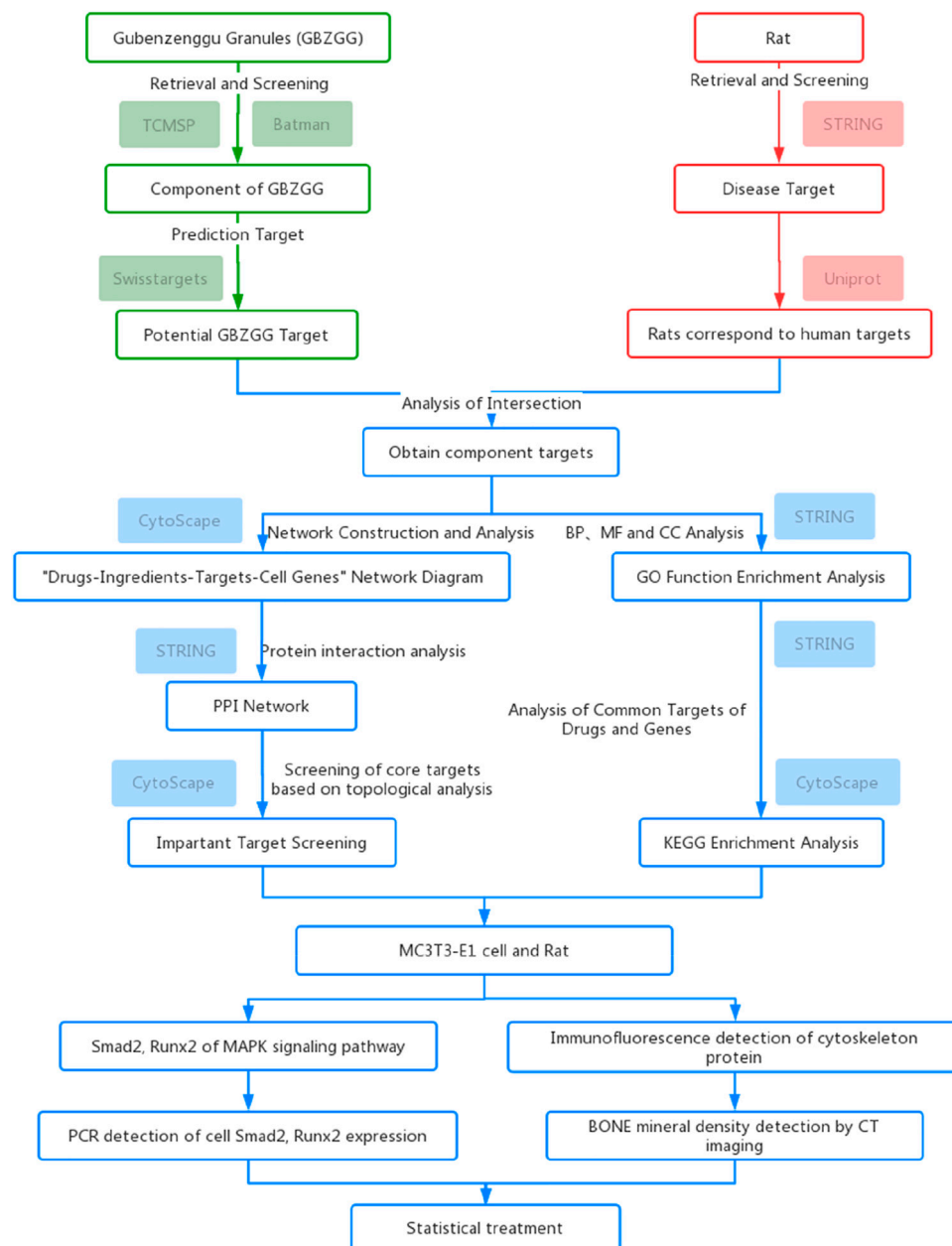


FIGURE 1

Shows the experimental design route in this paper.

Co., Ltd.), TrISE-Base (Biofroxx), HCl (Xinyang Chemical Reagents Co., Ltd.), DTT (Biofroxx), SDS (Sinopharm Chemical Reagents Co., Ltd.), Bromophenol blue (Sinopharm Chemical Reagents Co., Ltd.), Glycerin (Sinopharm Chemical Reagents Co., Ltd.), 30% acrylamide (Biosharp), TrisE-Base (Biofroxx), glycine (Biofroxx), SDS (Sinopharm Chemical Reagents Co., Ltd.), Tris-base (Biofroxx), Glycine (Biofroxx), Twain 20 (Sinopagic Chemical Reagents Co., Ltd.), Protein Marker (10-250kD), PVDF membrane (0.45  $\mu$ m) (Millipore),

PVDF membrane (0.22  $\mu$ m) (Millipore), mouse monoclonal antibody  $\beta$ -actin (40KD) (Wuhan Bod Bioengineering Co., Ltd.), rabbit polyclonal antibody NOX4 (62KD) (Wuhan Sanying Biotechnology Co., Ltd.), mouse monoclonal antibody RANKL (35KD) (Abcam), Rabbit polyclonal antibody OPG (60KD) (Abcam), rabbit polyclonal antibody ColI (129KD) (Abcam), rabbit polyclonal antibody OC (11KD) (Wuhan Sanying Biotechnology Co., Ltd.), rabbit polyclonal antibody OPN (60KD) (Abcam), mouse monoclonal antibody Runx2

(57KD) (Abcam), HRP labeling sheep fight two resistance in mice boster biological engineering co., LTD. (wuhan), HRP labeling sheep rabbit 2 resisting boster biological engineering co., Ltd. (wuhan), ECL substrate liquid Pulitzer gene technology co., Ltd. (Beijing), X-ray film (c sharp cosette door medical equipment co., Ltd.), developing fixing kit (tianjin hanzhong photographic materials plant).

## Experimental methods

### Cell culture and passage

MC3T3-E1 cells (Wuhan Penosai Life Science and Technology Co., Ltd., Article NO. Cl-0378) were taken out of liquid nitrogen using MEM - $\alpha$  + 10%FBS + 1% (Penicillin Streptomycin Solution) cell medium, and quickly put into 37°C water bath. After dissolved, transfer the cells to contain 5 ml medium in the centrifuge tube. Centrifugation was performed at 1000 RPM for 5 min at room temperature, and the supernatant was discarded. In containing 10% fetal bovine serum (Gibco, No. 10099-141) the complete culture medium of cell suspension, and inoculated into a petri dish. The gently blown and mixed cells were cultured at 37°C, 5% CO<sub>2</sub> saturation and humidity. When the cell density reached 80%, subculture, with 0.25% trypsin digestion, collect MC3T3-E1 cells after termination of digestive cells. With PBS washing cells twice, at 1500 rpm, 5 min; Complete culture medium was added, cells were blown, single cell suspension was prepared, and the culture was expanded at 37°C and 5% CO<sub>2</sub> saturation humidity at the ratio of 1:3.

### RNA extraction and gene sequencing

1ul RNA is extracted by TRizol method and quantified by Nanodrop instrument. According to the quantitative results, 500 ng 1% agarose is used for electrophoresis detection, dscDNA is synthesized and the end is supplemented, 12.5ul A-tailing buffer is added. 12.5ul A tail buffer was added to 17.5ul DNA, and the mixture was fully mixed. After 30min at 37°C, the splice was added for PCR enrichment and Qubit was used to quantify the RNA library. Illumina HiSeq3000 was sequenced on a Start CBOT instrument, and the HiSeq 2500 was run on the machine for 11 days before the data was converted into FASTQ format.

### Expression of differential genes in MC3T3-E1 cells under continuous static stress

Using DESeq2 software for screening differentially expressed genes between groups, with different meet | log<sub>2</sub>FC | 1 or higher

and Pvalue 0.05 or less scope of differentially expressed gene screening of the difference between the two groups. For the differentially expressed genes screened between sample groups, bidirectional hierarchical clustering of genes and samples was conducted and heat maps were used to display the clustering parameters (Distance metric: Pearson correlation; Linkage rule: Average Linkage). Mfuzz clustering method was used to classify the expression patterns into 10 groups for the sample size greater than or equal to 6.

### Drug composition and target screening

In TCMSP database (<https://tcmispw.com/tcmisp.php>) to retrieve the astragalus, codonopsis, angelica, epimedium, desertliving cistanche, rehmannia glutinosa, malaytea scurfpea fruit, ground beetle, dog ridge, radix linderae, antler glue ingredient, the composition of the filter is set to the OB 30% or higher, DL 0.18 or higher. For those not included in TCMSP, Batman database is used for retrieval, and Uniprot database is used for standardization and unification of target names. Will be gained by the composition by TCMSP database and Swisstargets database (<http://www.swisstargetprediction.ch/>) to obtain ingredients targets.

### TCM—Component-target-cell gene network construction and analysis

Use Cytoscape 3.7.2 software builds “pharmaceutical ingredients - target cell gene” Network diagram, using the Network Analyzer function to analyze the main effective ingredients of TCM compound. Network Analyzer is used to conduct topology analysis on the Network graph. The number of associations between components and targets is represented by degree values. The larger Degree value indicates that the component is more important.

### PPI network construction and core target analysis

Drug-induced disease will be the common target of the input STRING search the database, the protein type is set to “Mus muscus”, minimum threshold is set to 0.4, the interaction between PPI network build proteins interacting with each other.

### Core target screening based on topology analysis

With degree, median centrality, mean shortest path length and total centrality as reference standards, genes with higher

scores than average were selected as core targets by degree ranking, and bar charts of the first 30 targets were drawn using R3.6.1.

## Gene ontology enrichment analysis

The biological processes (BP), molecular functions (MF) and cellular components (CC) of GO are rich in common targets for drug cell genes and are referenced in the String database. Items with correction  $p < 0.05$  were screened out. Using R 3.6.3 software installation and reference clusterProfiler, rich plot and Ggplot2 package bar and bubble chart.

## KEGG enrichment analysis

Common targets of drug cell genes were analyzed by KEGG pathway enrichment, and the items with  $p < 0.05$  were screened by String database. Using R 3.6.3, after installing and referencing the clusterProfiler package, draw the bar and bubble charts.

## Selection of the dominant dose group

Take the treated MC3T3-E1 cells in good growth condition, adjust the cell density to  $5 \times 10^4$ /ml with MEM- $\alpha$  medium, and connect them to a 96-well plate with 100  $\mu$ l cell suspension per well. In the meantime, set a blank group at 37°C Cultivate overnight (add 100  $\mu$ l sterile PBS to the holes around the cell wells); Treat the cells separately according to the following different groupings and cell treatment settings, each group has 3 multiple wells, cultured at 37°C for 24h; control group; MC3T3-E1 + blank serum 5%, 10%, 20%, MC3T3-E1 + low-dose serum 1%, 5%, 10%, 15%, 20%; MC3T3-E1 + medium-dose serum 1%, 5%, 10%, 15%, 20%; MC3T3-E1 + high-dose serum 1%, 5%, 10%, 15%, 20%, after cell treatment, add 10  $\mu$ l MTT to each well and incubate at 37°C for 3h; aspirate the medium, add 150  $\mu$ l DMSO and shake for 10min; Microplate reader detects the absorbance value OD 568.

## Western blot detection

PVDF membranes were immersed in TBST (sealing fluid) containing 5% skim milk powder and sealed with a mixer at room temperature for 2 h. A diluted with sealing fluid resistance, the PVDF membrane were soaked in a resistance to the fluid of the incubation, 4°C incubation for the night. The PVDF membrane was thoroughly rinsed by TBST 5 times, 5min/time. Put up to 3 membranes in a dish. In the process of flushing, and pay

attention to whether film adhesion on board wall or membrane whether overlap. Corresponding HRP labeling 2 fight after diluted with TBST - 1:50,000, soaked the PVDF membrane in the fluid of the second antibody incubation, and incubated with the table 2 h at room temperature. Fully wash PVDF membrane with TBST 5 times, 5min/time. The enhancement solution in the ECL reagent was mixed with the stable peroxidase solution in a ratio of 1:1. The working droplets were added to the PVDF membrane and reacted for several minutes until the fluorescence band appeared. Then the excess substrate solution was absorbed with filter paper. Cover with plastic wrap, press X-ray film, put in developer solution for development, fixation solution, and rinse the film.

## PCR detection

Add 200  $\mu$ l L chloroform, mix thoroughly several times, and let stand at room temperature for 5 min. Centrifugation was performed at 12000rpm at 4°C for 15min, and the results were divided into three phases: upper (RNA), middle (protein) and lower (DNA). Transfer the upper water phase (about 400  $\mu$ l) into another 1.5 ml EP tube, add 400  $\mu$ l isopropyl alcohol, mix well, and let stand at room temperature for 10min. After centrifugation at 12000rpm at 4°C for 10min, white RNA precipitates could be seen at the bottom of the tube. Abandon the supernatant, add 1 ml without rnase 75% rnase-free 75% ethanol, vortex mixing, 4°C, 10,000 RPM, centrifugal 5 min. Repeat Step 6 once. The supernatant was discarded, the RNA was dried in air for precipitation for 5–10min, and the precipitation was dissolved in 20  $\mu$ l DEPC water. With microspectrophotometer dissolved RNA2  $\mu$ l OD260, OD280 and OD260/OD280 value, calculate the purity and concentration of RNA. According to OD260/OD280 ratio to estimate the quality of RNA, the ratio of between 1.8 ~ 2.0 conform to the requirements of the experiments. The total RNA concentration ( $\mu$ g/ $\mu$ l) = OD260  $\times$  40 $\times$ 10<sup>-3</sup>. Save the total RNA in - 80°C refrigerator to spare.

Primer sequences used for gene detection are as follows:

## Immunofluorescence detection of cytoskeleton protein

The climbing cell slides in the culture plate were immersed in PBS, and the slides are fixed with 4% paraformaldehyde, and normal goat serum is dripped on the slides, and the slides are sealed at room temperature; incubate in a humid box at 37°C, and add fluorescence. DAPI was added dropwise to incubate in the dark, the specimens were stained nucleus; the slides were fixed with fixing solution containing anti-fluorescence quenching agent and observed under fluorescence microscope and images were collected.



## Bone mineral density detection by CT imaging

The tissues were soaked in 4% paraformaldehyde and then fixed for 24 h. Micro-ct was used to detect the morphological indicators of the tissues. Skyscan1276 Micro-CT Scanner software was used for scanning, and the parameters were set as follows: Voltage: 100 KV; current: 100  $\mu$ A; scanning spatial resolution: 10  $\mu$ m; resolution: 4032  $\times$  2688 pixel; rotation Angle: 0.3°; exposure: 500 ms. After scanning, use Data Viewer software for calibration, and then use CT-AN software to select the area of interest. Finally, CT Vox software was used for 3D reconstruction and analysis.

## HE

Gradient alcohol is used to dehydrate the tissue. The tissue blocks must be transparent after dehydrated by alcohol. The transparent agent (xylene) can be mixed with dehydrating agent and paraffin wax at the same time. It replaces the dehydrating agent and the paraffin wax penetrates the tissue smoothly. The transparent tissue blocks were dipped in three cylinders of paraffin (60°C) successively. Embedding is to encase wax-impregnated tissue blocks in paraffin blocks. The temperature of embedding wax should be slightly higher than that of immersion wax to ensure that tissue blocks and embedded paraffin wax are completely integrated. The tissue slices were placed in a 40°C water bath after being sliced by a Leica pathological slicer. Dip anti-stripping slides into the water to scoop slice, slice was attached with the appropriate placement of the slide, at 60°C baking in the oven for 3 h. Paraffin section in xylene (20 min) i - xylene (20 min) ii - xylene iii (min) 20-i anhydrous ethanol (5 min) - ethanol (5 min) ii-95% alcohol (5 min) - 90% alcohol (5 min) - 80% alcohol (5 min) -70% alcohol (5min), soak in distilled water for 5min. Section into Mayer's hematoxylin (clean staining background, no differentiation required) dye for 5min, wash and soak in tap water and return to blue. With 1% water soluble eosin staining section 5 min, 30 s is washed with tap water.

## Statistical treatment

SPSS22.0 software was used for variance analysis of relevant data, and  $p < 0.05$  was statistically significant.

## Experimental results

### Differential gene analysis

According to experimental design, using DESeq2 software is not the same screening differentially expressed genes between groups,

according to  $|\log_2FC| \geq 1$  or higher and Pvalue 0.05 as differentially expressed or less range of filters, according to the results of 29,164 there were differences between two groups of genes, the expression level of 14,489, There were 14,675 down-regulated expressions, and Figures 2A,B were volcanic plots (Figure 2A) and scatter plots (Figure 2B) between the two groups of samples.

## Drug composition and target screening

OB  $\geq 30\%$  and DL  $\geq 0.18$  were set in THE TCMSP database to screen the active components and targets of *Angelica sinensis*, *Codonopsis pilosula*, *Astragalus membranaceus*, *Cistanche deserticola*, *rehmannia glutinosa*, *Aconitum aconitum* and *Herba fularii*. In the Batman database retrieval Dog's back, antler glue and tripelidae, received 98 potential active ingredients. Uniprot database was used to standardize and unify the target names and transform them into corresponding targets in mice. A total of 474 drug targets were screened out (Table 1).

## Drug-cell gene common targets

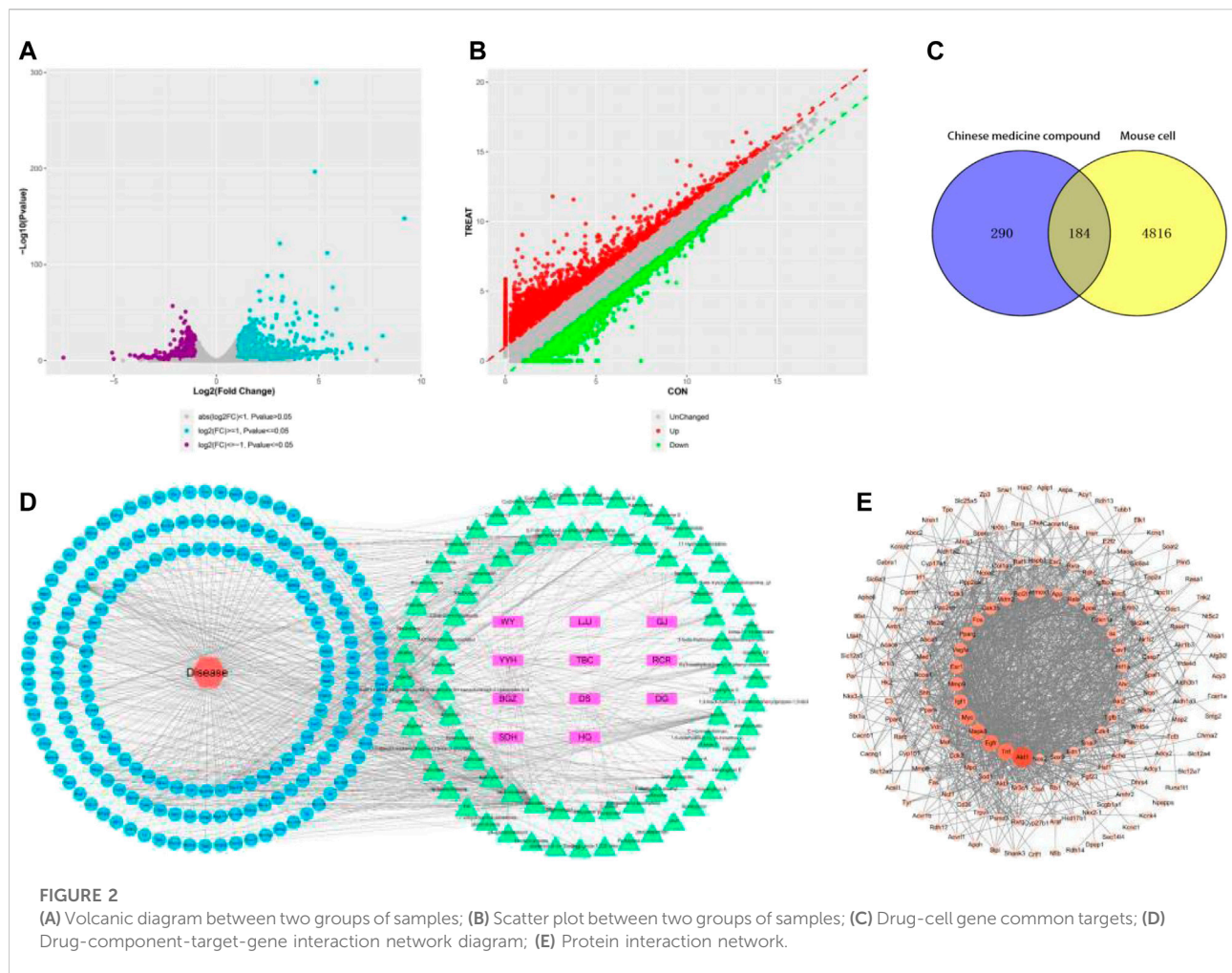
Use Venny2.1 online software drawing tool platform, input 474 of the 5,000 drug targets and disease targets, map Venny2.1, for after the intersection of 184 drug cell gene common targets (Figure 2C).

## Traditional chinese medicine—composition—target—cell gene construction and analysis

98 out of 184 potential active components and drug-disease common targets in Traditional Chinese medicine compound were input into Cytoscape software, targets were deleted, overlapping components were separated, and the interaction network diagram of "drug component-target-gene" was drawn (Figure 2D). In the figure, purple represents drugs, green represents 80 active ingredients in TCM compound (18 active ingredient targets have no intersection with cell gene points and have been deleted, 80 ingredients have been marked red in Table 2), blue represents 184 common targets, and red represents cell genes.

## PPI network construction

Enter the above 184 common targets in the STRING database for a search. Protein type was set to "mice", minimal interaction threshold value of 0.4. Get the target interaction network data, import Cytoscape software rendering protein interaction network diagram (Figure 2E). The Degree value is represented by the size, color, and shadow of the node.



## Gene ontology enrichment analysis

By GO enrichment analysis of David database, 184 common targets were obtained, and BP cross gene sets were enriched into 623 biological process pathways, mainly including positive regulation of RNA polymerase II promoter transcription, positive regulation of transcription and steroid hormone-mediated signaling pathways. CC intersection genes were concentrated in 71 cell components, mainly involving cytoplasm, cytoplasm, neuron cell body, protein complex, etc. In the process related to molecular function, the MF cross gene set was enriched to 140, mainly including protein binding, steroid receptor activity, protein heterodimerization activity, etc. See Figures 3A,B,C.

## KEGG enrichment analysis

By David database to enrichment of 184 common targets, received 107 KEGG pathways. Enrichment of top

20 results form the KEGG function bar chart (Figure 3D), including Pvalue representative enrichment of significance, the deeper the red color, the higher the significance. All the pathways shown here are Pvalue < 0.05, which is meaningful. You can choose the pathway you need based on the literature.

## Screening of the dominant dose group of drug-containing serum

After different concentrations of Guben Zenggu granule medicated serum on the proliferation of MC3T3-E1 cells, the results show that the concentration of medicated serum is too low, the promotion effect is not good, and the concentration is too high, its proliferation will be inhibited. The final 10% high-concentration medicated serum promoted the proliferation of MC3T3-E1 cells best, and its proliferation rate reached 116% (Table3).

TABLE 1 Drug composition of Guben Zenggu granule.

Seractical number	OB(%)	DL	herbs
MOL000358	36.91	0.75	DG、RCR、WY
MOL000449	43.83	0.76	DG、DS、SDH
MOL001006	42.98	0.76	DS
MOL002140	65.95	0.27	DS
MOL002879	43.59	0.39	DS
MOL003036	43.83	0.76	DS
MOL003896	42.56	0.20	DS
MOL004355	42.98	0.76	DS
MOL004492	38.72	0.58	DS
MOL005321	65.90	0.34	DS
MOL000006	36.16	0.25	DS、HYH
MOL006554	38.40	0.77	DS
MOL006774	37.42	0.75	DS
MOL007059	32.16	0.41	DS
MOL007514	39.67	0.23	DS
MOL008391	33.12	0.79	DS
MOL008393	38.33	0.29	DS
MOL008397	50.37	0.77	DS
MOL008400	50.48	0.24	DS
MOL008406	39.97	0.40	DS
MOL008407	45.40	0.76	DS
MOL008411	40.00	0.66	DS
MOL000211	55.38	0.78	HQ
MOL000239	50.83	0.29	HQ
MOL000296	36.91	0.75	HQ
MOL000033	36.23	0.78	HQ
MOL000354	49.60	0.31	HQ
MOL000371	53.74	0.48	HQ
MOL000374	41.72	0.69	HQ
MOL000378	74.69	0.30	HQ
MOL000379	36.74	0.92	HQ
MOL000380	64.26	0.42	HQ
MOL000387	31.10	0.67	HQ
MOL000392	69.67	0.21	HQ
MOL000398	109.99	0.30	HQ
MOL000417	47.75	0.24	HQ
MOL000422	41.88	0.24	HQ、HYH
MOL000433	68.96	0.71	HQ
MOL000438	67.67	0.26	HQ
MOL000439	49.28	0.62	HQ
MOL000442	39.05	0.48	HQ
MOL000098	46.43	0.28	HQ、RCR、WY、HYH
MOL005320	45.57	0.20	RCR
MOL005384	57.52	0.56	RCR
MOL007563	57.53	0.81	RCR
MOL008871	37.05	0.69	RCR
MOL000359	36.91	0.75	SDH、WY、HYH
MOL010495	31.93	0.30	WY

(Continued in next column)

TABLE 1 (Continued) Drug composition of Guben Zenggu granule.

Seractical number	OB(%)	DL	herbs
MOL010496	32.38	0.39	WY
MOL010907	40.92	0.46	WY
MOL010913	77.09	0.25	WY
MOL010916	42.55	0.19	WY
MOL010917	31.18	0.51	WY
MOL001510	37.58	0.71	HYH
MOL001645	42.10	0.20	HYH
MOL001771	36.91	0.75	HYH
MOL001792	32.76	0.18	HYH
MOL003044	35.85	0.27	HYH
MOL003542	38.04	0.39	HYH
MOL004367	62.23	0.41	HYH
MOL004373	45.41	0.44	HYH
MOL004380	39.14	0.49	HYH
MOL004382	56.96	0.77	HYH
MOL004384	45.67	0.50	HYH
MOL004386	51.63	0.55	HYH
MOL004388	60.64	0.66	HYH
MOL004391	48.54	0.25	HYH
MOL004394	41.58	0.61	HYH
MOL004396	52.31	0.22	HYH
MOL004425	41.58	0.61	HYH
MOL004427	31.91	0.86	HYH
MOL000622	63.71	0.19	HYH
The active ingredient			herbs
Corylifolinin			BGZ
Sophoracoumestan A			BGZ
Isopsoralidin			BGZ
Bavachin			BGZ
Bakuchiol			BGZ
Isobavachin			BGZ、GJ
Bavachalcone			BGZ
Bavachromene			BGZ
Psoralidin			BGZ
stigmaterol			BGZ
Xanthotoxin			BGZ
Backuchiol			BGZ
Angelicin			BGZ
Isobavachalcone			BGZ
Cudraphenone D			GJ
Kaempferol			GJ
Cudraphenone A			GJ
Bergapten			GJ
Naringenin			GJ
Aspidinol			GJ
Cudraphenone C			GJ
Cudraphenone B			GJ
Cudraflavanone B			GJ

(Continued on following page)

TABLE 1 (Continued) Drug composition of Guben Zenggu granule.

Seractical number	OB(%)	DL	herbs
Calcium Phosphate			LJJ
Calcium Carbonate			LJJ
Cholesterol			TBC

Note:DG:Angelica, RCR:Cistanche, WY:aconite, DS: Dangshen, SDH:cooked rehmannia glutinosa,RYH:epimedium,HQ:The root of remembranous milk vetch,BGZ: Psoraleae, GJ:dog spine,LJJ:Antler glue,TBC:Ground beetle.

WB detection

The results showed that the ratio of OC/ $\beta$ -actin, Coll/ $\beta$ -actin, OPN/ $\beta$ -actin, OPG/ $\beta$ -actin in the blank serum group was higher than that in the stress model group ( $p < 0.05$ ), while the drug-containing serum group was compared with its ratio increased ( $p < 0.05$ ). The ratios of Rankl/ $\beta$ -actin and Nox4/ $\beta$ -actin in the drug-containing serum group were significantly reduced ( $p < 0.05$ ). Therefore, it can be seen from the table that the intervention effect of the drug-containing serum group on sustained static pressure injury cells is more significant than that of the blank serum group. For details, see Table 4; Figure 4.

Guben Zenggu Granule drug-containing serum can inhibit the expression of Smad2 and Runx2/Cbfa1 genes in MC3T3-E1 cells under continuous static pressure overload

To analyze the influence of Smad2 and RUNx2/Cbfa1 gene expression in MC3T3-E1 cells under continuous static pressure, the cells were grouped to verify the expression levels of Smad2 and Runx2. Results show that the cell model (0.5 mpa) of Smad2 and RUNx2 expression level is lower than normal group ( $p < 0.05$ ). As shown in Figure 11-12. The above results showed that Smad2 and Runx2/Cbfa1 gene expressions were inhibited in MC3T3-E1 cells under continuous static pressure, while the expression of

Smad2 and Runx2/Cbfa1 gene was up-regulated by guben Zenggu granule containing serum, and the expression was normalized (Figures 4H,I).

Bone mineral density and bone trabecular examination

As shown in Figures 5, 6, with the increase of dose, bone mineral density, bone, bone volume fraction volume, trabecular thickness, trabecular number and trabecular spacing returns to normal.

HE dyed

The high dose group had a significant effect on the repair of bone tissue injury compared with the low dose group, indicating that the drug had a repair effect. (Figure 7).

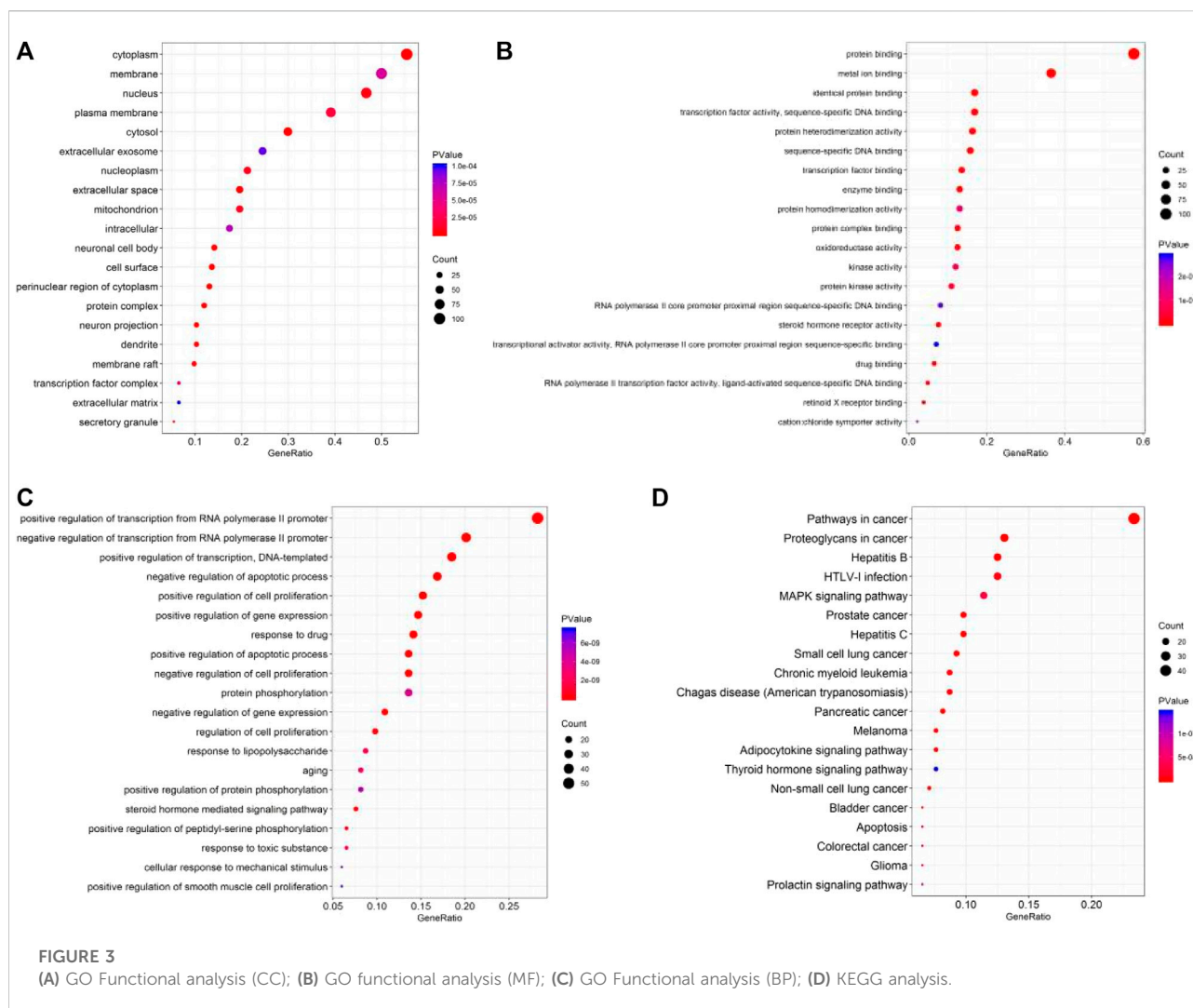
Discussion

The doctor of traditional Chinese medicine in the treatment of the disease has a unique advantage. It has multi-targets and multi-effects in the treatment of diseases. The principle of compatibility of monarchs and ministers makes its application more scientific. However, the dosage is also one of the factors that affect its effect in the process of using Chinese herbal compound. Through the influence on the proliferation rate of MC3T3-E1 cells, the optimal concentration and dose in the drug-containing serum of different concentrations and doses are screened. As the concentration increases, the effect of promoting proliferation is stronger. However, high doses of different doses of medicated serum will inhibit the proliferation of MC3T3-E1 cells. Finally, in a concentration gradient, a 10% dose of medicated serum has the best effect on cell proliferation.

Osteoporosis is a common systemic bone disease, mainly due to the imbalance of bone metabolism, resulting in bone mass reduction and bone microstructure destruction. This study is based on the effective clinical basis of TCM and is the core approach to realize the modernization of TCM. Traditional

TABLE 2 Primer sequence table.

Gene	Primer	Sequence (5'-3')	PCR Products
b-actin	Forward	CACGATGGAGGGGCCGACTCATC	240bp
	Reverse	TAAAGACCTCTATGCCAACACAGT	
Mus smad2	Forward	GACTACACCCACTCCATTCC	233bp
	Reverse	GCAGGTTCCGAGTAAGTAA	
Mus Runx2	Forward	AGATGGGACTGTGGTTACCG	203bp
	Reverse	TAGCTCTGTGGTAAGTGGCC	



Chinese medicine (TCM) is a kind of composition, ways and targets of natural medicine. In this study, cell sequencing genes were analyzed to clarify the basis and mechanism of drug action, and experimental verification was carried out.

In this study, KEGG and GO results showed that Guben Zenggu granule were involved in major signaling pathways including MAPK and positive regulation of cell apoptosis on MC3T3-E1 under continuous static pressure. The negative regulation of cell apoptosis process; Negative regulation of cell proliferation; Positive regulating cell proliferation; Negative regulation of gene expression; Regulate cell proliferation pathways.

MAPK signaling pathway, the bone morphogenetic protein (BMPs) signaling pathway is involved in a variety of bone metabolic processes through two typical Smad protein-dependent pathways (TGF- $\beta$ /BMP ligand, Receptors and Smad proteins) and atypical Smad independent signaling pathways (MAPK signaling TGF- $\beta$ /BMPs P38 mitogen-activated protein kinase signaling

pathway). The former is indispensable in cell stress transduction and osteogenesis, mainly reflected in Smad's regulation of TGF- $\beta$ /BMP signaling pathway, while the latter directly affects cytoskeleton depolymerization and rearrangement, mainly reflected in Smad's regulation of MAPK transduction signal. Runx2 is a downstream target gene of the TGF- $\beta$ /BMP pathway. Runx2 and Smads activated by bone morphogenetic protein jointly induce osteoblast specific gene expression and regulate bone metabolism. The target of bone morphogenetic protein signal is Runx, and BMPs signal transduction pathway is involved in the physiological response of osteoblasts to stress stimulation, and is very important in the level of information transmission in this process. The "endpoint" of mechanical stress stimulation in osteoblasts was the up-regulation of Runx2/Cbfa1 gene expression.

Continuous static pressure is a typical physical factor in the study of osteoporosis. It has been shown in the clinic that an appropriate amount of pressure can promote the therapeutic



TABLE 3 The effect of different concentrations of Gubenzenggu granule medicated serum on the proliferation of MC3T3-E1 cells.

Group	OD	Proliferation rate (%)
blank	0.0777 ± 0.0022	
The control group	1.0259 ± 0.0082	100 ± 0.01
5% Blank serum	1.0041 ± 0.0048	98 ± 0.02
10% Blank serum	1.0873 ± 0.0169	106 ± 0.01
20%Blank serum	0.9499 ± 0.0123	92 ± 0.01
1% Low concentration of drug-containing serum	1.0118 ± 0.0055	99 ± 0.04
5% Low concentration of drug-containing serum	1.0495 ± 0.0342	102 ± 0.02
10% Low concentration of drug-containing serum	1.1247 ± 0.0162	110 ± 0.03
15% Low concentration of drug-containing serum	1.0179 ± 0.0270	99 ± 0.01
20% Low concentration of drug-containing serum	0.8874 ± 0.0141	85 ± 0.01
1% Medium concentration drug containing serum	1.0357 ± 0.0123	101 ± 0.04
5% Medium concentration drug containing serum	1.0529 ± 0.0364	103 ± 0.03
10% Medium concentration drug containing serum	1.1456 ± 0.0288	113 ± 0.02
15% Medium concentration drug containing serum	1.0394 ± 0.0157	101 ± 0.02
20% Medium concentration drug containing serum	0.8950 ± 0.0147	86 ± 0.01
1% High concentration of drug serum	1.0115 ± 0.0071	98 ± 0.02
5% High concentration of drug serum	1.0554 ± 0.0155	103 ± 0.03
10% High concentration of drug serum	1.1748 ± 0.0279	116 ± 0.01
15% High concentration of drug serum	0.9971 ± 0.0097	97 ± 0.02
20% High concentration of drug serum	0.8651 ± 0.0214	83 ± 0.01

TABLE 4 Continuous static pressure damage model index WB detection (n = 3).

	Blank control group	Model group	Blank serum group	Drug containing serum group
OC/ $\beta$ -actin	0.793 ± 0.023	0.459 ± 0.053	0.498 ± 0.053	0.634 ± 0.082*
Coll/ $\beta$ -actin	0.633 ± 0.063	0.337 ± 0.050	0.359 ± 0.042	0.484 ± 0.020*
OPN/ $\beta$ -actin	0.466 ± 0.027	0.199 ± 0.043	0.247 ± 0.043	0.361 ± 0.052*
Rankl/ $\beta$ -actin	0.316 ± 0.033	0.915 ± 0.021	0.869 ± 0.040	0.413 ± 0.043*
Nox4/ $\beta$ -actin	0.829 ± 0.045	1.569 ± 0.069	1.309 ± 0.042	0.974 ± 0.061*
OPG/ $\beta$ -actin	0.937 ± 0.063	0.863 ± 0.25	0.746 ± 0.046	0.850 ± 0.050

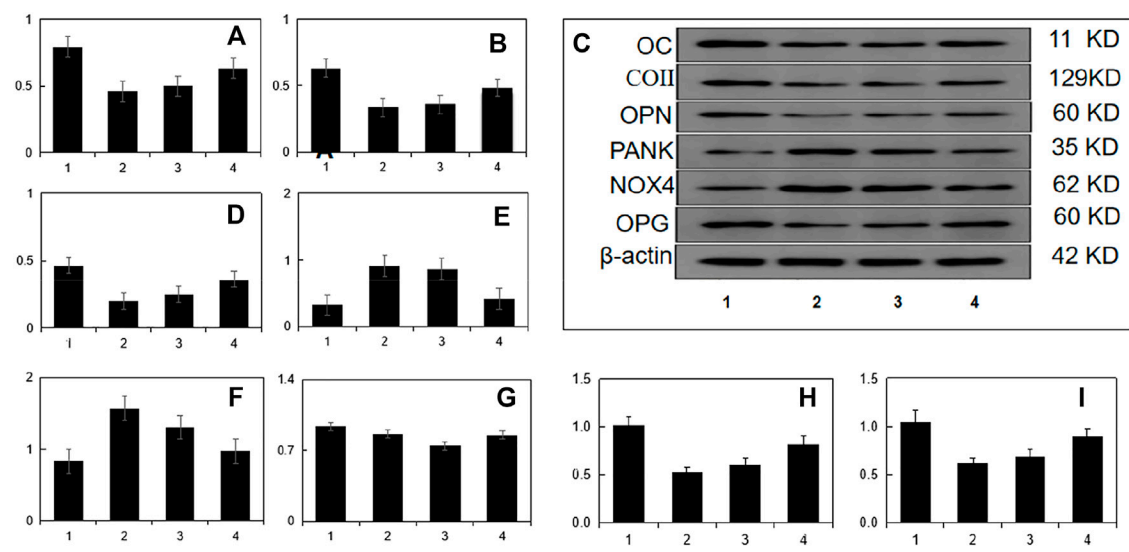
\*Represents  $p < 0.05$ .

effect of OP. The reason is that cells will produce a series of physiological and biochemical reactions after stress stimulation to resist the next pressure. Stimulate. In the related studies of continuous static pressure on MC3T3-E1 cells, OC, Coll, OPN, RANKL, NOX4, OPG, osteocalcin, ADAM28, TD, ALP, Runx2, Wnt1, DKK-1, and other genes and their Protein expression is the main research direction (Song, 2021; Pengjam, 2021; You et al., 2001; Zhao et al., 2021b).

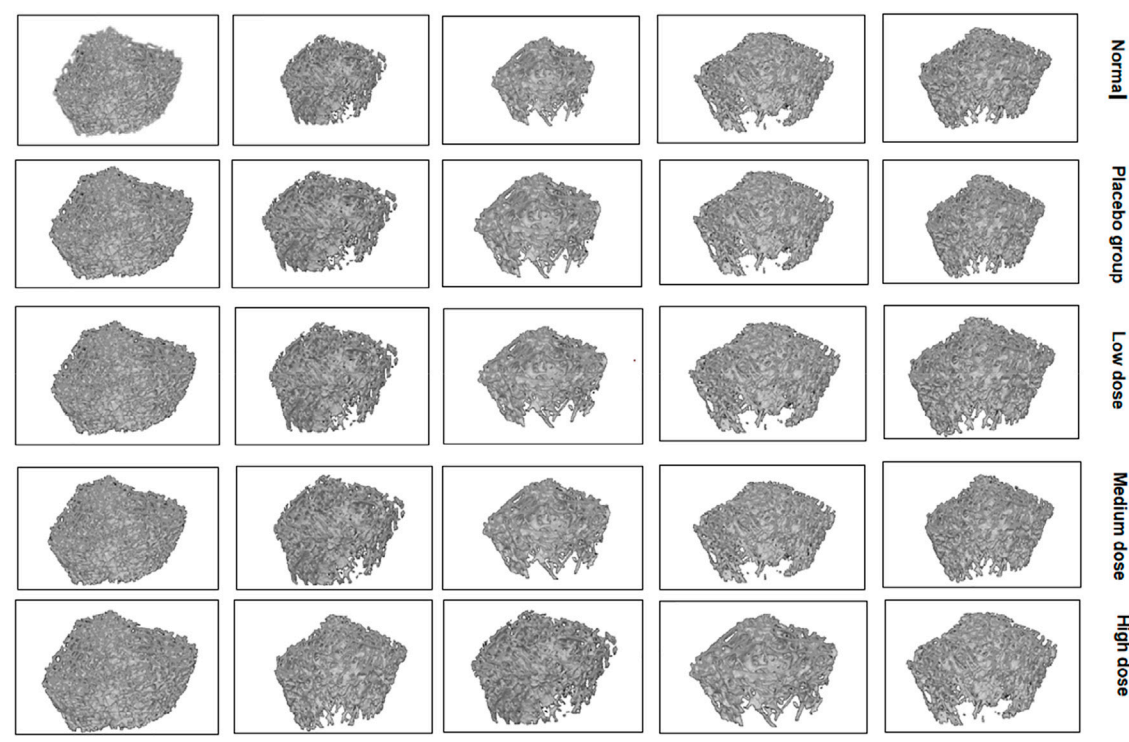
The osteoblasts responsible for bone formation activities are differentiated from bone marrow mesenchymal stem cells. Through exploring osteopontin (OPN), osteocalcin (OC), type II collagen (Col I), and osteoprotegerin (OPG) genes The expression of protein explores the molecular biological mechanism of cellular osteogenics. The results show that Guben Zenggu Granules

medicated serum can regulate the osteogenic markers of MC3T3-E1, indirectly promote the osteogenic differentiation of BMSCs, and affect other bone formation indicators of osteoblasts. It has a promoting effect and can effectively improve molecular biology and anti-oxidation. The results are correlated with previous studies (Feng et al., 2021; Dong et al., 2019a; Dong et al., 2018; Dong et al., 2019b; Wei et al., 2019), confirming that Guben Zenggu Granules medicated serum has a significant effect on osteoporosis, and effectively demonstrates the advantages of traditional Chinese medicine in the treatment of diseases. Advantages of Chinese herbal compound treatment.

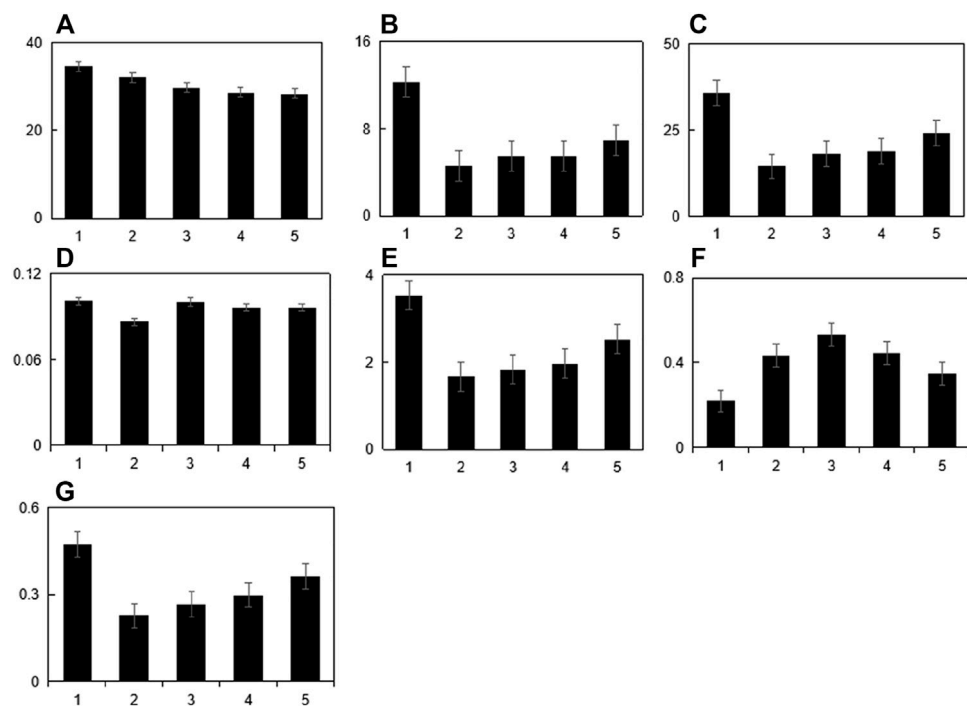
The positive regulatory signaling pathway of apoptosis is an important programmed cell death for metazoan development and internal environment stability. Apoptosis signaling pathway plays



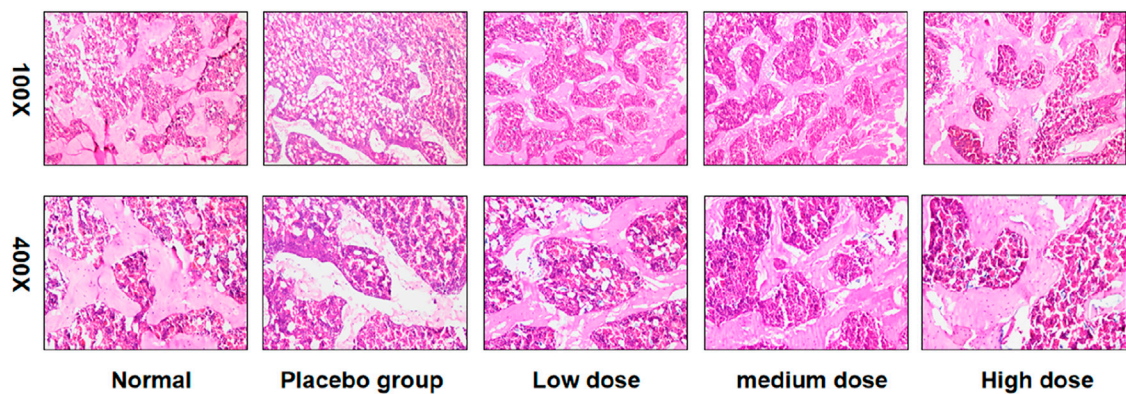
**FIGURE 4** Continuous static pressure damage model index WB detection. Note:1:Blank control group 2:Model group3:Blank serum group 4:Drug containing serum group.(A):OC/β-actin,(B):ColI/β-actin,(D):OPN/β-actin,(E):Rankl/β-actin,(F):Nox4/β-actin,(G):OPG/β-actin;(H):RT-QPCR was used to detect TGF-β/BMP signaling pathway Smad2 in MC3T3-E1 cells. (I)Rt-qpcr was used to detect runx2 gene expression of TGF-β/BMP signaling pathway in MC3T3-E1 cells.



**FIGURE 5** Image of trabecular bone.



**FIGURE 6**  
Bone mineral density index and bone trabecular index. Note: (A):TV MM3 (selected volume of ROI),(B):BV MM3 (bone volume),(C):BV/TV % (bone volume fraction),(D):Tb.Th mm (bone trabecular thickness),(E):Tb.N 1/mm (bone trabecular number),(F):Tb (Bone trabecular separation), (G):BMD G/cm3 (bone density),1. Normal rats group,2. Rat model + placebo group, 3. Rat model + low dose,4. Rat model + medium dose,5. Rat model + high dose\* \*represents  $p < 0.05$ .



**FIGURE 7**  
HE staining of bone tissue with different doses of drugs.

an important role in osteoclast - induced bone loss. Jijie Chai suggested that Smac/DIABLO could not only promote protein hydrolysis activation of procaspase-3, but also promote the enzyme activity of mature caspase-3 (Chai et al., 2000). As a targeted anti-apoptotic drug, Guben Zenggu granule may positively regulate the apoptosis process through MAPK. Negative regulation of

apoptosis process; Negative regulation of cell proliferation; Positive regulating cell proliferation; Negative regulation of gene expression; Regulating cell proliferation pathway and Runx2, Smad gene protein acts on MC3T3-E1 cells to treat induction of bone resorption related diseases such as osteoporosis and fracture.

## The conclusion

As a targeted anti-apoptotic drug, Guben Zenggu granule may positively regulate the apoptosis process through MAPK. Negative regulation of apoptosis process; Negative regulation of cell proliferation; Positive regulating cell proliferation; Negative regulation of gene expression; Regulating cell proliferation pathway and Runx2, Smad gene protein acts on MC3T3-E1 cells to treat induction of bone resorption related diseases such as osteoporosis and fracture.

Too high a dose of medicated serum will inhibit the proliferation of MC3T3-E1 cells. In the final concentration gradient, a 10% dose of medicated serum has the best effect on cell proliferation. Guben Zenggu Granules medicated serum may directly participate in inducing the generation and differentiation of OB and OC by regulating TGF- $\beta$ /BMP and BMP-2/Smads signaling pathways, and regulate bone metabolism in a continuous static pressure overload environment. This may be one of the effects of Guben Zenggu Granules medicated serum through regulating bone metabolism.

Guben Zenggu Granules medicated serum can promote the expression of Tubulin and actin proteins in MC3T3-E1 cells in a continuous static pressure overload environment, thereby promoting bone metabolism.

## 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

ZS, HZ, YJ, RZ, XP, HN, and HC: Editing data curation, Supervision. JP, YG, MS, and WW: Writing- original draft preparation.

## References

- Ai, X., Yu, P., Hou, Y., Song, X., Luo, J., Li, N., et al. (2020). A review of traditional Chinese medicine on treatment of diabetic retinopathy and involved mechanisms. *Biomed. Pharmacother.* 132, 110852. doi:10.1016/j.biopha.2020.110852
- Chai, J., Du, C., Wu, J. W., Kyin, S., Wang, X., and Shi, Y. (2000). Structural and biochemical basis of apoptotic activation by Smac/DIABLO. *Nature* 406 (6798), 855–862. doi:10.1038/35022514
- Cui, X., Wang, S., Cao, H., Guo, H., Li, Y., Xu, F., et al. (2018). A review: The bioactivities and pharmacological applications of polygonatum sibiricum polysaccharides. *Molecules* 23 (5), 1170. doi:10.3390/molecules23051170
- Dong, W., et al. (2018). Effects of Guben Zenggu Recipe on serum osteocalcin and free [Ca-(2+)]i in NEI network tissue in ovariectomized rats[J]. *Pharmacol. Chin. Med. Clin.* 34 (01), 121
- Dong, W., Gong, Y. L., Song, M., Huang, K., Dong, P., Song, Z. J., et al. (2019). Effects of guben zenggu decoction on bone metabolism and bone microstructure in ovariectomized rats. *J. Sichuan Univ. Med. Ed.* 50 (05), 679–683.
- Dong, W., Huang, K., Song, M., Zhou, L. T., Hou, H. Y., Liu, T., et al. (2019). Effect of guben zenggu decoction on expressions of serum BALP and CaM, CaMK mRNA in NEI network of ovariectomized rats. *J. Sichuan Univ. Med. Sci. Ed.* 50 (1), 27–33.
- Feng, Yilei, et al. (2021). Effects of Guben Zenggu Recipe combined with hyperbaric oxygen therapy on bone metabolism balance in osteoporosis model rats[J]. *J. Traditional Chin. Med.* 62 (05), 433
- Gali, J. C. (2001). Osteoporosis[J]. *Acta Ortopédica Bras.* 9, 53–62. doi:10.1590/S1413-78522001000200007
- Gallagher, J., et al. (1994). Diagnosis, prophylaxis, and treatment of osteoporosis [J]. *Am. J. Med.* 90 (90), 646
- He, R., Jin, Z., Ma, R. y., Hu, F. d., and Dai, J. y. (2021). Network pharmacology unveils spleen-fortifying effect of Codonopsis Radix on different gastric diseases based on theory of “same treatment for different diseases” in traditional Chinese medicine. *Chin. Herb. Med.* 13 (2), 189–201. doi:10.1016/j.chmed.2020.12.005

## Funding

1. National Natural Science Foundation of China: Response mechanism of MC3T3-E1 cell stress transduction signal and cytoskeleton and intervention effect of Guben Zenggu granules under continuous static pressure, No. 81960877. 2. Natural Science Foundation of Gansu Province: The effect of high altitude environment on osteogenesis of osteoporosis rats and the intervention effect of Guben Zenggu Granules. No. 21JR7RA561. 3. Innovation Fund of Universities in Gansu Province: Study on the mechanism of low mechanical stress affecting MC3T3-E1 cell activity through TRIM21/F-Actin signal and the intervention effect of Guben Zenggu Pill containing drug serum. No. 2021A-076. 4. Open Fund project of Dunhuang Key Laboratory of Medicine and Transformation of Ministry of Education, Research and Development of Dunhuang Mofeng ointment and *in vitro* pharmacodynamics evaluation, No. DHYX20-16.

## 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.

- Kanis, J. A., et al. (2019). European guidance for the diagnosis and management of osteoporosis in postmenopausal women[J]. *Osteoporos. Int.* 30 (1), 3–44. doi:10.1007/s00198-018-4704-5
- Li, H., et al. (2014). A network pharmacology approach to determine active compounds and action mechanisms of ge-gen-qin-lian decoction for treatment of type 2 diabetes[J]. *Evidence-based complementary Altern. Med.* 2014, 495840. doi:10.1155/2014/495840
- Li, J., Wang, T., Zhu, Z., Yang, F., Cao, L., and Gao, J. (2017). Structure features and anti-gastric ulcer effects of inulin-type fructan CP-A from the roots of *codonopsis pilosula* (franch.) nannf. *Molecules* 22 (12), 2258. doi:10.3390/molecules22122258
- Li, Y., Sun, J., Wu, R., Bai, J., Hou, Y., Zeng, Y., et al. (2020). Mitochondrial mptp: A novel target of ethnomedicine for stroke treatment by apoptosis inhibition. *Front. Pharmacol.* 11, 352. doi:10.3389/fphar.2020.00352
- Liu, Z., Guo, F., Wang, Y., Li, C., Zhang, X., Li, H., et al. (2016). BATMAN-TCM: A bioinformatics analysis tool for molecular mechanism of traditional Chinese medicine. *Sci. Rep.* 6 (1), 21146. doi:10.1038/srep21146
- Pengjam, Y., Syazwani, N., Inchai, J., Numit, A., Yodthong, T., Pitakpornprecha, T., et al. (2021). High water-soluble curcuminoids-rich extract regulates osteogenic differentiation of MC3T3-E1 cells: Involvement of Wnt/ $\beta$ -catenin and BMP signaling pathway[J]. *Chin. Herb. Med.* 13 (4), 534–540. doi:10.1016/j.chmed.2021.01.003
- Qiao, H., Dai, Z., Ge, S., Xu, S., and Yang, L. (2015). Application progress of molecular docking technology in the field of new drug research and development[J]. *J. Nanyang Normal Univ.* 14 (12), 29.
- Ru, J., Li, P., Wang, J., Zhou, W., Li, B., Huang, C., et al. (2014). Tcmsp: A database of systems pharmacology for drug discovery from herbal medicines. *J. Cheminform.* 6 (1), 13–16. doi:10.1186/1758-2946-6-13
- Song, M., Gong, Y., Dong, P., Dong, W., and Jiang, L. (2020). Based on the BMP-Smad/RUNX2 signaling pathway to explore the effect of Guben Zenggu Recipe containing serum on the proliferation and osteogenic differentiation of rat BMSCs [J]. *World Sci. Technology-Chinese Med. Mod.* 22 (04), 1159–1165.
- Song, M., Gong, Y., Dong, P., Dong, W., Liu, X., Dong, P., et al. (2020). Effect of Guben Zenggu Recipe containing serum on osteogenic differentiation of bone marrow mesenchymal stem cells[J]. *Chin. J. Osteoporos.* 26 (04), 511.
- Song, M., Gong, Y., Dong, W., Huang, K., Dong, P., Hou, H., et al. (2020). Effects of Guben Zenggu Recipe on serum BGP, TRACP-5b and bone quality in ovariectomized rats[J]. *World Sci. Technology-Modernization Traditional Chin. Med.* 22 (10), 3682–3687.
- Song, Z., Wang, W., Song, M., Dong, W., Gong, Y., and Wang, K. (2021). Research progress on the osteogenic expression mechanism of MC3T3-E1 cells under different mechanical stress stimulation[J]. *Chin. J. Osteoporos.* 27 (02), 289–292. +307.
- Soudy, M., Anwar, A. M., Ahmed, E. A., Osama, A., Ezzeldin, S., Mahgoub, S., et al. (2020). UniprotR: Retrieving and visualizing protein sequence and functional information from Universal Protein Resource (UniProt knowledgebase). *J. Proteomics* 213, 103613. doi:10.1016/j.jpro.2019.103613
- Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47 (D1), D607–D613. doi:10.1093/nar/gky1131
- Wan, Y., Xu, L., Liu, Z., Yang, M., Jiang, X., Zhang, Q., et al. (2019). Utilising network pharmacology to explore the underlying mechanism of Wumei Pill in treating pancreatic neoplasms. *BMC Complement. Altern. Med.* 19 (1), 158–212. doi:10.1186/s12906-019-2580-y
- Wei, M., Song, M., Dong, W., and Feng, Y. (2019). Guben Zenggu Decoction combined with calcium adjuvant treatment of 46 cases of vertebral compression fractures of spleen and kidney deficiency osteoporosis [J]. *J. Gansu Univ. Traditional Chin. Med.* 36 (06), 57–61. doi:10.16841/j.issn1003-8450.2019.06.13
- Xu, Q., Guo, Q., Wang, C. X., Zhang, S., Wen, C. B., Sun, T., et al. (2021). Network differentiation: A computational method of pathogenesis diagnosis in traditional Chinese medicine based on systems science. *Artif. Intell. Med.* 118, 102134. doi:10.1016/j.artmed.2021.102134
- Yaacobi, E., Sanchez, D., Maniar, H., and Horwitz, D. S. (2017). Surgical treatment of osteoporotic fractures: An update on the principles of management. *Injury* 48, S34–S40. doi:10.1016/j.injury.2017.08.036
- You, J., Reilly, G. C., Zhen, X., Yellowley, C. E., Chen, Q., Donahue, H. J., et al. (2001). Osteopontin gene regulation by oscillatory fluid flow via intracellular calcium mobilization and activation of mitogen-activated protein kinase in MC3T3-E1 osteoblasts. *J. Biol. Chem.* 276 (16), 13365–13371. doi:10.1074/jbc.M009846200
- Zhang, N. D., Han, T., Huang, B. K., Rahman, K., Jiang, Y. P., Xu, H. T., et al. (2016). Traditional Chinese medicine formulas for the treatment of osteoporosis: Implication for antiosteoporotic drug discovery. *J. Ethnopharmacol.* 189, 61–80. doi:10.1016/j.jep.2016.05.025
- Zhang, W., Xue, K., Gao, Y., Huai, Y., Wang, W., Miao, Z., et al. (2019). Systems pharmacology dissection of action mechanisms of Dipsaci Radix for osteoporosis. *Life Sci.* 235, 116820. doi:10.1016/j.lfs.2019.116820
- Zhao, J., Zeng, L., Wu, M., Huang, H., Liang, G., Yang, W., et al. (2021). Efficacy of Chinese patent medicine for primary osteoporosis: A network meta-analysis. *Complement. Ther. Clin. Pract.* 44, 101419. doi:10.1016/j.ctcp.2021.101419
- Zhao, M., Li, S., Ahn, D. U., and Huang, X. (2021). Phosvitin phosphopeptides produced by pressurized hea-trypsin hydrolysis promote the differentiation and mineralization of MC3T3-E1 cells via the OPG/RANKL signaling pathways. *Poult. Sci.* 100 (2), 527–536. doi:10.1016/j.psj.2020.10.053
- Zhu, Y., Yip, R., and Wang, J. (2021). Opportunistic CT screening of osteoporosis on thoracic and lumbar spine: A meta-analysis[J]. *Clin. Imaging* 80, 382–390. doi:10.1016/j.clinimag.2021.08.005





## OPEN ACCESS

## EDITED BY

Li Zhang,  
Royal Holloway, University of London,  
United Kingdom

## REVIEWED BY

Anastasios Doulamis,  
National Technical University of Athens,  
Greece  
Peng Gao,  
Children's Hospital of Philadelphia,  
United States

## \*CORRESPONDENCE

Qiang Lin,  
qiang.lin2010@hotmail.com  
Haijun Wang,  
1718315929@qq.com

\*These authors have contributed equally  
to this work

## SPECIALTY SECTION

This article was submitted to Molecular  
Diagnostics and Therapeutics,  
a section of the journal  
Frontiers in Molecular Biosciences

RECEIVED 30 May 2022

ACCEPTED 13 October 2022

PUBLISHED 28 October 2022

## CITATION

Lin Q, Gao R, Luo M, Wang H, Cao Y,  
Man Z and Wang R (2022), Semi-  
supervised segmentation of metastasis  
lesions in bone scan images.  
*Front. Mol. Biosci.* 9:956720.  
doi: 10.3389/fmolb.2022.956720

## COPYRIGHT

© 2022 Lin, Gao, Luo, Wang, Cao, Man  
and Wang. This is an open-access article  
distributed under the terms of the  
[Creative Commons Attribution License](#)  
(CC BY). The use, distribution or  
reproduction in other forums is  
permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original  
publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or  
reproduction is permitted which does  
not comply with these terms.

# Semi-supervised segmentation of metastasis lesions in bone scan images

Qiang Lin<sup>1,2,3\*</sup>, Runxia Gao<sup>2,3†</sup>, Mingyang Luo<sup>2,3†</sup>, Haijun Wang<sup>4\*</sup>,  
Yongchun Cao<sup>1,2,3</sup>, Zhengxing Man<sup>1,2,3</sup> and Rong Wang<sup>4</sup>

<sup>1</sup>School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou, China, <sup>2</sup>Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, China, <sup>3</sup>Key Laboratory of Streaming Data Computing Technologies and Application, Northwest Minzu University, Lanzhou, China, <sup>4</sup>Department of Nuclear Medicine, Gansu Provincial Hospital, Lanzhou, China

To develop a deep image segmentation model that automatically identifies and delineates lesions of skeletal metastasis in bone scan images, facilitating clinical diagnosis of lung cancer-caused bone metastasis by nuclear medicine physicians. A semi-supervised segmentation model is proposed, comprising the feature extraction subtask and pixel classification subtask. During the feature extraction stage, cascaded layers which include the dilated residual convolution, inception connection, and feature aggregation learn the hierarchical representations of low-resolution bone scan images. During the pixel classification stage, each pixel is first classified into categories in a semi-supervised manner, and the boundary of pixels belonging to an individual lesion is then delineated using a closed curve. Experimental evaluation conducted on 2,280 augmented samples (112 original images) demonstrates that the proposed model performs well for automated segmentation of metastasis lesions, with a score of 0.692 for DSC if the model is trained using 37% of the labeled samples. The self-defined semi-supervised segmentation model can be utilized as an automated clinical tool to detect and delineate metastasis lesions in bone scan images, using only a few manually labeled image samples. Nuclear medicine physicians need only attend to those segmented lesions while ignoring the background when they diagnose bone metastasis using low-resolution images. More images of patients from multiple centers are typically needed to further improve the scalability and performance of the model *via* mitigating the impacts of variability in size, shape, and intensity of bone metastasis lesions.

## KEYWORDS

bone scan, metastasis, lesion segmentation, deep learning, semi-supervised learning

# 1 Introduction

A bone scan (skeletal scintigraphy) with technetium-99 methylenediphosphonic acid ( $^{99m}\text{Tc}$ -MDP) is one of the most commonly used clinical tools for screening bone metastasis (Sderlund, 1996; Costelloe et al., 2009). When a primary solid tumor invades into the bone tissue, there will be one or more areas of increased radionuclide uptake in a  $^{99m}\text{Tc}$ -MDP single-photon emission computed tomography ( $^{99m}\text{Tc}$ -MDP SPECT) image. Compared to positron emission tomography (PET) imaging, a  $^{99m}\text{Tc}$ -MDP SPECT bone scan is more available and affordable for surveying skeletal metastases due to its high sensitivity and low cost (Moon et al., 1998).

However,  $^{99m}\text{Tc}$ -MDP SPECT imaging suffers from low specificity typically caused by inferior spatial resolution, accumulation of radiopharmaceuticals in the normal bones, soft tissues or viscera, and uptake in benign processes (Nathan et al., 2013). Low specificity combined with the normal variation of uptake and technical artifacts often brings misinterpretation to human experts when they manually diagnose bone metastasis.

Automated analysis of  $^{99m}\text{Tc}$ -MDP SPECT images is desired for accurate and efficient diagnosis of bone metastasis. Conventional machine learning algorithms have been adopted to develop methods for identifying bone metastasis (Aslanta et al., 2016; Elfarra et al., 2019; Mac et al., 2021; Sadik et al., 2006; Sadik et al., 2008) or delineating metastasis lesions (Cheimariotis et al., 2018; Thorwarth et al., 2013; Zhu et al., 2008). However, the handcrafted image features often have insufficient capability and unsatisfied performance for clinical tasks (Shan et al., 2020).

Deep learning has been revolutionizing the field of machine learning for the past decades. As a mainstream branch of deep learning, convolutional neural networks (CNNs) have gained huge success in computer vision due to their superiority in automatically learning hierarchical features from images in an optimal way. Several excellent review articles present a holistic perspective on the recent progress of deep learning in medical image segmentation (AsgariTaghanaki et al., 2021; Minaee et al., 2022; Litjens et al., 2017; Lei et al., 2020). Semi-supervised learning is becoming one of the hot research topics in this field due to the reduced requirement of large-scale labeled images (Christoph et al., 2017; Doulamis and Doulamis, 2014; Tarvainen and Valpola, 2017).

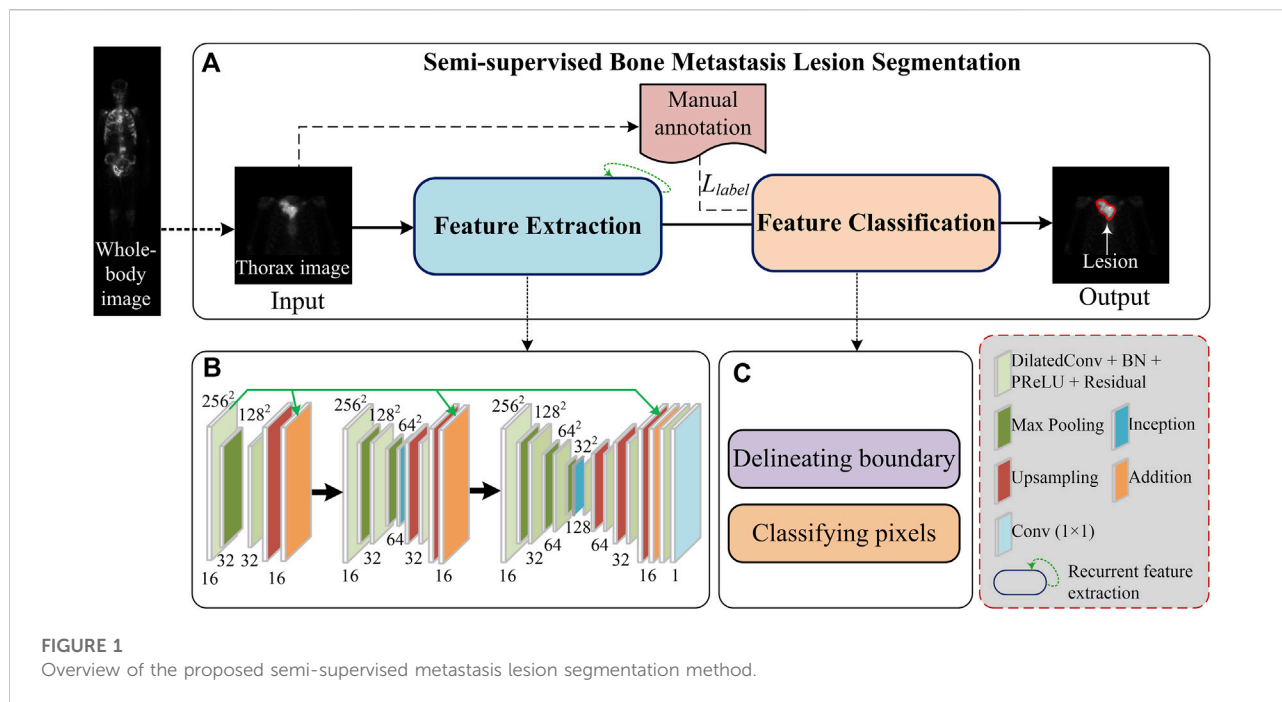
There has recently been a substantial amount of CNN-based work aimed at developing image classification methods for automated detection or diagnosis of metastasis (Bochkovski et al., 2020; Cheng et al., 2021a; Cheng et al., 2021b; Dang, 2016; Guo et al., 2022; Lin et al., 2021a; Lin et al., 2021b; Lin et al., 2021c; Li et al., 2022; Papandrianos et al., 2020a; Papandrianos et al., 2020b; Papandrianos et al., 2020c; Pi et al., 2020; Redmon and Farhadi, 2018; Zhao et al., 2020) by classifying  $^{99m}\text{Tc}$ -MDP SPECT images into categories. By contrast, segmenting  $^{99m}\text{Tc}$ -

MDP SPECT images to detect and delineate metastasis lesions is still in its infancy. Using the recurrent CNN (RCNN) (Liang and Hu, 2015) as a backbone network, Chen and Frey (2020) proposed a semi-supervised segmentation model to delineate bone structures in the pelvis. Their model reported a score of 0.593 for the Dice metric on the simulated instead of real clinical SPECT images. MaligNet (Apiparakoon et al., 2020) is a two-stage network used for semi-supervised segmentation of chest metastasis lesions, consisting of a feature extraction subnetwork (ResNet-50) and feature classification subnetwork (ladder feature pyramid network). Their proposed network achieved a mean score of 0.848 for the F-1 score, without segmentation metrics [e.g., Dice and IoU (intersection over union)] being reported. Based on the classical networks U-Net (Ronneberger et al., 2015) and R-CNN (He et al., 2020), we investigated supervised segmentation of bone metastasis lesions in clinical  $^{99m}\text{Tc}$ -MDP SPECT images (Lin et al., 2020), obtaining a best score of 0.6103 for the IoU metric.

Using clinical  $^{99m}\text{Tc}$ -MDP SPECT bone scans, we have propose an RCNN-based method for segmenting bone metastasis lesions in a semi-supervised way in this work. The proposed segmentation method can automatically identify a lesion and delineate its boundary using only few manually labeled samples (i.e., semi-supervised learning). Our work is based on the following observations.

First, image segmentation involving partitioning images into multiple segments or objects (e.g., organs, tissues, and lesions) is routinely conducted in clinical diagnosis. This thereby enables the extraction of bone metastatic lesions and measurement of lesion volumes. Second, 2D bone scan is characterized by inferior spatial resolution. The size of a whole-body image is 256 (width)  $\times$  1024 (height). This brings a huge challenge to manual analysis by nuclear medicine physicians. Lastly, the lack of large labeled data sets of bone scan images is an especially prevalent challenge in supervised image segmentation because it is very time-consuming, laborious, and subjective to manually labeling lesions in large-sized low-resolution  $^{99m}\text{Tc}$ -MDP SPECT images. On the contrary, CNN-based semi-supervised segmentation has the potential to automatically divide an image into regions of concerns with only a small number of partially labeled samples.

Given that the spine and ribs are the common areas where primary tumors frequently invade, a whole-body SPECT image was first cropped to extract the thoracic region in this work. With the extracted regional images, an end-to-end semi-supervised segmentation network was built to learn the hierarchical representations of  $^{99m}\text{Tc}$ -MDP SPECT images and classify all pixels into categories (i.e., the lesion and background). The pixels falling into each lesion were then surrounded by an irregular closed curve. The delineated lesions could help the human expert focus on bone metastasis lesions while ignoring the background area, thereby enabling to improve the accuracy and efficiency of the diagnosis.



The main contributions of this work are first, we define the research problem of segmenting metastasis lesions in  $^{99m}\text{Tc}$ -MDP SPECT images in a semi-supervised way. Second, we propose a semi-supervised segmentation method consisting of image feature extraction and pixel classification. Last, we utilize a set of clinical  $^{99m}\text{Tc}$ -MDP SPECT images to evaluate the proposed method. The experimental results show that our method performs well in the automated segmentation of bone metastasis lesions with only few manually labeled samples being used.

The remaining part of this article is organized as follows: the proposed semi-supervised segmentation method is detailed in Section 2. Experimental evaluation conducted on clinical data is provided in Section 3. A brief discussion is presented in Section 4. We have concluded this work and pointed out the future research directions in Section 5.

## 2 Methodology

An overview of the proposed semi-supervised segmentation framework is depicted in Figure 1A, comprising two stages, namely, image feature extraction (Figure 1B) and pixel classification (Figure 1C). During the feature extraction stage, a set of cascaded layers is used to learn low-to-high representations of low-resolution images, enabling us to extract features of bone metastasis lesions as much as possible. During the pixel classification stage, the pixels within an image are classified into categories with the partially labeled samples

(semi-supervised) and the boundary of pixels belonging to a lesion is delineated.

### 2.1 Image cropping

Whole-body SPECT bone scanning outputs large images with a size of 256 pixels  $\times$  1024 pixels. This brings a heavy computational burden to the pixel-level classification task. On the other hand, the thoracic region that covers the sternum, clavicles, scapulae, and ribs is the most common site of metastases in a variety of solid cancers (Nathan et al., 2013). Focusing on the automated segmentation of bone metastasis lesions in the thoracic skeleton, an empirical image-cropping method (Li et al., 2022) is used to extract the thoracic region from the whole-body images. The cropping method contains three main steps: “whole-body image  $\rightarrow$  body area,” “body area  $\rightarrow$  upper body,” and “upper body  $\rightarrow$  thoracic region.” A cropped regional image has the size of 256  $\times$  256. Each “pixel” in this image is a 16-bit unsigned integer, representing the detected count of the radiotracer’s uptake.

A regional image can be viewed as a count matrix  $\mathbf{CM}$ , which can be formally represented as

$$\mathbf{CM} = (c_{ij}) | 1 \leq i, j \leq 256 \\ c_{ij} \in [0, c_{\max}] \quad (1)$$

Typically,  $c = 0$  denotes the background pixels, and  $c_{\max}$  varies largely from patient to patient. As mentioned previously, apart

from the metastasized bone, the high uptake of radiopharmaceuticals is commonly seen in normal bones and benign processes. It is thus difficult to normalize the count,  $c$ , into a fixed range, such as it is done in natural image analysis.

## 2.2 Lesion labeling

Semi-supervised learning tasks still need human manual labels to train a segmentation model, where the labeled lesions act as ground truth in the experiments. To help human experts (a chief physician aged 45 years, an associate chief physician aged 40 years, and a resident physician aged 33 years) to manually label a low-resolution SPECT image, we developed an annotation system based on the open-source online tool LabelMe (<http://labelme.csail.mit.edu/Release3.0/>).

Using the LabelMe-based annotation system, three experienced nuclear medicine physicians in our group independently labeled each image. Let  $l = \langle p_1, p_2, \dots, p_m = p_1 \rangle$  denote a manual label, which is a closed curve consisting of points. For a bone metastasis lesion, if the difference between the areas surrounded by any two closed curves is not larger than the threshold  $t_{AA}$ , we randomly select any one from the three labels as the ground truth; a new annotation process will start otherwise. Specifically, the area difference  $\Delta A$  is defined using the Intersection over Union (IoU) in Eq. 2.

$$\Delta A = 1 - \frac{A_{l1} \cap A_{l2}}{A_{l1} \cup A_{l2}}, \quad (2)$$

where  $A_{lk}$  ( $k = 1, 2$ ) represents the area of the closed curve  $lk$ . We assign a value of 5% for  $t_{AA}$  in the experiments.

During the supervised training stage, the manual annotation  $L_{label}$  in Figure 1A will be fed into the segmentation model.

## 2.3 Feature extraction

As depicted in Figure 1B, cascaded layers are used in the feature extraction stage which include the residual dilated convolution, pooling, feature aggregation, inception connection, upsampling, and traditional convolution to extract multi-scale features of lesions from low-resolution images.

### 2.3.1 Dilated residual convolution

Bone metastasis lesions typically demonstrate variability in size, shape, and intensity. Extracting hierarchical features of lesions from low-resolution images is significantly challenging. Compared to the conventional convolution, a dilated convolution has the potential to systematically aggregate multi-scale contextual information without losing resolution (Yu and Koltun, 2015). On the other hand, the residual connection can alleviate the gradient vanishing and explosion problems.

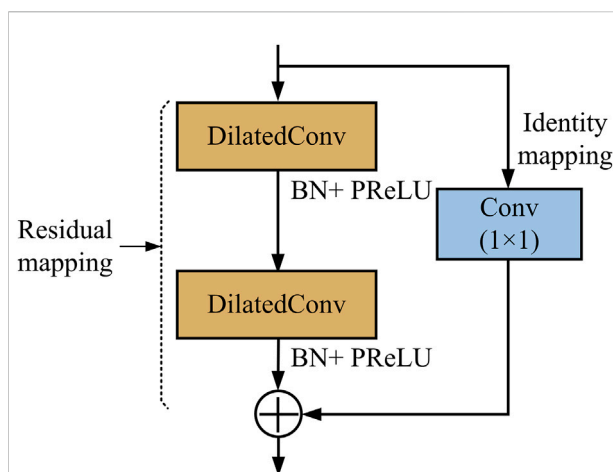


FIGURE 2  
Structure of the residual dilated convolution block.

We used the residual dilated convolution block in the feature extraction stage to extract the multi-scale features of metastasis lesions, which is illustrated in Figure 2.

As shown in Figure 2, there are two paths in the residual dilated convolution block. In the residual mapping path, two dilated convolution (DilatedConv) layers are used to extract multi-scale features, with each followed by a batch normalization (BN) operation and parametric rectified linear unit (PReLU) operation. The BN operation has the potential to make a network utilize much higher learning rates and be less careful about initialization, enabling the acceleration of network training. As a typical activation function, PReLU performs a threshold operation to bring nonlinearity into the network. In the identity path (also skip path), a  $1 \times 1$  conventional convolution is used to reduce the number of depth channels by simply mapping an input pixel to output pixel.

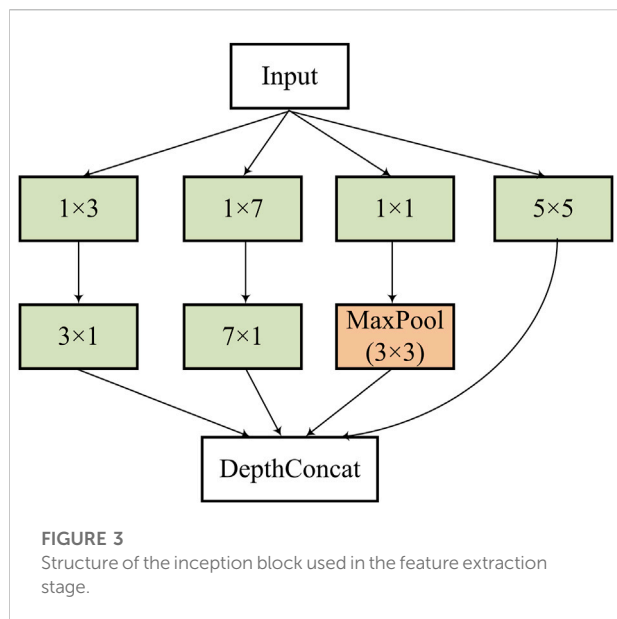
Given the  $i \times i$  feature map  $I_{IN}$  with a kernel size of  $k$  as the input, for any given dilation rate of  $d$ , the size  $o$  of the feature map  $I_{OUT}$  after dilated convolution can be calculated according to Eq. 3.

$$\begin{aligned} o &= \left\lceil \frac{i - 2p - n}{s} \right\rceil + 1, \\ n &= k + (k - 1)(d - 1) \end{aligned} \quad (3)$$

where  $p$  and  $s$  denote the padding and stride, respectively.

### 2.3.2 Inception architecture

Increasing the size of a network (depth and width) is one of the alternatives to improve the performance of a convolutional neural network. However, an enlarged network with a larger size is often prone to overfitting. The inception architecture (Szegedy et al., 2014) can find an optimal local sparse structure in the deep network while allowing for significantly increasing the number of



units at each stage without an uncontrolled blow-up in computational complexity.

We define an inception block (see Figure 3) in this work to further extract multi-scale features of bone metastasis lesions.

The defined inception block consists of several convolutions with different kernels ( $1 \times 3/3 \times 1$ ,  $1 \times 7/7 \times 1$ ,  $1 \times 1$ , and  $5 \times 5$ ) and a max pooling with stride = 2 after a  $1 \times 1$  convolution to capture hierarchal features and halve the resolution of the grid. The DepthConcat concatenates the outputs from the previous layers.

### 2.3.3 Feature aggregation

The dual views of 2D SPECT imaging are exploited to enhance the metastasis lesions by aggregating the anterior and

posterior views of an image. Specifically, let  $I_{Ant}$  and  $I_{Post}$  be the anterior- and posterior-view images (features) of a patient, respectively, and an aggregated image  $I_{Agg}$  can be calculated according to Eq. 4.

$$I_{Agg} = f[I_{Ant}, \text{Mirr}(I_{Post})], \quad (4)$$

where  $f$  is the pixel-wise addition, and  $\text{Mirr}(\cdot)$  is the image horizontal flipping/mirroring operation.

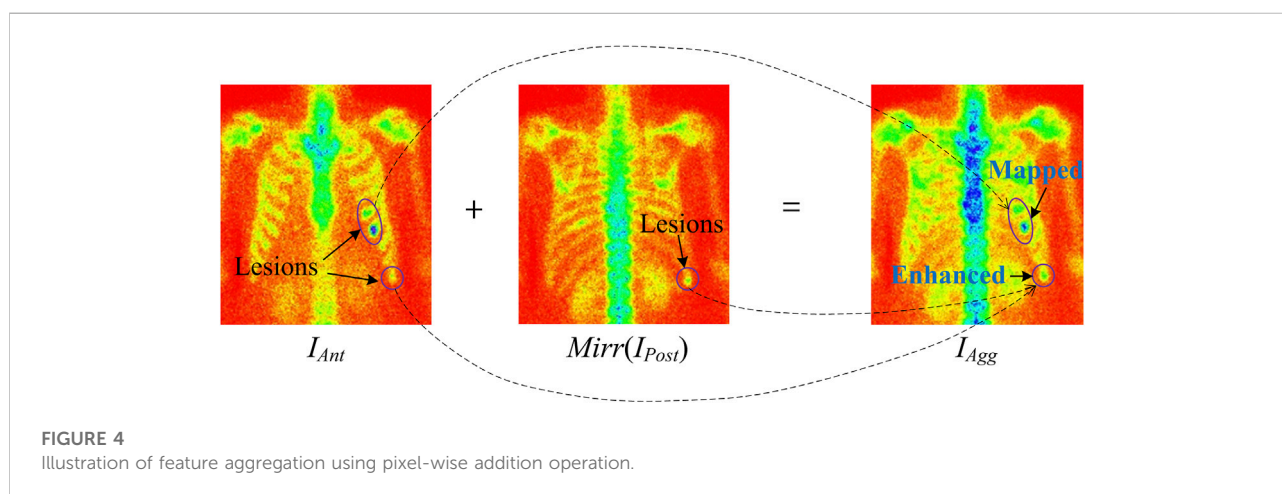
For instance, Figure 4 illustrates the image aggregation by using a pixel-wise addition operation (see “Addition” in Figure 1). Lesions in the aggregated image are either enhanced by adding the lesion areas in both the anterior and posterior views or mapping from the anterior or posterior views. For a patient with diagnosed bone metastasis, the lesion(s) can always be displayed in the aggregated image rather than only in the anterior or posterior one. The dual-view aggregation will be helpful for CNN-based classifiers in detecting lesions.

### 2.3.4 Recurrent feature extraction

Inspired by the structure of the recurrent convolutional layer in RCNN (Liang and Hu, 2015), we let the feature extraction network run in a recurrent way, indicated by the green dotted line with an arrow in Figure 1A, to integrate the context information for lesion segmentation. An illustration of a recurrent feature extraction subnetwork is depicted in Figure 5.

As shown in Figure 5, the feature extraction subnetwork comprises the feedforward network and recurrent network. Let  $u(t)$  be the input of the feedforward network at time  $t$ , and  $x(t-1)$  be the input of the recurrent network at time  $t-1$ , the output of the network at time  $t$  can be calculated according to Eq. 5.

$$z_{ijk}(t) = [W_{F(k)}]^T u^{(ij)}(t) + [W_{R(k)}]^T x^{(ij)}(t-1) + b_k \quad (5)$$





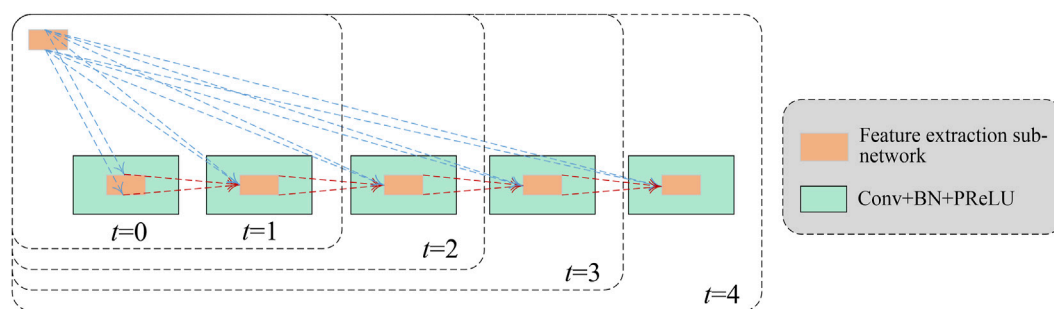


FIGURE 5

Structure of the recurrent feature extraction network with  $t$  indicating recurrence time.

where  $(i, j)$  indicates the pixel in the  $k$ -th feature map,  $W_{F(k)}$  and  $W_{R(k)}$  are the convolutional parameters of the feedforward and recurrent nets, and  $b_k$  is the bias.

## 2.4 Feature classification

Typically, image segmentation performs portioning of objects by automatically classifying pixels into the corresponding categories. In the paradigm of supervised learning, the labeled samples of images are used to train a segmentation model, which is then used to predict the class,  $y$ , from a pixel,  $x$ , of any new image. On the contrary, an unsupervised model is trained without manual labels.

Our semi-supervised segmentation model is trained using a large number of unlabeled samples together with only few labeled samples to build a pixel-level feature classifier. A core consideration of the semi-supervised segmentation model is in how to determine a segmentation loss function, which would probabilistically predict the class for each pixel and be defined according to the unsupervised segmentation loss and manual labels. The semi-supervised segmentation loss  $\ell$  consists of two parts: unsupervised loss  $\ell_U$  and supervised loss  $\ell_S$ , which is defined in Eq. 6.

$$\ell = \ell_U + \alpha \cdot \ell_S. \quad (6)$$

For a given image  $g$ , let  $c = f(g)$  be a closed curve that delineates a bone metastasis lesion, the unsupervised loss  $\ell_U$  in Eq. 6 can be defined as

$$\ell_U = \nu \cdot \text{Area}(f(g) > 0) + \sum_{f(g) > 0} |g(x, y) - c_1|^2 + \sum_{f(g) < 0} |g(x, y) - c_2|^2. \quad (7)$$

where  $c_1$  and  $c_2$  are the average of image  $g$  inside curve  $c$  and outside curve  $c$ , respectively; and  $\text{Area}(\cdot)$  is the function used for

measuring the area inside the curve, which has been defined by Chan and Vese (2001).

Let  $\Omega$  be the image filed and  $u$  be the label, the supervised segmentation loss  $\ell_S$  can be calculated according to Eq. 8.

$$\ell_S = \sum_{f(i)} |\nabla(f(g))| + \sum_{\Omega} ((1 - \mu)^2 - (0 - \mu)^2) f(g). \quad (8)$$

In the aforementioned Eqs. 7, 8, the constants are assigned as  $\alpha = 0.4$ ,  $\nu = 0.004$ , and  $u = 0$  in the experiments.

With the semi-supervised segmentation loss function defined in Eq. 6, the segmentation model classifies each pixel into one of the categories (i.e., the background and lesion regions). The boundary of pixels that falls into an individual lesion is then delineated using a closed curve. For an input image, the segmented lesions act as the output of a segmentation model.

## 3 Results

### 3.1 Experimental data

The SPECT images used in this retrospective study were acquired from the Department of Nuclear Medicine, Gansu Provincial Hospital. A total of 724 whole-body images were collected from 362 patients, who were clinically diagnosed with bone metastasis by using a single-head equipment (GE SPECT Millennium MPR) with a parallel-beam low-energy high-resolution (LEHR) collimator (energy peak = 140 keV, intrinsic energy resolution  $\leq 9.5\%$ , energy window = 20%, and intrinsic spatial resolution  $\leq 6.9$  mm). The SPECT imaging was taken between 2 and 3 h after the intravenous injection of  $^{99m}\text{Tc}$ -MDP (20–25 mCi). The imaging size was  $256 \times 1024$  with a pixel size of 2.26 mm. The acquisition time was 10–15 min for each whole-body bone scan image.

We selected 112 images that contained bone metastasis lesions in the thorax from all images. The selected images were cropped to extract the regional images using the image

TABLE 1 Overview of the data sets used in this work.

Data set	Number of samples	Annotation
D1	112	—
D2	2280	—
D3	2280	—
D4	4560	D2 + D3

cropping technique as mentioned in Section 2.1. We organized these 112 regional images into the data set D1, which is outlined in Table 1.

The subsequent section details the process of augmenting the size of D1, which would be helpful for training a better segmentation model since deep learning-based models often perform well on the “big” data set.

## 3.2 Data augmentation

### 3.2.1 Geometric transformation

Geometric transformations such as flipping, cropping, rotation, and translation are frequently used in the field of deep learning-based image augmentation (Shorten, Khoshgoftaar). In this work, image flipping, rotation, and translation were applied on the images in data set D1 to obtain more samples.

The augmented samples that were obtained by using the aforementioned geometric transformations and the images in data set D1 were grouped into data set D2, which is outlined in Table 1.

### 3.2.2 Adversarial learning

The generative adversarial network (GAN) (Goodfellow et al., 2014) is one of the most emerging deep learning techniques that is used to generate new “fake” samples with the given images. The generated samples have an entirely different distribution from the original ones. The GAN consists of a generator ( $G$ ) and discriminator ( $D$ ). Specifically, generator  $G$  generates “fake” data  $G(z)$  from a distribution of  $P_Z$  and discriminator  $D$  distinguishes fake data from real data  $X$ .

The deep convolutional GAN (DCGAN) (Radford et al., 2015) as a variant of GAN has the potential to improve the stability of training and alleviate mode collapse that the original GAN may suffer from. Assume that the distribution of real data is  $P_D$ , both the generator and discriminator are iteratively optimized against each other in a minimax game as follows (Radford et al., 2015):

$$\max_{\theta_G} \max_{\theta_D} E_{x \sim P_D} [\log D(x)] + E_{z \sim P_Z} [\log (1 - D(G(z)))], \quad (9)$$

where  $\theta_G$  and  $\theta_D$  denote the parameters of  $G$  and  $D$ , respectively.

In this work, we used a DCGAN-based technique to generate the simulated samples of SPECT images. Figure 6 shows the diagram of training such a network, where the iteration parameter is set as  $k = 3$  in the experiment.

The samples generated by using the aforementioned DCGAN-based technique and the images in data set D1 are grouped into data set D3, which is outlined in Table 1.

## 3.3 Experimental setup

The evaluation metrics used include Dice similarity coefficient ( $DSC$ ), class pixel accuracy ( $CPA$ ), and *Recall*, which are defined in Eqs 10–12.

$$DSC = \frac{2 \cdot TP}{FP + 2 \cdot TP + FN}, \quad (10)$$

$$CPA = \frac{TP}{TP + FP}, \quad (11)$$

$$Recall = \frac{TP}{TP + FN}, \quad (12)$$

where  $TP$  = true positive,  $TN$  = true negative,  $FP$  = false positive, and  $FN$  = false negative.

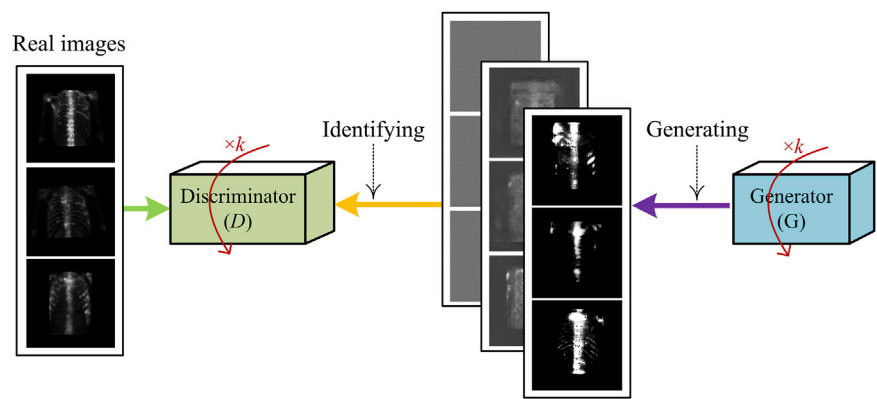
The parameter setting of the proposed deep segmentation model is provided in Table 2.

In the experiment, we divided each data set in Table 1 into two parts: subset 1# for unsupervised learning and subset 2# for supervised learning. Specifically, in each subset, we have randomly chosen 70% of the samples to train and the rest (30%) to test the developed segmentation model. It is worth noting that samples without manual labels are used to train the proposed model in an unsupervised manner, and samples with a varied number of manual labels are used to train the model in a semi-supervised manner. Images which include the augmented ones from the same patient were not divided into different subsets because they would show similarities. The trained model was run 10 times on the test subset to reduce the effects of randomness. For the aforementioned defined evaluation metrics, the final outputs of the model are the average of the 10 running results. The experimental results reported in the next section are the averaged ones, unless otherwise specified.

The experiments are run in TensorFlow 2.0 on an Inter Xeon(R) Silver 4110 PC with 16 Kernels 62 GB RAM running on Ubuntu 16.04 equipped with a GeForce RTX 2080 × 2.

## 3.4 Experimental results

In this subsection, we evaluated the segmentation performance of the proposed semi-supervised model with respect to the evaluation metrics on several data sets as shown in Table 1.



**FIGURE 6**  
Illustration of generating fake samples of SPECT images using DCGAN-based sample generation technique.

**TABLE 2** Parameter setting of the proposed deep segmentation model.

Parameter	Value
Input size	256 × 256
Optimizer	Adam
Learning rate	0.0005
Learning momentum	0.9
Weight decay	0.0001
Epoch	400

**TABLE 3** Scores of evaluation metrics obtained by the proposed semi-supervised model.

Data set	$L_{label}$	DSC	CPA	Recall
D1	10	0.582	0.618	<b>0.547</b>
D2	210	<b>0.586</b>	<b>0.621</b>	0.539
D3	210	0.481	0.507	0.471
D4	210	0.483	0.514	0.487

The bold value in each column indicates the maximal one.

### 3.4.1 Quantitative scores

Table 3 reports the scores of evaluation metrics obtained by the semi-supervised model on test samples in data sets D1–D4, where  $L_{label}$  refers to the number of labeled samples used while training the model.

On the whole, the proposed semi-supervised model performs best on data set D2. This tells us, on the one hand, that data augmentation positively contributes to improving segmentation performance; on the other hand, geometric transformation is more suitable to be used for augmenting the size of the SPECT

data set when compared to the adversarial learning-based augmentation since image flipping, translation, and rotation operations can preserve label post-transformation.

Using test samples in data set D2, we examined how the number of labeled samples (i.e.,  $L_{label}$ ) affects the segmentation performance, by providing experimental results in Figure 7.

From the experimental results in Figure 7, we can see that as expected, the scores of evaluation metrics keep increasing as  $L_{label}$  increases. When 37% of the labeled samples were used for the training model, the best segmentation performance ( $DSC = 0.683$ ,  $CPA = 0.715$ , and  $Recall = 0.601$ ) was obtained. An exception is that the unsupervised model ( $L_{label} = 0$ ) obtained the highest score for the *Recall* metric, which is mainly contributed by background pixels during testing the model. This reveals the major difference between our SPECT and the natural images: objects (i.e., lesions) in the former are far smaller than those in the background.

### 3.4.2 Ablation experiments

The reported aforementioned experimental results were obtained when the complete model was recurrently run thrice on data set D2 without image aggregation. In this subsection, a set of scores of ablation experiments are reported.

Impact of recurrent feature extraction on segmentation performance: as mentioned in subsection 2.4.1, the feature extraction network can run in a recurrent manner. It is necessary to examine the impact of parameter  $t$  on the segmentation performance, which is illustrated in Figure 8.

Figure 8 demonstrates that recurrent feature extraction is of great necessity, and a value of 3 for  $t$  is optimal when the complete model is used during the feature extraction stage. In Table 4, we further present the number of model parameters and the test time for different recurrence times.

It can be seen that the proposed model can segment bone metastatic lesions efficiently with a maximum test time of 4.21 s.

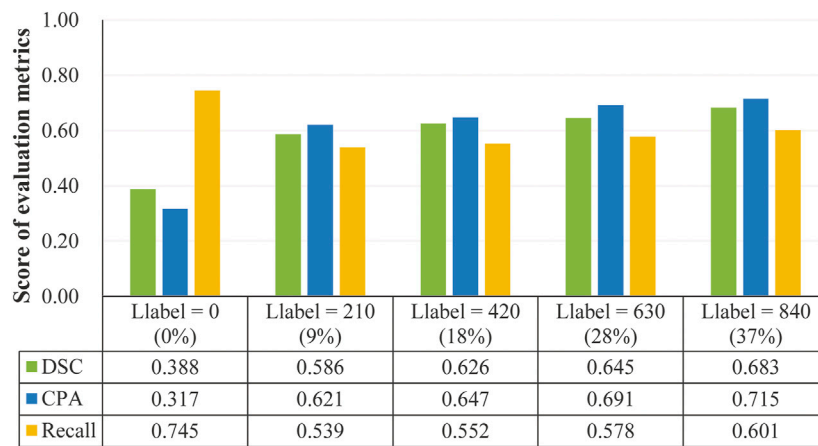


FIGURE 7

Segmentation performance obtained by the proposed model when varying numbers of labeled samples were used during training the model.

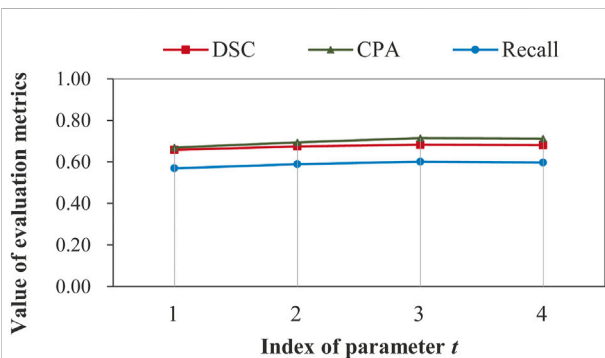


FIGURE 8

Illustration of impact of recurrent feature extraction on segmentation performance obtained by the complete model on data set D2 without image aggregation.

TABLE 4 Number of model parameters and test time for different recurrence times.

<i>t</i>	Number of model parameters (million)	Test time (sec)
1	3.654	2.19
2	9.808	2.45
3	23.545	2.60
4	46.826	4.21

Despite the significant increases of parameters, the time grows slowly.

Impact of network structure on segmentation performance: until now, the best aforementioned segmentation performance

TABLE 5 Impacts of structure changes of the feature extraction network on the segmentation performance.

Case	Dilated conv.	Inception	Residual unit	DSC
1#	—	—	—	0.632
2#	✓	—	—	0.656
3#	✓	✓	—	0.667
4#	✓	✓	✓	<b>0.683</b>

The bold value in each column indicates the maximal one.

was obtained when the feature extraction network was structured as stacked dilated residual convolutions combined with the inception block. We call such a model a *complete model*. It is required to explore whether the segmentation performance varies as the structure of the feature extraction network changes. Table 5 reports the scores of *DSC* under different cases, with each indicating one type of setting, where Case 4# denotes the one with which our model achieved the best segmentation performance.

We can see from Table 5 that the value of *DSC* increases steadily as the feature extraction network approaches the “perfect” structure as shown by Case 4#. An absolute increase of 0.051 for *DSC* shows superiority of stacked dilated residual convolution layers for automatically extracting the representative features of bone metastasis lesions from low-resolution SPECT images.

Impact of image aggregation on segmentation performance: image aggregation operation conducted on ‘optimal’ data set D2 outputs an aggregated data set D2\_Agg. There are 1,140 aggregated samples in data set D2\_Agg. Image or feature aggregation is detailed in subsection 2.1.3. In Figure 9,

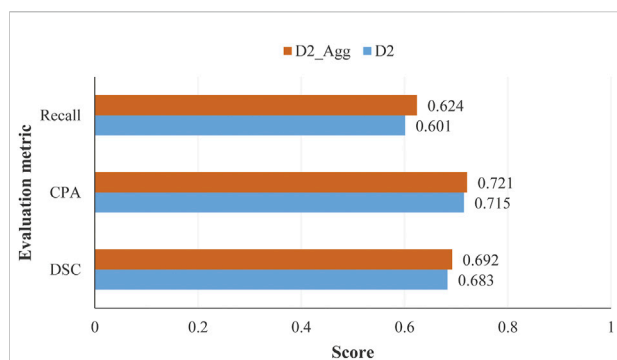


FIGURE 9

Scores of evaluation metrics obtained by the proposed model on test samples in data sets D2 and D2\_Agg, with the number of labeled samples used for training the model being  $L_{label} = 840$  and  $L_{label} = 420$  for D2 and D2\_Agg, respectively.

TABLE 6 Comparative analysis between the proposed model and existing models with test samples in data set D2\_Agg.

$L_{label}$	Model	DSC	CPA	Recall
798	U-Net	0.695	0.722	0.608
798	nnU-Net	0.757	0.772	0.712
420	Proposed	0.692	0.721	0.624

we have shown the scores of several evaluation metrics on data sets D2 and D2\_Agg.

We can see from Figure 9 that the image aggregation operation slightly improves the performance, with an increase of 0.006 for CPA when the proposed model was trained in a semi-supervised manner with 840 labeled samples.

### 3.4.3 Comparative analysis

A comparative analysis was conducted between the proposed model and existing classical models which included U-Net (Ronneberger et al., 2015) and its variant nnU-Net (Isensee et al., 2021). The U-Net and its variants are well-known supervised models. We therefore selected 70% of the samples (i.e.,  $L_{label} = 798$ ) for training U-Net and nnU-Net. Our semi-supervised model was trained using 37% of samples (i.e.,  $L_{label} = 420$ ). Table 6 reports the scores of evaluation metrics achieved by several models.

It can be seen that the existing supervised models perform slightly better than the proposed semi-supervised model. On the whole, however, our model obtains comparable performance with  $DSC = 0.692$ . The model U-Net and its variants were originally designed for supervised learning, which cannot be trained in a semi- or unsupervised manner. The encoder-decoder structure combined with the skip connection that the U-Net has greatly inspires us to develop better semi-

supervised models using more samples of SPECT images in the near future.

## 4 Discussion

In this section, we provide a brief discussion about the proposed semi-supervised segmentation model on identifying and delineating bone metastasis lesions in  $^{99m}\text{TC}$ -MDP SPECT images. This section begins with a visual presentation of the segmented images, which is followed by an analysis on the reasons that account for the imperfect performance.

With regional images in the thorax acquired from two patients with metastasis, Figure 10 shows the segmented areas by our model, where the original images and manual labels are also presented.

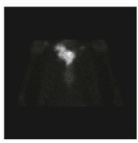
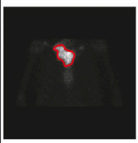
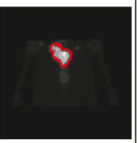
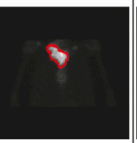
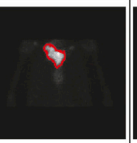
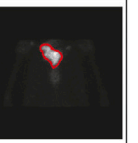
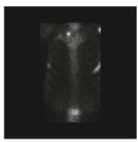
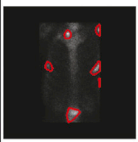
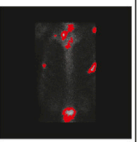
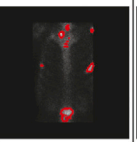
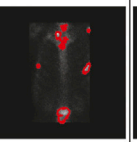
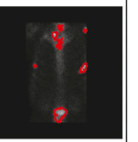
The visual presentation depicted in Figure 10 shows that our model performs better on segmenting the single-lesion metastasis (patient 1#) than the multi-lesion one (Patient 2#). There is almost no difference between the manually labeled and the automatically delineated regions of the image acquired from Patient 1#, who was clinically diagnosed with bone metastasis in the collarbone. There is also no noticeable improvement in performance as the number of labeled samples significantly increases during the training model. By contrast, there is quite a distinction in terms of the size of the lesion between the manually labeled and automatically delineated areas of the image acquired from Patient 2#, who was clinically diagnosed with bone metastasis in the collarbone, left scapula, ribs, and lumbar vertebrae simultaneously.

Now, we present the possible reasons that negatively affect segmentation performance as follows.

**Imperfect manual annotation:**  $^{99m}\text{TC}$ -MDP SPECT imaging is typically characterized by the inferior spatial resolution, which brings a significant challenge to the pixel-level annotation by human researchers. The situation will get much worse when multiple tiny lesions are present in an image (e.g., the one from Patient 2# in Figure 10). Any error of manual annotation may result in incorrect classification of pixels by the automated segmentation model. The more the pixels have been correctly classified, the higher the values for DSC, CPA, and Recall are. Therefore, the imperfect manual annotation mainly accounts for the decreased segmentation performance.

**Insufficient feature representation:** the proposed deep segmentation model segments metastatic lesions via first extracting the hierarchical features of lesions and then classifying the pixel-level features into classes. However, learning the representative features of metastasis lesions from a small-scale data set of low-resolution images is significantly challenging as metastasis lesions are commonly distributed irregularly and typically show variability in size, shape, and intensity of radiopharmaceutical uptake. Although data augmentation, especially the geometric transformation-based operations, positively contributes to improving segmentation



Patient	Original image	Manual label	$L_{label}=210$	$L_{label}=420$	$L_{label}=630$	$L_{label}=840$
Patient 1#						
Patient 2#						

**FIGURE 10**  
Visual presentation of manually labeled lesions and automatically segmented lesions by our semi-supervised deep segmentation model.

performance, more samples are still needed to be used for extracting deeper features of bone metastasis lesions.

Now, we can summarize that the proposed semi-supervised segmentation model has the potential to be used to automatically detect and delineate bone metastasis lesions with low-resolution SPECT images. A score of 0.683 for DSC has been obtained by the proposed deep model on the augmented data set if only 37% ( $\approx 840/2280$ ) of the labeled samples were used for training the model. In the case that the model was trained with the aggregated samples, i.e., 37% ( $\approx 420/1140$ ) of the labeled images, a score of 0.692 was obtained for *DSC*, achieving comparable segmentation performance.

5 Conclusion

To facilitate the clinical diagnosis of skeletal metastasis by nuclear medicine physicians, in this work, we have proposed a semi-supervised segmentation model to automatically detect and delineate bone metastasis lesions in the regional SPECT images. The proposed model was presented by detailing the structures of feature extraction and pixel-level feature classification stages. Experimental data of clinical SPECT bone scan images and the data augmentation methods used were also elaborated. The experimental evaluation conducted on these images has shown that the proposed model has the potential to be used as a clinical tool for automatically delineating the boundaries of bone metastasis lesions in low-resolution images, achieving a best mean score of 0.692 for *DSC*, if the model was trained using 37% of the aggregated samples with manual labels.

We plan to extend our work in two directions in the future. First, we intend to collect more data of clinical SPECT images and fine-tune the proposed semi-supervised model such that it can work in computer-aided diagnosis systems. Second, we plan to develop models for whole-body SPECT image segmentation, enabling automated detection and delineation of multi-lesion bone metastasis.

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.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of Gansu Provincial Hospital.

Author contributions

Conceptualization: QL, RG, and ML; methodology: QL, RG, and ML; software: RG and ML; validation: QL, RG, and ML; formal analysis: QL, ZM, and YC; investigation: QL; resources: HW and RW; data curation: HW, RW, and QL; writing—original draft preparation: QL and ML; writing—review and editing: QL, RG, and ML; visualization: RG and ML; supervision: QL; project administration: QL; funding acquisition: QL and ZM. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Key R&D Plan of Gansu Province (21YF5GA063), the Youth Ph.D. Foundation of Education Department of Gansu Province (2021QB-063), the Natural Science Foundation of Gansu Province (20JR5RA511), the Fundamental Research Funds for the Central Universities (31920220020, 31920220054, 31920210013), the National Natural Science Foundation of China (61562075), the Gansu Provincial First-class Discipline Program of Northwest Minzu

University (11080305), and the Program for Innovative Research Team of SEAC ([2018] 98).

## 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.

## References

- Apiparakoon, T., Rakratchatakul, N., Chantadisai, M., Vutrpongwatana, U., Kingetch, K., Sirisalipoch, S., et al. (2020). MaligNet: Semisupervised learning for bone lesion instance segmentation using bone scintigraphy. *IEEE Access* 8, 27047–27066. doi:10.1109/access.2020.2971391
- Asgari Taghanaki, S., Abhishek, K., Cohen, J. P., Cohen-Adad, J., and Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: A review. *Artif. Intell. Rev.* 54, 137–178. doi:10.1007/s10462-020-09854-1
- Aslanta, A., Dandl, E., Akroli, M., and Cakiroğlu, M. (2016). Cadboss: A computer-aided diagnosis system for whole-body bone scintigraphy scans. *J. Cancer Res. Ther.* 12 (2), 787–792. doi:10.4103/0973-1482.150422
- Bochkovskiy, A., Wang, C. Y., and Liao, H. (2020). YOLOv4: Optimal speed and accuracy of object detection.
- Chan, T. F., and Vese, L. A. (2001). Active contours without edges. *IEEE Trans. Image Process.* 10 (2), 266–277. doi:10.1109/83.902291
- Cheimariotis, G., Al-Mashat, M., Haris, K., Aletras, A. H., Jogi, J., Bajc, M., et al. (2018). Automatic lung segmentation in functional SPECT images using active shape models trained on reference lung shapes from CT. *Ann. Nucl. Med.* 32, 94–104. doi:10.1007/s12149-017-1223-y
- Chen, J., and Frey, E. C. (2020). *Medical image segmentation via unsupervised convolutional neural network*.
- Cheng, D. C., Hsieh, T. C., Yen, K. Y., and Kao, C. H. (2021). Lesion-based bone metastasis detection in chest bone scintigraphy images of prostate cancer patients using pre-train, negative mining, and deep learning. *Diagnostics* 11 (3), 518. doi:10.3390/diagnostics11030518
- Cheng, D. C., Liu, C. C., Hsieh, T. C., Yen, K. Y., and Kao, C. H. (2021). Bone metastasis detection in the chest and pelvis from a whole-body bone scan using deep learning and a small dataset. *Electronics* 10, 1201. doi:10.3390/electronics10101201
- Christoph, B., Albarqouni, S., and Navab, N. (2017). “Semi-supervised deep learning for fully convolutional networks,” in International Conference on Medical Image Computing and Computer-Assisted Intervention, 311–319.
- Costelloe, C. M., Rohren, E. M., Madewell, J. E., Hamaoka, T., Theriault, R. L., Yu, T. K., et al. (2009). Imaging bone metastases in breast cancer: Techniques and recommendations for diagnosis. *Lancet. Oncol.* 10 (6), 606–614. doi:10.1016/S1470-2045(09)70088-9
- Dang, J. (2016). *Classification in bone scintigraphy images using convolutional neural networks*. master's thesis: Lund University.
- Doulamis, N., and Doulamis, A. (2014). “Semi-supervised deep learning for object tracking and classification,” in 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014, 848–852.
- Elfarra, F. G., Calin, M. A., and Parasca, S. V. (2019). Computer-aided detection of bone metastasis in bone scintigraphy images using parallelepiped classification method. *Ann. Nucl. Med.* 33 (11), 866–874. doi:10.1007/s12149-019-01399-w
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial networks. *Adv. Neural Inf. Process. Syst.* 3, 2672–2680.
- Guo, Y., Lin, Q., Zhao, S., Li, T., Cao, Y., Man, Z., et al. (2022). Automated detection of lung cancer-caused metastasis by classifying scintigraphic images using convolutional neural network with residual connection and hybrid attention mechanism. *Insights Imaging* 13, 24. doi:10.1186/s13244-022-01162-2
- He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2020). Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* 42 (2), 386–397. doi:10.1109/TPAMI.2018.2844175
- Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi:10.1038/s41592-020-01008-z
- Lei, T., Wang, R., Wan, Y., Du, X., Meng, H., and Nandi, A. (2020). *Medical image segmentation using deep learning: A survey*, 13120.
- Li, T., Lin, Q., Guo, Y., Zhao, S., Zeng, X., Man, Z., et al. (2022). Automated detection of skeletal metastasis of lung cancer with bone scans using convolutional nuclear network. *Phys. Med. Biol.* 67, 015004. doi:10.1088/1361-6560/ac4565
- Liang, M., and Hu, X. L. (2015). “Recurrent convolutional neural network for object recognition,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 07–12 June 2015, 3367–3375.
- Lin, Q., Cao, C., Li, T., Cao, Y., Man, Z., and Wang, H. (2021). Multiclass classification of whole-body scintigraphic images using a self-defined convolutional neural network with attention modules. *Med. Phys.* 48 (10), 5782–5793. doi:10.1002/mp.15196
- Lin, Q., Cao, C., Li, T., Man, Z., Cao, Y., and Wang, H. (2021). dSPIC: A deep SPECT image classification network for automated multi-disease, multi-lesion diagnosis. *BMC Med. Imaging* 21, 122. doi:10.1186/s12880-021-00653-w
- Lin, Q., Li, T., Cao, C., Cao, Y., Man, Z., and Wang, H. (2021). Deep learning based automated diagnosis of bone metastases with SPECT thoracic bone images. *Sci. Rep.* 11, 4223. doi:10.1038/s41598-021-83083-6
- Lin, Q., Luo, M., Gao, R., Li, T., Man, Z., Cao, Y., et al. (2020). Deep learning based automatic segmentation of metastasis hotspots in thorax bone SPECT images. *PLoS ONE* 15 (12), e0243253. doi:10.1371/journal.pone.0243253
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi:10.1016/j.media.2017.07.005
- Mac, A., Fgeb, C., and Svp, D. (2021). Object-oriented classification approach for bone metastasis mapping from whole-body bone scintigraphy. *Phys. Med.* 84, 141–148. doi:10.1016/j.ejmp.2021.03.040
- Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. (2022). Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (7), 3523–3542. doi:10.1109/TPAMI.2021.3059968
- Moon, D. H., Maddahi, J., Silverman, D. H., Glaspy, J. A., Phelps, M. E., and Hoh, C. K. (1998). Accuracy of whole-body fluorine-18-FDG PET for the detection of recurrent or metastatic breast carcinoma. *J. Nucl. Med.* 39 (3), 431–435.
- Nathan, M., Gnanasegaran, G., Adamson, K., and Fogelman, I. (2013). *Bone scintigraphy: Patterns, variants, limitations and artefacts*. Berlin Heidelberg: Springer.
- Papadrianos, N., Papageorgiou, E., and Anagnostis, A. (2020). Development of Convolutional Neural Networks to identify bone metastasis for prostate cancer patients in bone scintigraphy. *Ann. Nucl. Med.* 34, 824–832. doi:10.1007/s12149-020-01510-6
- Papadrianos, N., Papageorgiou, E., Anagnostis, A., and Papageorgiou, K. (2020). Bone metastasis classification using whole body images from prostate cancer patients based on convolutional neural networks application. *PLoS ONE* 15 (8), e0237213. doi:10.1371/journal.pone.0237213
- Papadrianos, N., Papageorgiou, E., Anagnostis, A., and Papageorgiou, K. (2020). Efficient bone metastasis diagnosis in bone scintigraphy using a fast convolutional neural network architecture. *Diagnostics* 10 (8), 532. doi:10.3390/diagnostics10080532
- Pi, Y., Zhao, Z., Xiang, Y., Li, Y., Cai, H., and Yi, Z. (2020). Automated diagnosis of bone metastasis based on multi-view bone scans using attention-augmented deep neural networks. *Med. Image Anal.* 65, 101784. doi:10.1016/j.media.2020.101784
- Radford, A., Metz, L., and Chintala, S. (2015). *Unsupervised representation learning with deep convolutional generative adversarial networks*.
- Redmon, J., and Farhadi, A. (2018). YOLOv3: An incremental improvement.

## 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, editors, and 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.

- Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional networks for biomedical image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 234–241.
- Sadik, M., Hamadeh, I., Nordblom, P., Suurkula, M., Hoglund, P., Ohlsson, M., et al. (2008). Computer-assisted interpretation of planar whole-body bone scans. *J. Nucl. Med.* 49, 1958–1965. doi:10.2967/jnumed.108.055061
- Sadik, M., Jakobsson, D., Olofsson, F., Ohlsson, M., Suurkula, M., and Edenbrandt, L. (2006). A new computer-based decision-support system for the interpretation of bone scans. *Nucl. Med. Commun.* 27, 417–423. doi:10.1097/00006231-200605000-00002
- Sderlund, V. (1996). Radiological diagnosis of skeletal metastases. *Eur. Radiol.* 6, 587–595. doi:10.1007/BF00187654
- Shan, H., Jia, X., Yan, P., Li, Y., Paganetti, H., and Wang, G. (2020). Synergizing medical imaging and radiotherapy with deep learning. *Mach. Learn. Sci. Technol.* 1, 021001. doi:10.1088/2632-2153/ab869f
- Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 60. doi:10.1186/s40537-019-0197-0
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2014). "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Tarvainen, A., and Valpola, H. (2017). "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in Proceedings of the 31st International Conference on Neural Information Processing Systems, 1–10.
- Thorwarth, R., Dewaraja, Y., Wilderman, S., Kaminski, M., Avram, A., and Roberson, P. (2013). SU-E-J-186: Automated SPECT based segmentation for quality assurance of CT-delineated tumor volumes for 131I tositumomab therapy of non-hodgkins lymphoma. *Med. Phys.* 40, 194. doi:10.1118/1.4814398
- Yu, F., and Koltun, V. (2015). "Multi-scale context aggregation by dilated convolutions," in International Conference on Learning Representations.
- Zhao, Z., Pi, Y., Jiang, L. S., Xiang, Y., Wei, J., Yang, P., et al. (2020). Deep neural network based artificial intelligence assisted diagnosis of bone scintigraphy for cancer bone metastasis. *Sci. Rep.* 10, 17046. doi:10.1038/s41598-020-74135-4
- Zhu, C., Tian, L., Chen, P., Wang, L., Ye, G., and Mao, Z. (2008). Application of GVF snake model in segmentation of whole body bone SPECT image. *J. Biomed. Eng.* 25 (1), 27–29.



## OPEN ACCESS

## EDITED BY

Deepak Kumar Jain,  
Chongqing University of Posts and  
Telecommunications, China

## REVIEWED BY

Sameet Mehta,  
Yale University, United States  
Stephen R. Piccolo,  
Brigham Young University, United States

## \*CORRESPONDENCE

Rahee Walambe,  
rahee.walambe@sitpune.edu.in  
Ketan Kotecha,  
drketankotecha@gmail.com  
Satyajeet P. Khare,  
satyajeet.khare@ssbs.edu.in

## SPECIALTY SECTION

This article was submitted to Molecular  
Diagnostics and Therapeutics,  
a section of the journal  
Frontiers in Molecular Biosciences

RECEIVED 29 March 2022

ACCEPTED 28 September 2022

PUBLISHED 07 November 2022

## CITATION

Bhandari N, Walambe R, Kotecha K and  
Khare SP (2022), A comprehensive  
survey on computational learning  
methods for analysis of gene  
expression data.  
*Front. Mol. Biosci.* 9:907150.  
doi: 10.3389/fmolb.2022.907150

## COPYRIGHT

© 2022 Bhandari, Walambe, Kotecha  
and Khare. This is an open-access article  
distributed under the terms of the  
[Creative Commons Attribution License](#)  
(CC BY). The use, distribution or  
reproduction in other forums is  
permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original  
publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or  
reproduction is permitted which does  
not comply with these terms.

# A comprehensive survey on computational learning methods for analysis of gene expression data

Nikita Bhandari<sup>1</sup>, Rahee Walambe<sup>2,3\*</sup>, Ketan Kotecha<sup>1,3\*</sup> and  
Satyajeet P. Khare<sup>4\*</sup>

<sup>1</sup>Computer Science Department, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India, <sup>2</sup>Electronics and Telecommunication Department, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India, <sup>3</sup>Symbiosis Center for Applied AI (SCAAI), Symbiosis International (Deemed University), Pune, India, <sup>4</sup>Symbiosis School of Biological Sciences, Symbiosis International (Deemed University), Pune, India

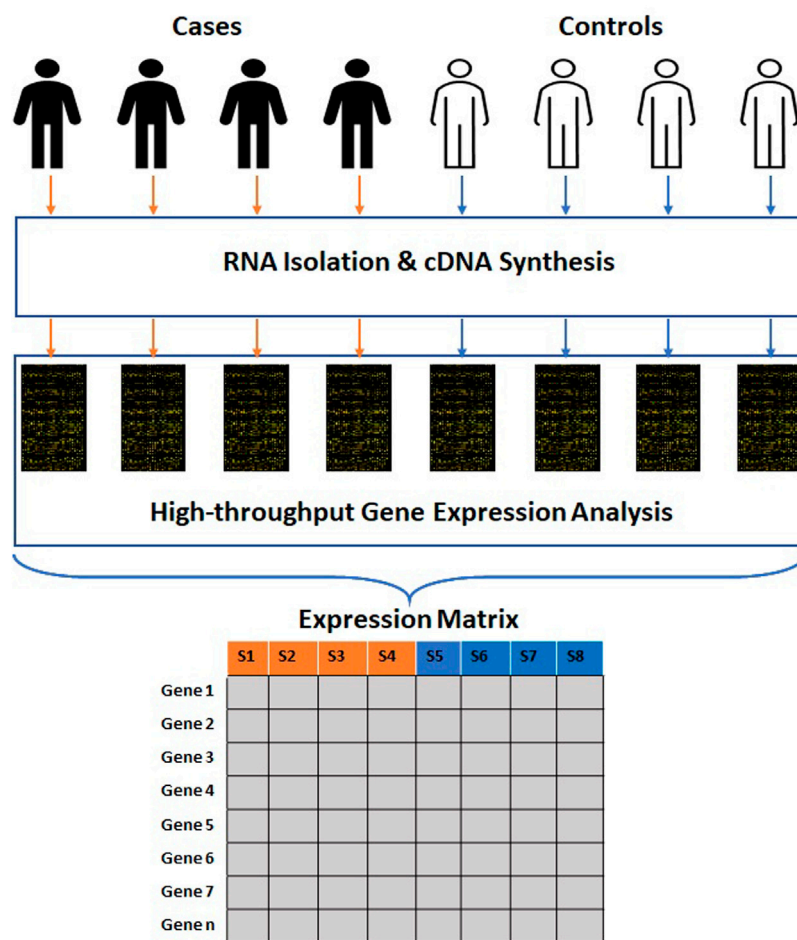
Computational analysis methods including machine learning have a significant impact in the fields of genomics and medicine. High-throughput gene expression analysis methods such as microarray technology and RNA sequencing produce enormous amounts of data. Traditionally, statistical methods are used for comparative analysis of gene expression data. However, more complex analysis for classification of sample observations, or discovery of feature genes requires sophisticated computational approaches. In this review, we compile various statistical and computational tools used in analysis of expression microarray data. Even though the methods are discussed in the context of expression microarrays, they can also be applied for the analysis of RNA sequencing and quantitative proteomics datasets. We discuss the types of missing values, and the methods and approaches usually employed in their imputation. We also discuss methods of data normalization, feature selection, and feature extraction. Lastly, methods of classification and class discovery along with their evaluation parameters are described in detail. We believe that this detailed review will help the users to select appropriate methods for preprocessing and analysis of their data based on the expected outcome.

## KEYWORDS

gene expression, microarray, machine learning, deep learning, missing value imputation, feature selection, interpretation, explainable techniques

## 1 Introduction

A genome is a complete set of genes in an organism. Genomics is a study of the information structure and function programmed in the genome. Genomics has applications in multiple fields, including medicine (Chen et al., 2018; Lai et al., 2020; Huang et al., 2021), agriculture (Abberton et al., 2016; Parihar et al., 2022), industrial biotechnology (Alloul et al., 2022), synthetic biology (Baltes and Voytas, 2015), etc.

**FIGURE 1**

Process of generation of high-throughput gene expression data. The clinical samples are subjected to RNA isolation and cDNA synthesis. The cDNAs are subjected to high-throughput gene expression analysis. The raw data obtained from these methods is further transmuted into a numerical matrix where rows and columns represent genes and samples.

Researchers working in these domains create and use a variety of data such as DNA, RNA, and protein sequences, gene expression, gene ontology, protein-protein interactions (PPI), *etc.*

Genomics data can be broadly classified into sequence and numeric data (e.g., gene expression matrix). The DNA sequence information can be determined by first generation (Sanger, Nicklen and Coulson, 1977), second generation sequencing (Margulies et al., 2005; Shendure et al., 2005; Bentley et al., 2008; Valouev et al., 2008) or third generation sequencing (Harris et al., 2008; Eid et al., 2009; Eisenstein, 2012; Rhoads and Au, 2015) methods. The second and third generation sequencing are together referred to as Next Generation Sequencing (NGS). Applications of DNA sequence analysis include prediction of protein sequence and structure, molecular phylogeny, identification of intrinsic features, sequence variations, *etc.* Common implementations of these applications include splice

site detection (Nguyen et al., 2016; Fernandez-Castillo et al., 2022), promoter prediction (Umarov and Solovyev, 2017; Bhandari et al., 2021), classification of diseased related genes (Peng, Guan and Shang, 2019; Park, Ha and Park, 2020), identification of protein binding sites (Pan and Yan, 2017; Uhl et al., 2021), biomarker discovery (Arbitrio et al., 2021; Frommlet et al., 2022), *etc.* The numeric data often generated from functional genomics studies include gene expression, single nucleotide polymorphism (SNP), DNA methylation, *etc.* Microarray and NGS technologies are the tools of choice for functional genomics studies. The functional genomics that deals with high-throughput study of gene expression is referred to as transcriptomics.

Gene expression data, irrespective of the platform used (e.g., microarray, NGS, *etc.*), contains the expression levels of thousands of genes experimentally evaluated in various



TABLE 1 Expression array repositories.

Name	Link	References
<b>Primary databases</b>		
Gene Expression Omnibus (GEO)	<a href="https://www.ncbi.nlm.nih.gov/geo/">https://www.ncbi.nlm.nih.gov/geo/</a>	Barrett <i>et al.</i> (2013)
ArrayExpress (AE)	<a href="https://www.ebi.ac.uk/arrayexpress/">https://www.ebi.ac.uk/arrayexpress/</a>	Brazma <i>et al.</i> (2003)
Genomic Expression Archive (GEA)	<a href="https://www.ddbj.nig.ac.jp/gea/">https://www.ddbj.nig.ac.jp/gea/</a>	Kodama <i>et al.</i> (2019)
<b>Secondary and domain specific databases</b>		
The Cancer Genome Atlas (TCGA)	<a href="https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga">https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga</a>	Tomczak, Czerwińska and Wiznerowicz, (2015)
BioDataome	<a href="http://dataome.mensxmachina.org/">http://dataome.mensxmachina.org/</a>	Lakiotaki <i>et al.</i> (2018)
RefDIC	<a href="http://refdic.rcai.riken.jp/welcome.cgi">http://refdic.rcai.riken.jp/welcome.cgi</a>	Hijikata <i>et al.</i> (2007)

conditions. Gene expression analysis helps us understand gene networks and molecular pathways. Gene expression information can be utilized for basic as well as clinical research (Behzadi, Behzadi and Ranjbar, 2014; Chen *et al.*, 2016; Karthik and Sudha, 2018; Kia *et al.*, 2021). In disease biology, gene expression analysis provides an excellent tool to study the molecular basis of disease as well as the identification of markers for diagnosis, prognosis, and drug discovery. Therefore, for this review, we will focus on computational methods in the analysis of gene expression data.

The data produced by microarray as well as NGS-based RNA sequencing goes through multiple phases of quality check before analysis. This data is further transmuted to a numerical matrix (Figure 1) where rows and columns represent genes and samples. The numeric value in each cell of a matrix links the expression level of a specific feature gene to a particular sample. The expression matrix is generally a flat dataset as the number of features is very high compared to the number of samples. Some of the standard DNA microarray platforms available are Affymetrix (Pease *et al.*, 1994), Agilent (Blanchard, Kaiser and Hood, 1996), *etc.* Some of the standard commercial NGS platforms are Illumina (Bentley *et al.*, 2008), Ion torrent (Rothberg *et al.*, 2011) *etc.* The massive amount of data generated from publicly funded research is available through open access repositories such as Gene Expression Omnibus (GEO), ArrayExpress, Genomic Expression Archive (GEA), *etc.* (Table 1).

Identification of differentially expressed genes is the most common application in gene expression analysis. This type of class comparison analysis can be achieved using basic statistical techniques, for example, chi-squared test, *t*-test, ANOVA, *etc.* (Segundo-Val and Sanz-Lozano 2016). Commonly used packages for microarray-based gene expression analysis include limma (Smyth, 2005), affy (Gautier *et al.*, 2004), lumi (Du, Kibbe and Lin, 2008), oligo (Carvalho and Irizarry, 2010); whereas, those for RNA sequencing analysis include EdgeR

(Robinson, McCarthy and Smyth, 2009) and DESeq2 (Love, Huber and Anders, 2014). The classification and regression problems on the other hand depend on classical linear and logistic regression analysis. However, the data typically generated by the transcriptomic technologies creates a need for penalized or modified prospects as a solution to the problems of high dimensionality and overfitting (Turgut, Dagtekin and Ensari, 2018; Morais-Rodrigues *et al.*, 2020; Tabares-Soto *et al.*, 2020; Abapihi *et al.*, 2021). The development of high-end computational algorithms, such as machine learning techniques, has created a new dimension for gene expression analysis.

Machine learning (ML) is an artificial intelligence-based approach that emphasizes building a system that learns automatically from data and improves performance without being explicitly programmed. ML models are trained using a significant amount of data to find hidden patterns required to make decisions (Winston, 1992; Dick, 2019; Micheuz, 2020). Artificial Neural Network (ANN), Classification and regression Trees (CART), Support vector machine (SVM), and vector quantization are some of the architectures used in ML. Recent advancement in the ML domain is deep learning (DL) which is based on artificial neural networks (ANN) (Deng and Yu, 2014; LeCun, Bengio and Hinton, 2015). ANN architectures comprise input, hidden, and output layers of neurons. When more than one hidden layer is used, the ANN method is referred to as the DL method. Basic ML and DL models can work on lower-end machines with less computing power; however, DL models require more powerful hardware to process vast and complex data.

ML techniques, in general, are broadly categorized into supervised and unsupervised learning methods (Jenike and Albert, 1984; Dayan, 1996; Kang and Jameson, 2018; Yuxi, 2018). Supervised learning, which makes use of well-labelled data, is applied for classification and regression analysis. A labelled dataset is used for the training process, which later

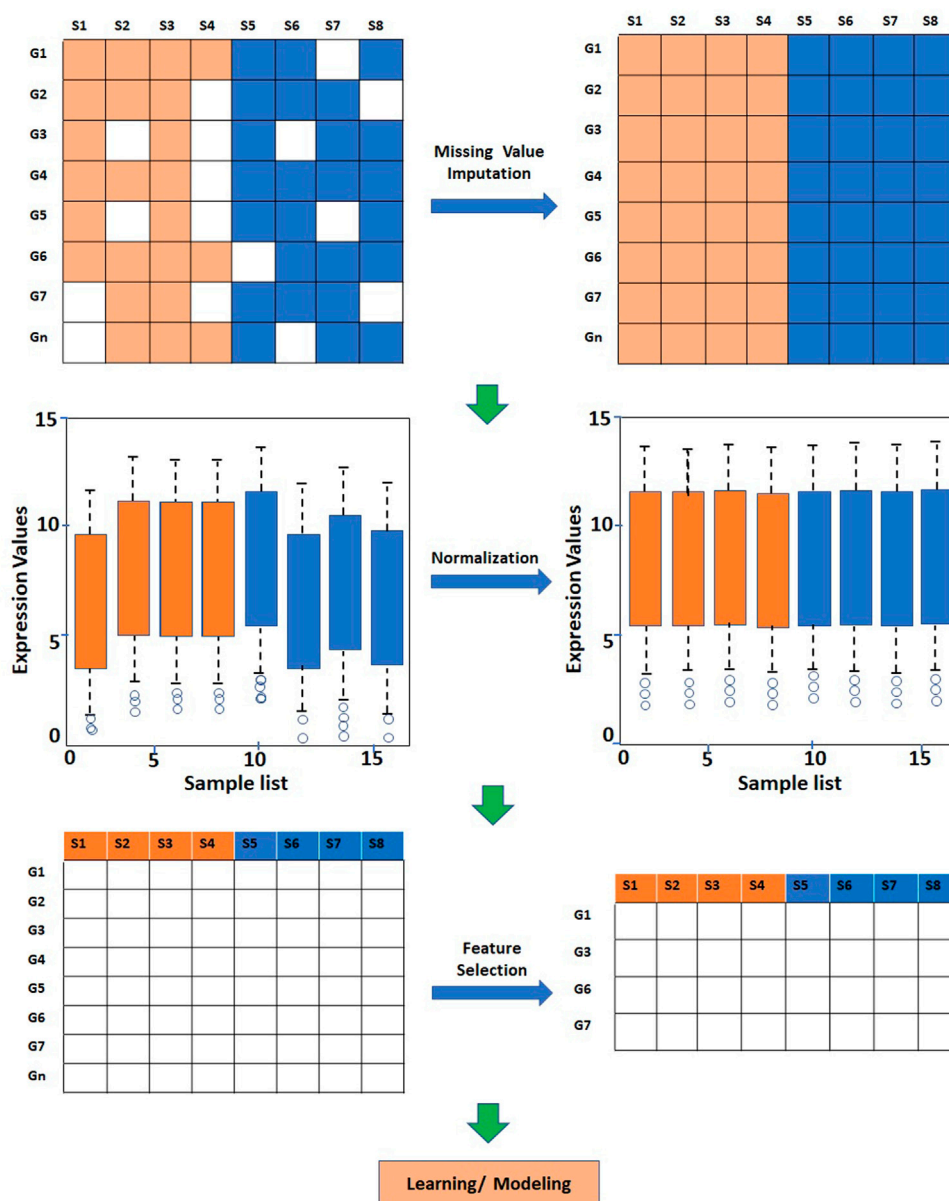


FIGURE 2

Steps involved in preprocessing and analysis of gene expression data. The raw gene expression data may get subjected to missing value imputation, normalization, and feature reduction depending on the type of analysis.

produces an inferred function to make predictions about unknown instances. Classification techniques train the model to separate the input into different categories or labels (Kotsiantis, 2007). Regression techniques train the model and predict continuous numerical value as an output based on input variables (Fernández-Delgado et al., 2019). Unsupervised techniques, on the other hand, let the model discover information or unknown patterns from the data. We can roughly divide unsupervised learning into clustering and association rules. Clustering used for class discovery is the

task of grouping a set of instances in such a way that samples in the same group or cluster are more similar in their properties than the samples in other groups or clusters. Association rules associate links between data instances inside large databases (Kotsiantis and Kanellopoulos, 2006).

The supervised ML techniques have been used for binary classification e.g., identification of cases in clinical studies, as well as multiclass classification analysis e.g., grading and staging of the disease. ML techniques have been extensively used to analyze gene expression patterns in various complex diseases, such as cancer

(Sharma and Rani, 2021), Parkinson's Disease (Peng, Guan and Shang, 2019), Alzheimer's disease (Kong, Mou and Hu, 2011; Park, Ha and Park, 2020), diabetes (Li, Luo and Wang, 2019), arthritis (Liu et al., 2009; Zhang et al., 2020), etc. The classification algorithms have also contributed to biomarker identification (Jagga and Gupta, 2015), precision treatment (Toro-Domínguez et al., 2019), drug toxicity evaluation (Vo et al., 2020) etc. The unsupervised learning techniques for clustering are routinely used in transcriptomics. The clustering analysis is applied for the study of expression relationships between genes (Liu, Cheng and Tseng, 2011), extracting biologically relevant expression features (Kong et al., 2008), discovering frequent determinant patterns (Prasanna, Seetha and Kumar, 2014), etc.

In supervised and unsupervised learning, the data is subjected to preprocessing, e.g., missing value imputation, normalization, etc. (Figure 2). In supervised learning for classification analysis, the entire dataset is divided into two subsets *viz.* training and testing/validation. The training dataset, which typically comprises 70–80% of the samples, is used for the construction of a model. The training data can first be subjected to missing value imputation and feature scaling. The preprocessed data is then subjected to feature selection/extraction and model development. The model is then applied to the test/validation dataset, which is also preprocessed in a similar fashion. The preprocessing and feature selection steps are applied to the training dataset after the train-test split to avoid “data leakage”. The unsupervised learning which is based on unlabeled data, may include preprocessing steps and data-driven techniques for feature reduction.

Though missing value imputation, normalization, feature selection, and modelling are important steps in classification analysis, there appears to be very limited literature that reviews them together. Most of the reviews focus either on missing value imputation, features selection, or learning/modelling (Quackenbush, 2001; Dudoit and Fridlyannnd, 2005; Chen et al., 2007; Liew, Law and Yan, 2011; Sahu, Swarnkar and Das, 2011; Yip, Amin and Li, 2011; Khatri, Sirota and Butte, 2012; Tyagi and Mishra, 2013; Bolón-Canedo et al., 2014; Li et al., 2015; Manikandan and Abirami, 2018; Hambali, Oladele and Adewole, 2020; Zhang, Jonassen and Goksøyr, 2021). This creates gaps in understanding of the complete pipeline of the analysis process for researchers from different domains. The objective of this review is to bridge these gaps. Here we discuss various ways to analyze gene expression data and computational methods used at each step. Through this comprehensive review, we also discuss the need for interpretability to provide insights and bring trust to the predictions made. The review is organized into 6 sections. Section 2 broadly covers different missing value imputation approaches along with their advantages and limitations. Section 3 discusses feature scaling techniques applied to gene expression data. In Section 4, broad categories of feature selection and dimensionality reduction techniques are discussed. Section 5 covers the different types of gene expression analyses, including class comparison, classification (class prediction), and class discovery. In Section 6, we discuss conclusions and future directions.

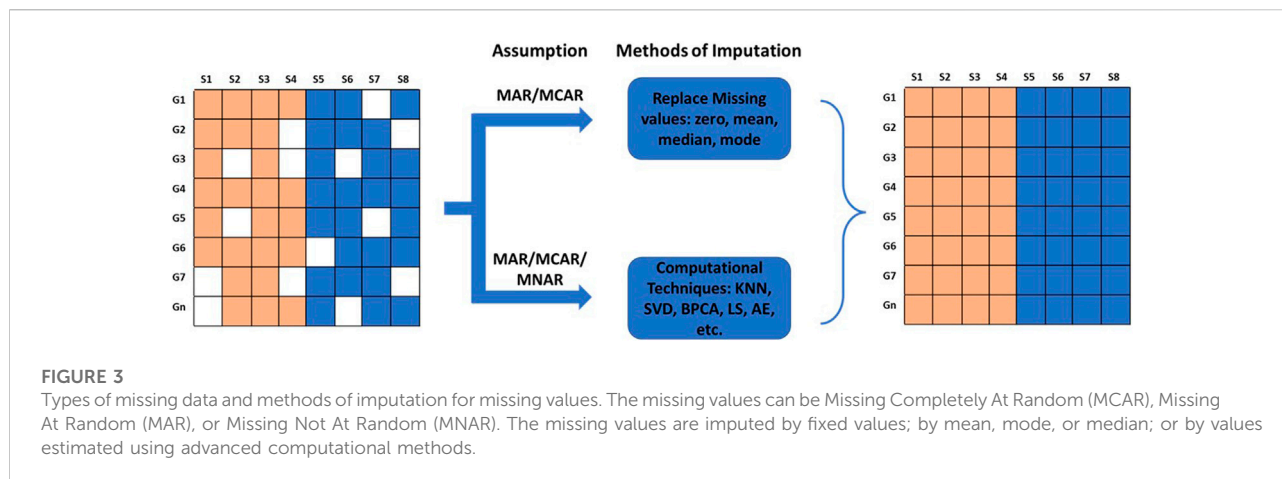
## 2 Missing value imputation

Gene expression matrices are often riddled with missing gene expression values due to various reasons. In this section, we will discuss sources of missing values and various computational techniques utilized to perform the imputation of missing values. Missing data are typically grouped into three categories: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) (Rubin, 1976; Schafer and Graham, 2002; Aydılek and Arslan, 2013; Mack, Su and Westreich, 2018) (Figure 3). In MCAR, the missing data is independent of their unobserved values and independent of the observed data. In other words, the data is completely missing at random, independent of the nature of the investigation. MAR is a more general class of MCAR where conditional dependencies are accounted for. In MAR, the missingness of data is random but conditionally dependent on observed and unobserved values. In transcriptomics, it can be assumed that all MAR values are also MCAR (Lazar et al., 2016); for example, a channel signal obscured accidentally by a dust particle. However, in meta-analysis, a missing value can be attributable to a specific dataset due to its architecture. In this case, the missing values are MAR and not MCAR. In MNAR, the missingness depends on the observed and/or unobserved data. In microarray analysis, values missing due to their low signal intensities are an example of MNAR data.

Missing values can be imputed using two different approaches. MCAR/MAR values are either embedded with a fixed value, or mean, median, or mode. However, this method creates lots of similar values if missing data is high. MCAR/MAR and MNAR values can be imputed using advanced computational techniques. The choice of imputation method depends on the accuracy of the results obtained from the downstream analysis. Computational techniques for estimating missing values can be categorized into four different approaches: Global, Local, Hybrid, and Knowledge Assisted (García-Laencina et al., 2008; Moorthy et al., 2019; Farswan et al., 2020) (Table 2).

### 2.1 Global approaches

Global approaches assume homogeneity of data and use global correlation information extracted from the entire data matrix to estimate missing values. The Bayesian framework for Principal Component Analysis (BPCA) is based on a probabilistic model that can handle large variations in the expression matrix (Oba et al., 2003; Jörnsten et al., 2005; Souto, Jaskowiak and Costa, 2015). In BPCA, the missing value is replaced with a set of random values that are estimated using the Bayesian principle to obtain the relevant principal axes for regression. Singular Value Decomposition (SVD) is another global approach for missing value imputation. SVD is a matrix decomposition method for reducing a matrix to its three constituent parts (Figure 4A). A



new matrix that is similar to the original matrix is reconstructed using these constituents in order to reduce noise and impute missing values (Troyanskaya et al., 2001).

Other than the above mentioned techniques, ANN-based techniques are also being utilized for the imputation of missing gene expression values. ANN-based methods for imputation include ANNimpute (García-Laencina et al., 2008), RNNimpute (Bengio and Gingras, 1995), etc. ANNimpute utilizes MLP (Multi-Layered Perceptron) based architecture that is trained with complete observed data (Saha et al., 2017) (Figure 4D). The final weight matrix generated through this process is further used for missing value imputation. RNNimpute utilizes Recurrent Neural Network architecture-based imputation (Bengio and Gingras, 1995) (Figure 4E). Since RNN has feedback connections from its neurons, it can preserve the long-term correlation between parameters.

## 2.2 Local approaches

Local approaches utilize a potential local similarity structure to estimate missing values. For heterogeneous data, the local approach is considered to be very effective. Many local imputation methods have been proposed since 2001. These techniques use a subset of the entire data by estimating underlying heterogeneity. K-Nearest Neighbor (KNN) is a standard ML-based missing-value imputation strategy (McNicholas and Murphy, 2010; Ryan et al., 2010; Pan et al., 2011; Dubey and Rasool, 2021) (Figure 4B). A missing value is imputed by finding the samples closest to the sample from which the gene expression value is missing. It should be noted that a lower number of neighboring points (K) may lead to overfitting of data (Batista and Monard, 2002) whereas a higher K may result in underfitting. Least Square (LS) imputation technique selects a number of most correlated genes using the L2-norm and/or Pearson's correlation (Bo, Dysvik and Jonassen, 2004; Liew, Law and Yan, 2011; Dubey and Rasool, 2021). Support Vector Regression (SVR) method is a non-linear generalization of the

linear model used for the imputation of missing gene expression values (Wang et al., 2006; Oladejo, Oladele and Saheed, 2018) (Figure 4C). A significant advantage of the SVR model is that it requires less computational time than other techniques mentioned above (Wang et al., 2006). However, the change in the missing data patterns and the high fraction of missing data limits the effects of SVR. Gaussian Mixture Clustering (GMC) is another technique used for the imputation of missing values that works with highly observable data (Ouyang, Welsh and Georgopoulos, 2004).

Some studies have compared the global and local approaches for their performances. SVD and KNN require re-computation of a matrix for every missing value, which results in prolonged evaluation time (Aghdam et al., 2017). SVR, BPCA, and LS try to mine the hidden pattern from the data and seem to perform better than SVD and KNN (Sahu, Swarnkar and Das, 2011) (Tuikkala et al., 2008; Subashini and Krishnaveni, 2011; Qiu, Zheng and Gevaert, 2020).

## 2.3 Hybrid approaches

The internal correlation among genes affects the homogeneity and heterogeneity of data and, therefore, the performance of global and local imputation approaches (Liew, Law and Yan, 2011). In order to cover both homogeneous and heterogeneous data, a hybrid approach can be very effective. LinCmb is one such hybrid approach for data imputation. LinCmb (Jörnsten et al., 2005) puts more weight on local imputation if data is heterogeneous and has fewer missing values. In contrast, it puts more weight on global methods if data is homogeneous with higher missing values. LinCmb takes an ensemble of row mean, KNN, SVD, BPCA, and GMC. When evaluated, LinCmb's performance was found to be better than each technique it has ensembled. Ensemble missing data imputation method EMDI is another hybrid imputation approach composed of BPCA, matrix completion, and two types of LS and KNN estimators (Pan et al., 2011). It utilizes high-level diversity of data for the imputation of missing values. Recursive Mutual Imputation (RMI) is also a hybrid approach that comprises BPCA and LS to

TABLE 2 Various approaches of missing value imputation.

Approach	Advantages	Limitations	Methods	References
Global	Optimal performance when data is homogeneous	Poor performance when data is heterogeneous	BPCA	Jörnsten et al. (2005), Oba et al. (2003), Souto et al. (2015)
			SVD	Troyanskaya et al. (2001)
			ANNImpute	García-Laencina et al. (2008)
			RNNImpute	Bengio and Gingras (1995)
Local	Optimal performance when data is heterogeneous	Poor performance when data is homogeneous	KNNImpute	Dubey and Rasool (2021), McNicholas and Murphy (2010), Pan et al. (2011), Ryan et al. (2010)
			LSImpute	Bo et al. (2004)
			SVRimpute	Wang et al. (2006)
			GMCImpute	Ouyang et al. (2004)
Hybrid	Optimal performance regardless of local or global correlation	Sub-optimal performance when data is noisy and has high missing rates	LinCmb	Jörnsten et al. (2005)
			EMDI	Pan et al. (2011)
			RMI	Li et al. (2015)
			VAE, DAPL	Qiu et al. (2020), Qiu et al. (2018)
Knowledge-assisted	Optimal performance in presence of noisy data	Sub-optimal performance when data has high missing rates	iMISS	Hu et al. (2006)
			GOImpute	Tuikkala et al. (2006)
			POCSimpute	Gan et al. (2006)
			HALimpute	Xiang et al. (2008)

exploit global and local structures in the dataset, respectively (Li et al., 2015). ANN based autoencoders (AE) denoising autoencoder with partial loss (DAPL) (Qiu, Zheng and Gevaert, 2018) and variable autoencoders (VAE) (Qiu, Zheng and Gavaert, 2020) consist of encoder, and decoder layers. The encoder converts the input into the hidden representation and the decoder tries to reconstruct the input from the hidden representation. Hence, AE aims to produce output close to the input (García-Laencina et al., 2008).

## 2.4 Knowledge-assisted approaches

Knowledge-assisted approaches incorporate domain knowledge or external information into the imputation process. These approaches are applied when there exists a high missing rate, noisy data, or a small sample size. The solution obtained through this approach is not dependent on the global or local correlation structure that exists in the data but on the domain knowledge. Commonly used domain knowledge includes sample information such as experimental conditions, clinical information, and gene information which includes gene ontology, epigenetic profile, etc. Integrative MISSING Value Estimation (iMISS) (Hu et al., 2006) is one such knowledge-assisted imputation technique. iMISS incorporates knowledge from multiple related microarray datasets for missing value imputation. It obtains coherent neighbors set of genes for

every gene with missing data by considering reference dataset. GOImpute (Tuikkala et al., 2006) is another knowledge-assisted imputation technique that uses GO database for knowledge assistance. This method integrates the semantic similarity in the GO with the expression similarity estimated using the KNN imputation algorithm. Projection onto convex sets impute (POCSimpute) (Gan, Liew and Yan, 2006) formulates every piece of prior knowledge into a corresponding convex set to capture gene-wise correlation, array-wise correlation, and known biological constraint. After this, a convergence-guaranteed iterative procedure is used to obtain a solution in the intersection of all these sets. HALimpute (Xiang et al., 2008) utilizes epigenetic information e.g. histone acetylation knowledge for the imputation of missing values. First, it uses the mean expression values of each gene from each cluster to form an expression pattern. It obtains missing values in the sample by applying linear regression as a primary imputation and uses KNN or LS for secondary imputation. Since knowledge-based methods strongly rely on domain-specific knowledge, they may fail to estimate missing values from under-explored cases with low knowledge available (Wang et al., 2019).

Although a large number of missing value imputation methods are available to the users, there are still quite a few challenges when it comes to the application of imputation methods to the data. Firstly, there is only limited knowledge on the performance of different imputation methods on different types of missing data.



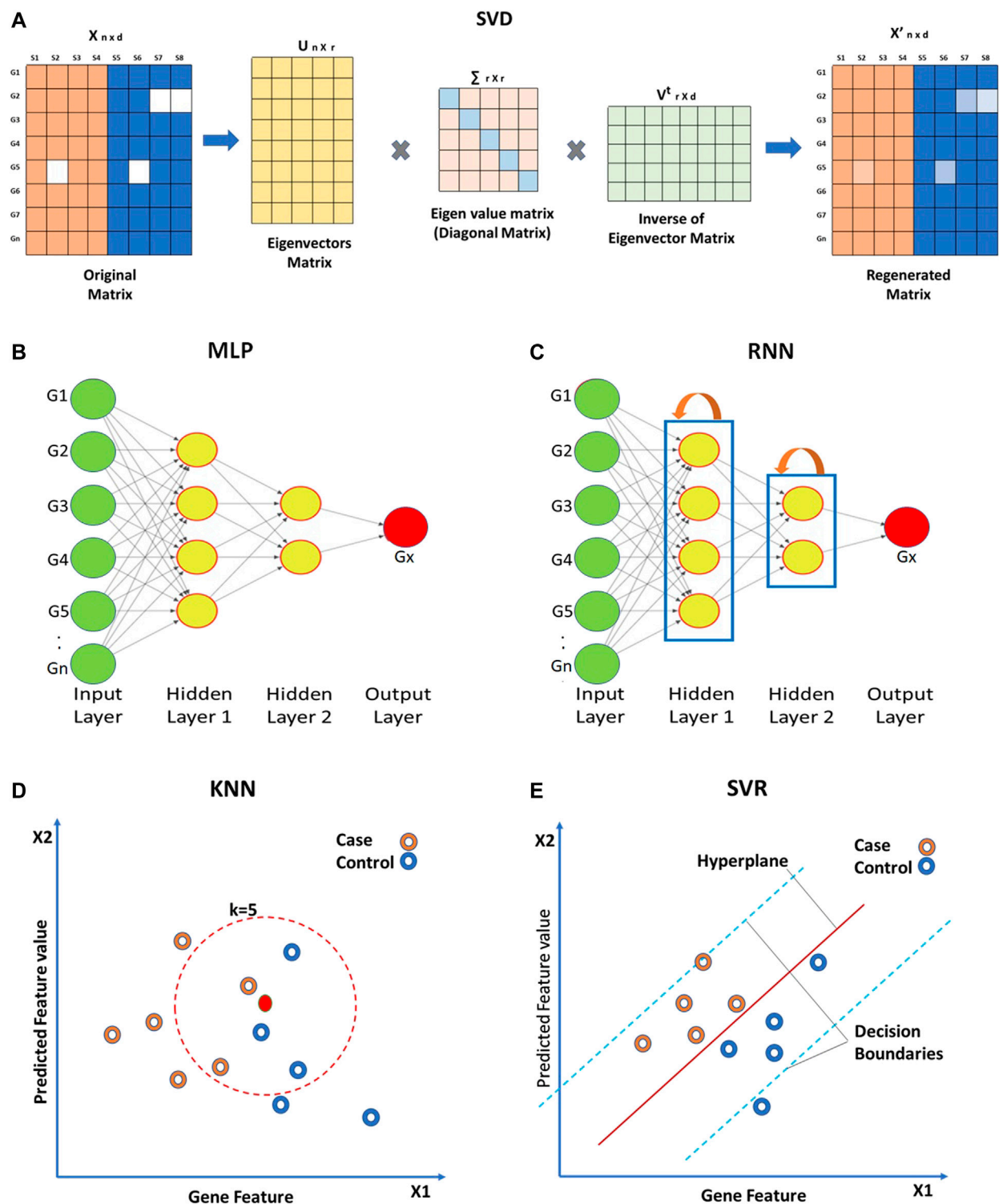


FIGURE 4

Approaches for missing value imputation. Global approaches such as (A) Singular value decomposition (SVD), (B) Multi-layered perceptron (MLP), and (C) Recurrent Neural Network use global correlation information extracted from gene expression array to estimate missing values. Local approaches such as (D) K-nearest neighbor (KNN) and (E) Support vector regression (SVR) utilize local similarity in the gene expression data. The solid green circles in (B,C) represent known gene expression values and solid red circles represent missing ones. Hollow circles in (D,E) represent samples.

The performance of the imputation methods may vary significantly depending on the experimental settings. Therefore, it is important to systematically evaluate the existing methods for their performance on different platforms and experimental settings (Aittokallio, 2009). Secondly, despite the many recent advances, better imputation algorithms that can adapt to both global and local characteristics of the data are still needed. Thirdly, the knowledge-based approaches can also be hybridized with local and/or global approaches to data imputation. More sophisticated algorithms which handle this combinatorial information may work better on the dataset with a higher rate of missing values and can be expected to perform better than those working on transcriptomics data alone (Liew, Law and Yan, 2011).

### 3 Data normalization

Once the missing values are imputed, the datasets can be subjected to downstream analysis. Efficacy of some of the classification methods, e.g., tree-based techniques, linear discriminant analysis, naïve Bayes, *etc.*, does not get affected by variability in the data. However, the performance of class comparison, class discovery, and classification methods, e.g., KNN, SVM *etc.*, may get affected due to technical variations in gene expression signals. The gene expression signals may vary from sample to sample due to technical reasons such as the efficiency of labeling, amount of RNA, and platform used for the generation of data. It is important to reduce the variability due to technical reasons but preserve the variability due to biological reasons. This can be achieved using data normalization or scaling techniques (Brown et al., 1999) (Table 3).

Quantile normalization (Bolstad et al., 2003; Hansen, Irizarry and Wu, 2012) is a global mean or median technique utilized for the normalization of single channel expression array data. It arranges all the expression values of samples in order, takes average across probes, substitutes probe intensity with average value, and goes back to the original order. Low computational cost is the advantage of quantile normalization. Robust Multi-chip Average (RMA) is a commonly used technique to generate an expression matrix from Affymetrix data (Gautier et al., 2004) or oligonucleotide microarray (Carvalho and Irizarry, 2010). RMA obtains background corrected, quantile normalized gene expression values (Irizarry et al., 2003). Robust Spline Normalization (RSN) used for Illumina data also makes use of quantile normalization (Du, Kibbe and Lin, 2008). Quantile normalization is also used for single color Agilent data (Smyth, 2005). Loess is a local polynomial regression-based approach which can be utilized to adjust intensity levels between two channels (Yang et al., 2002; Smyth and Speed, 2003; Bullard et al., 2010; Baans et al., 2017). Loess normalization performs local regression for each pair of arrays which are composed of the difference and average of the log-transformed intensities derived from the two channels. Two color Agilent data (Smyth, 2005) (Du, Kibbe and Lin, 2008) use loess normalization. Log-transformation is the simplest

and very common data normalization technique applied to gene expression data (Pochet et al., 2004; Li, Suh and Zhang, 2006; Aziz et al., 2017). This method does not shuffle the relative order of expression values, therefore, does not affect the rank-based test results. Log transformation is often applied to data subjected to prior normalization by other methods such as quantile and loess.

Standardization is a normalization technique that does not bind values to a specific range. Standardization is commonly applied by subtracting the mean value from each expression value. Z-score is one of the most frequently used methods of standardization. The Z-score transformation modifies expression values such that the expression value of each gene is denoted as a unit of standard deviation from the normalized mean of zero (Cheadle et al., 2003). The standardization can also be used with the median instead of the mean (Pan, Lin and Le, 2002). The use of the median is more robust against outliers. Standardization techniques are often used for data visualization.

Feature normalization can have positive and negative effects on the expression array analysis results. It lowers the bias but also decreases the sensitivity of the analysis (Freyhult et al., 2010). Existing normalization methods for microarray gene expression data generally assume a similar global expression pattern among samples being studied. However, scenarios of global shifts in gene expressions are dominant in the datasets of complex diseases, for example, cancers which makes the assumption invalid. Therefore, when applying it should be kept in mind that normalization techniques such as RMA or Loess may arbitrarily flatten the differences between sample groups which may lead to biased gene expression estimates.

### 4 Feature selection and feature extraction

High dimension data often results in the sparsity of information which is less reliable for prediction analysis. As a result, feature selection or feature extraction techniques are typically used to find informative genes and resolve the curse of dimensionality. The dimensionality reduction not only speeds up the training process but also helps in data visualization. Dimensionality reduction is achieved by either selection or extraction of features by transforming the original set of features into new ones. Dimensionality reduction serves as an important step in classification and class discovery analysis. For classification, the dataset is split into training and testing sets, and feature selection/extraction is carried out only on the training set to avoid data leakage. Feature selection and extraction techniques are broadly divided into four categories: filter methods, wrapper methods, embedded methods, and hybrid methods (Tyagi and Mishra, 2013; Dhote, Agrawal and Deen, 2015; Almugren and Alshamlan, 2019) (Figure 5) (Table 4).

TABLE 3 List of data transformation and feature scaling techniques prior to dimensionality reduction.

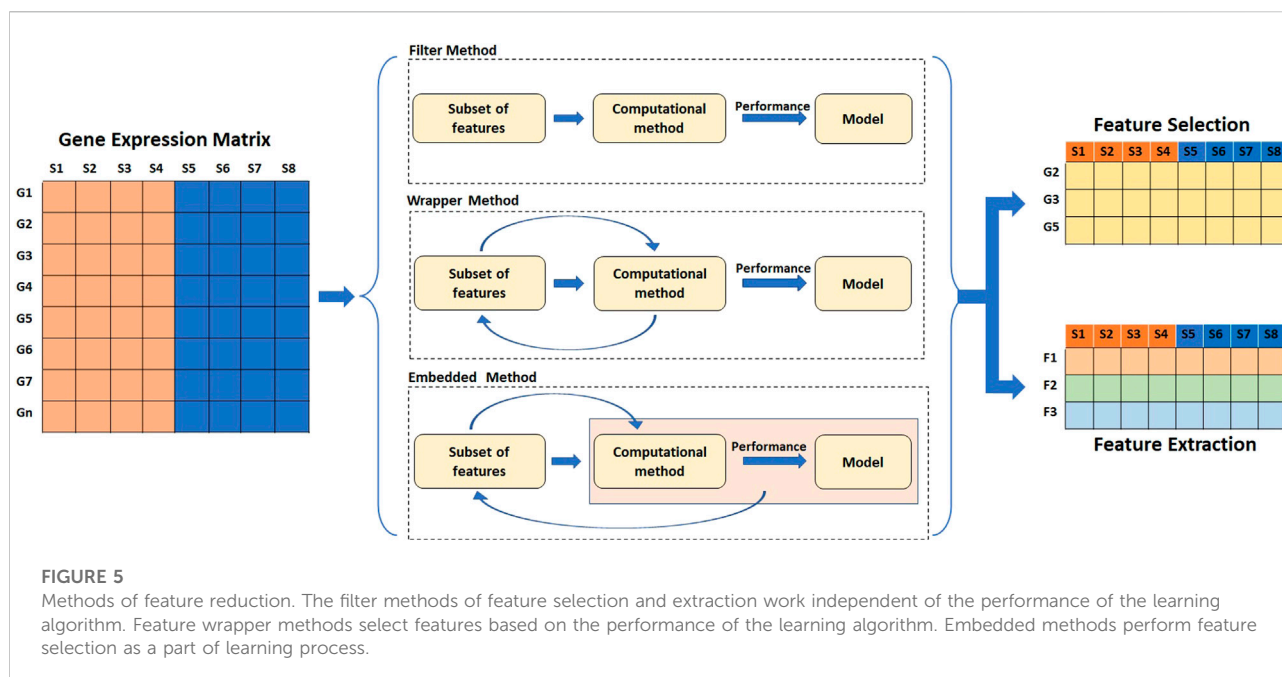
Type	Advantages	Limitation	Technique	Reference
Normalization	Identifies and removes systematic variability. Increases the learning speed.	Less effective if high number of outliers exist in the data.	Quantile	Larsen et al. (2014) Smyth and Speed (2003) Schmidt et al. (2004)
			Loess	Franks et al. (2018) Karthik and Sudha (2021) Larsen et al. (2014) Huang et al. (2018) Bolstad et al. (2003) Doran et al. (2007)
Data transformation	Reduces the variance and reduces the skewness of the distribution of data points.	Data do not always approximate the log-normal distribution.	Log transformation	Pirooznia et al. (2008) Pan et al. (2002) Doran et al. (2007)
Standardization	Ensures feature distributions have mean = 0. Applicable to datasets with many outliers.	Less effective when data distribution is not Gaussian, or the standard deviation is very small.	z-score	Peterson and Coleman (2008) Cheadle et al. (2003) De Guia et al. (2019) Chandrasekhar et al. (2011) Pan et al. (2002)

## 4.1 Filter approaches

The filter methods are independent of the performance of the learning algorithm. Statistical methods such as ANOVA, chi-square, *t*-test, *etc.* (Pan, Lin and Le, 2002; Saeys, Inza and Larrañaga, 2007; Land et al., 2011; Önskog et al., 2011; Kumar et al., 2015) which are often used for class comparison are also used for the feature selection for prediction analysis. The fold change or *p*-value is often used as a cutoff parameter for the selection of features. Correlation-based unsupervised learning algorithms are also used for the features selection process (Figure 6A). In correlation-based features selection (CFS), Pearson's coefficient is utilized to compute the correlation among feature genes (Al-Batah et al., 2019). As a next step, the network of genes that has a moderate to high positive correlation with the output variable is retained. Statistical approaches have also been coupled with correlation analysis for feature selection on Maximum Relevance and Minimum Redundancy (MRMR) principles (Radovic et al., 2017). MRMR is a filter approach that helps to achieve both high accuracy and fast speed (Ding and Peng, 2005; Abdi, Hosseini and Rezghi, 2012). The method selects genes that correlate with the condition but are dissimilar to each other. Another commonly used tool is Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder and Horvath, 2008). This approach is utilized to find the correlation

patterns in gene expression across samples as an absolute value of Pearson's correlation (Langfelder and Horvath, 2008). WGCNA groups genes into clusters or modules depending on their co-expression patterns (Agrahari et al., 2018). The eigenvectors generated through clustering can be thought of as a weighted average expression profile, also called eigengenes. These eigengenes can be used to study the relationship between modules and external sample traits. WGCNA is used more often in class comparison analysis for the identification of "hub" genes associated with a trait of interest. Another correlation-based technique, Fast Correlation Feature Selection (FCFS) utilizes a predominant correlation to identify relevant features and redundancy among them without pairwise correlation analysis (Yu and Liu, 2003) (Figure 6B).

The entropy-based methods are supervised learning methods that are used for feature selection. The entropy-based method selects features such that the probability distribution function across external traits have the highest entropy. Information Gain (IG) is a commonly used entropy-based method for feature selection applied to expression array data (Nikumbh, Ghosh and Jayaraman, 2012; Bolón-Canedo et al., 2014; Ayyad, Saleh and Labib, 2019). IG calculates the entropy of gene expression for the entire dataset. The entropy of gene expression for each external trait is then calculated. Based on entropy values, the information gain is calculated for



each feature. Ranks are assigned to all the features and a threshold is used to select the features genes. The information gained is provided to the modeling algorithm as heuristic knowledge.

Feature extraction methods are multivariate in nature and are capable of extracting information from multiple feature genes. Classical Principal Component Analysis (PCA), an unsupervised linear transformation technique has been used for dimensionality reduction (Jolliffe, 1986; Pochet et al., 2004; Ringnér, 2008; Adiwijaya et al., 2018) (Figure 6C). PCA builds a new set of variables called principal components (PCs) using original features. To obtain principal components, PCA finds linear projection of gene expression levels with maximal variance over a training set. The PCs with the highest eigenvalues which explain the most variance in data are usually selected for further analysis. Independent component analysis (ICA), another unsupervised transformation method, generates a new set of features from the original ones by assuming them to be linear mixtures of latent variables (Lee and Batzoglou, 2003; Zheng, Huang and Shang, 2006). All features generated using ICA are considered to be statistically independent and hence equally important. As a result, unlike PCA, all components from ICA are used for further analysis. (Hyvärinen, 2013), however, as compared to PCA, ICA is slower. Linear Discriminant Analysis (LDA), on the other hand, is a supervised linear transformation feature reduction method that takes class labels into account and maximizes the separation between classes (Guo and Tibshirani, 2007; Sharma et al., 2014) (Figure 6C). The projection vectors are generated from

original features. The projection vectors corresponding to the highest eigenvalue are used for downstream analysis. Similar to PCA, LDA also uses second order statistics. However, as compared to PCA and ICA, LDA offers faster speed and scalability.

All filter approaches (both simple filter and feature extraction methods) ignore the interface with classifier which can result in poor classification performance. This limitation can be overcome by wrapper and embedded approaches.

## 4.2 Wrapper approaches

The wrapper approach is a feature selection approach that wraps a specific machine learning technique applied to fit the data (Figure 7). The wrapper approach overcomes the limitation of the filter approach by selecting a subset of features and evaluating them based on the performance of the learning algorithm. The process of feature selection repeats itself until the best set of features is found.

Sequential Forward Selection (SFS) is an iterative method of feature selection (Figure 7A). It calculates the performance of each feature and starts with the best performing feature. It then keeps adding a feature with each iteration and keeps checking the performance of the model. A set of features that will produce the highest improvement will be retained, and others will be discarded (Park, Yoo and Cho, 2007; Fan, Poh and Zhou, 2009). Sequential Backward Elimination (SBE), on the other hand, initiates the feature selection process by including all the features in the first iteration and by removing one feature with each iteration (Figure 7B). The effect of elimination of each feature is evaluated based on the prediction

TABLE 4 List of different feature selection and feature extraction techniques.

Approach	Advantages	Limitation	Feature Selection Techniques	Reference
Filter	Datasets are easily scalable. Perform simple and fast computation. Independent of the prediction-outcome. Only one-time feature selection.	Ignores the interface with the classifier. Every feature is separately considered. Ignores feature dependencies. Poor classification performance compared to other feature selection techniques.	t-statistics ( <i>t</i> -test)  Chi-square ANOVA CFS FCFS WGCNA  PCA ICA LDA	Pan et al. (2002), Önskog et al. (2011) Dittman et al. (2010) Kumar et al. (2015) Al-Batah et al. (2019) Yu and Liu (2003) Langfelder and Horvath (2008) Pochet et al. (2004) Zheng et al. (2006) Sharma et al. (2014)
Wrapper	Interaction between selected features and learning model taken into account. Considers feature dependencies.	Higher risk of overfitting compared to filter approach. Computationally intensive.	SFS SBE RFE GA ABC ACO PSO	Park et al. (2007) Dhote et al. (2015) Guyon et al. (2002) Ram and Kuila (2019) Li et al. (2016) Alshamlan et al. (2016) Sahu and Mishra (2012)
Embedded	Requires less computation than wrapper methods.	Very specific to learning technique.	k-means clustering  LASSO GLASSO SGLASSO AE RF	Aydadenta and Adiwijaya (2018) Tibshiranit (1996) Meier et al. (2008) Ma et al. (2007) Danaee et al. (2017) Diaz-Uriarte and Alvarez de Andrés (2006)
Hybrid	Combines filter and wrapper methods. Reduces the risk of overfitting. Lower error rate.	Computationally expensive. Can be less accurate: the filter and the wrapper both being used in different steps.	SVM-RFE MIMAGA-Selection Co-ABC	Guyon et al. (2002) Lu et al. (2017) Alshamlan (2018)

performance (Guyon et al., 2002; Dhote, Agrawal and Deen, 2015). Selection or elimination of features in SFS and SBE is based on a scoring function, e.g., *p*-value, *r*-square, or residuals sum of squares of the model to maximize performance. A Genetic Algorithm (GA) is a stochastic and heuristic search technique used to optimize a function based on the concept of evolution in biology (Pan, Zhu and Han, 2003) (Figure 7C). Evolution works on mutation and selection processes. In GA, the Information Index Classification (IIC) value for each gene feature is calculated. The IIC value for the feature gene represents its prediction power. As a first step, top gene features with high IIC values are selected for further processing. The selected feature genes are randomly assigned a binary form (0 or 1) to

represent a ‘chromosome’. A set of chromosomes of the select genes with randomly assigned 0s and 1s creates a ‘chromosome population’. The fitness power of each chromosome is calculated by considering only the genes which are assigned a value of 1. ‘Fit’ chromosomes are selected using techniques such as Roulette-wheel selection, rank selection, tournament selection, etc. The select set of chromosomes is subjected to crossover or mutagenesis to generate the offspring. Upon crossover and mutagenesis, the chromosomes exchange or mutate their information contents. The offspring chromosomes are used for further downstream analysis (Aboudi and Benhlila, 2016; Sayed et al., 2019). There are quite a few variants of GAs to handle the feature selection problem (Liu,



2008, 2009; Ram and Kuila, 2019; Sayed et al., 2019). Other stochastic and heuristic methods are Artificial Bee Colony (ABC) (Li, Li and Yin, 2016), Ant Colony Optimization (ACO) (Alshamlan, Badr and Alohal, 2016), Particle Swarm Optimization (PSO) (Sahu and Mishra, 2012), etc.

Though, the wrapped methods provide optimized prediction results as compared to the filter methods they are computationally expensive. This limitation of wrapped methods is addressed by the embedded methods.

### 4.3 Embedded approaches

The embedded approaches perform feature selection as a part of the learning process and are typically specific to the learning algorithm. They integrate the importance of both wrapper and filter methods by including feature interaction at a low computational cost. The embedded approach extracts the most contributing features from iterations of training. Commonly used embedded techniques for feature selection are LASSO (Least Absolute Shrinkage and Selection Operator) and Ridge regression (Figure 8A). Both these techniques are regularized versions of multiple linear regression and can be utilized for feature selection (Tibshirani, 1996). These techniques perform feature selection by eliminating weights of the least important features (Hoffmann, 2007; Ma, Song and Huang, 2007; Meier, Van De Geer and Bühlmann, 2008; Algamal and Lee, 2015). Other than LASSO and Ridge Regression, K-means clustering, Random Forest and ANN-based techniques are also used.

The K-means clustering technique is an unsupervised method that is utilized to eliminate redundancy in high-dimensional gene expression data (Aydadenta and Adiwijaya, 2018) (Figure 8B). In K-means clustering, an arbitrary K number of points from the data are selected as centroids, and all the genes are allocated to the nearest centroid (MacQueen, 1967; Kanungo et al., 2002). After clustering, a scoring algorithm such as Relief (Kira and Rendell, 1992) is utilized and high-scoring gene features of each cluster are selected for further analysis. The computational complexity of K-means is linear with respect to the number of instances, clusters, and dimensions. Though it is one of the fastest clustering techniques, it may also lead to an incorrect result due to convergence to a local minimum. The Random Forest (RF) is a supervised approach applied to obtain very small sets of non-redundant genes by preserving predictive accuracy (Díaz-Uriarte and Alvarez de Andrés, 2006; Moorthy and Mohamad, 2012) (Figure 8C). RF is an ensemble of decision trees constructed by randomly selecting data samples from the original data (Breiman, 2001). The final classification is obtained by combining results from the decision trees passed by vote. The bagging strategy of RF can effectively decrease the risk of overfitting when applied to large dimension data. RF can also incorporate connections among predictor features. The

prediction performance of RF is highly competitive when compared with SVM and KNN. An important limitation of RF is that many trees can make the model very slow and unproductive for real-time predictions.

ANN-based Autoencoders (AE) (Kramer, 1991) is an unsupervised encoder and decoder technique (Figure 8D). It tries to obtain output layer neuron values as close as possible to input layer neurons using lower-dimensional layers in between. AE can obtain both linear and nonlinear relationships from the input information. AE such as Denoising Autoencoders (DAE) (Vincent and Larochelle, 2008), Stacked Denoising Autoencoder (SDAE) (Vincent et al., 2010; Danaee, Ghaeini and Hendrix, 2017) are utilized to extract functional features from expression arrays and are capable of learning from the dense network. Convolutional Neural Network (CNN) is another ANN-based architecture that is utilized for the feature extraction process in order to improve classification accuracy (Zeebaree, Haron and Abdulazeez, 2018; Almugren and Alshamlan, 2019) (Figure 8E). CNN can extricate local features from the data (LeCun et al., 1998; O'Shea and Nash, 2015). The convolutional layer of CNN extracts the high-level features from the input values. The pooling layer is utilized to reduce the dimensionality of feature maps from the convolution layer.

### 4.4 Hybrid approaches

A hybrid approach is considered as a combination of two or more filter and wrapper methods. It can reduce the error rate and the risk of overfitting. A well-known feature selection hybrid approach is Recursive Feature Elimination with a linear SVM (SVM-RFE) (Guyon et al., 2002). SVM-RFE utilizes SVMs classification capability and, from the ranked list, recursively deletes the least significant features. This method was taken as a benchmark feature selection method due to its performance. However, its main disadvantage is that it ignores the correlation hidden between the features and requires high computational time (Li, Xie and Liu, 2018). A combination of the mutual information maximization (MIM) and the adaptive genetic algorithm (AGA) has also been proposed for feature selection (Lu et al., 2017). MIM is able to select the advanced feature subset, and AGA speeds up the search in the identification of the substantial feature subsets. This combination of methods is more efficient and robust compared to the individual component (Lu et al., 2017). This technique streamlines the feature selection procedure without getting into classification accuracy on the reduced dataset. MIMAGA-Selection technique can reduce datasets with the number of genes up to 20,000 to below 300 with high classification accuracies. It also removes redundancy from the data and results in a lower error rate (Bolón-Canedo et al., 2014). This technique is an iterative feature reduction technique. Therefore, with an increase in the size of the microarray dataset, the computational time increases.

Co-ABC is a hybrid approach for feature selection based on the correlation Artificial Bee Colony (ABC) algorithm (Alshamlan, 2018). The first step utilizes correlation-based feature selection to filter noisy and redundant genes from high dimensionality domains and the second step utilizes ABC technique to select the most significant genes.

Feature selection or feature extraction process can generate high quality data for classification and predication analysis. It should be noted that for classification analysis, feature selection is carried out only on the training dataset. For clinical applications, it should be noted that model interpretation is important, and feature extraction technique may cause the model interpretation challenging as compared to feature selection techniques.

## 5 Modeling/learning and analysis

The final step of analysis of microarray gene expression data is statistical analysis and model learning through computational techniques. Methods used for normalization, gene selection and analyses exhibit a synergistic relationship (Önskog et al., 2011). Class Comparison is one of the most common types of gene expression data analysis for the identification of differentially expressed genes (O'Connell, 2003). To solve the class comparison problems most researchers use standard statistical techniques e.g., *t*-test, ANOVA, etc. (Storey and Tibshirani, 2003). Scoring enrichment techniques such as z-score or odds ratio are hit-counting methods utilized to describe either the pathway or the functional enrichment of a gene list (Curtis, Orešič and Vidal-Puig, 2005). A higher number of hits shows a higher score and represents greater enrichment.

### 5.1 Classification (class prediction)

Classification is the process of classifying microarray data into categories or systematic arrangement of microarray data into different classes, e.g., cases and controls. For classification analysis, the entire dataset is divided into two subsets, viz. training and testing. The training dataset, which typically comprises 70–80% of the samples, is used for the construction of a model. To improve the efficiency of classification, it is essential to assess the performance of models. A common way to improve the performance of a model during training is to include an additional validation subset (Refaeilzadeh, Tang and Liu, 2009). The validation dataset comprises 10–15% of the total sample observations used for parameter optimization. The remaining samples are used as a testing dataset. (Refaeilzadeh, Tang and Liu, 2009). However, to assess the generalization ability and prevent model overfitting, instead of setting aside a single validation set, k-fold cross-validation can be an effective solution. Various ML algorithms have been used for classification analysis.

K-Nearest Neighbor (KNN) is one of the techniques that can be utilized for the classification of expression array data (Kumar et al., 2015; Ayyad, Saleh and Labib, 2019). The classification of a sample is achieved by measuring its distance (e.g., Euclidean distance etc.) from all training samples using the distance metric. The performance of KNN is dependent on the threshold of the feature selection method and is subject to the distance function (Deegalla and Bostr, 2007). An increase in sample size has been shown to increase the computational and time complexity of KNN (Begum, Chakraborty and Sarkar, 2015). Another classification technique for expression array data is Nearest Shrunken Centroid (NSC) (Tibshirani et al., 2003; Dallora et al., 2017). It calculates the centroid for each class and tries to shrink each of the class centroids toward the global centroid by threshold. A sample is classified into a class whose centroid is nearest to it based on the distance metric. This method can reduce the effects of noisy genes. However, an arbitrary choice of shrinkage threshold is a limitation of NSC.

A Decision Tree (DT) (Safavian and Landgrebe, 1991) approach can also be utilized for the classification of gene expression data (Peng, Li and Liu, 2006; Krętowski and Grześ, 2007; Chen et al., 2014). A decision tree is also a versatile ML technique that can perform classification as well as regression operations (Safavian and Landgrebe, 1991). DT requires less effort for data preparation during preprocessing. However, a slight variation in the input information can result in a significant variation in the optimal decision tree structure. Also, overfitting is a known limitation of the DT models. Random Forest (RF) (Breiman, 2001) is another algorithm used for the classification and regression analysis of gene expression data. RF is an ensemble of decision trees (Statnikov, Wang and Aliferis, 2008; Aydadenta and Adiwijaya, 2018). While Random Forest has lesser chances of overfitting and provides more accurate results, it is computationally expensive and more difficult to interpret as compared to DT.

Another technique that is utilized for classification analysis using expression arrays is an SVM (Brown et al., 2000; Furey et al., 2000; Ben-Hur, 2001; Abdi, Hosseini and Rezghi, 2012; Adiwijaya et al., 2018; Turgut, Dagtekin and Ensari, 2018). For complex non-linear data, higher degree polynomials can be added to the cost function of SVM. This will increase the combination of a number of features; however, this results in the reduction of computation speed. To overcome this situation, 'kernel trick' is used, which can handle complex non-linear data without the addition of any polynomial features. Various kernel types can be used with SVM, such as linear, polynomial, radial, etc. In some studies, SVMs performed better than DT and ANN-based techniques (Önskog et al., 2011), whereas, in others the performance of SVM was poor (Tabares-Soto et al., 2020) (Motieghader et al., 2017).

Multilayered CNN, a deep learning algorithm typically applied where the data can be visualized as an image (Neubauer, 1998; Collobert and Weston, 2008), has also been proposed for the analysis of microarray data (Zeebaree, Haron and Abdulazeez, 2018). Each neuron is scanned

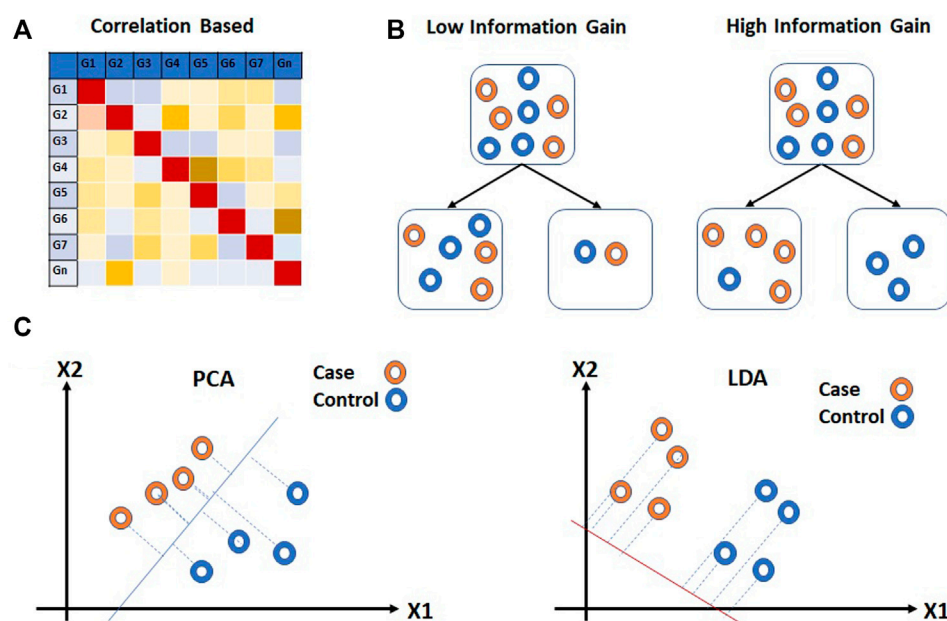


FIGURE 6

Filter approaches for feature reduction. (A) Correlation based feature selection (CFS) and (B) Information Gain (IG) are feature selection approaches for feature reduction. (C) Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) extract information from multiple feature genes. Hollow circles in (B,C) represent samples.

throughout the input matrix, and for every input, the CNN calculates the locally weighted sum and produces an output value. CNN can deal with insufficient data. CNN involves much less preprocessing and can do far better in terms of results as compared to other supervised techniques.

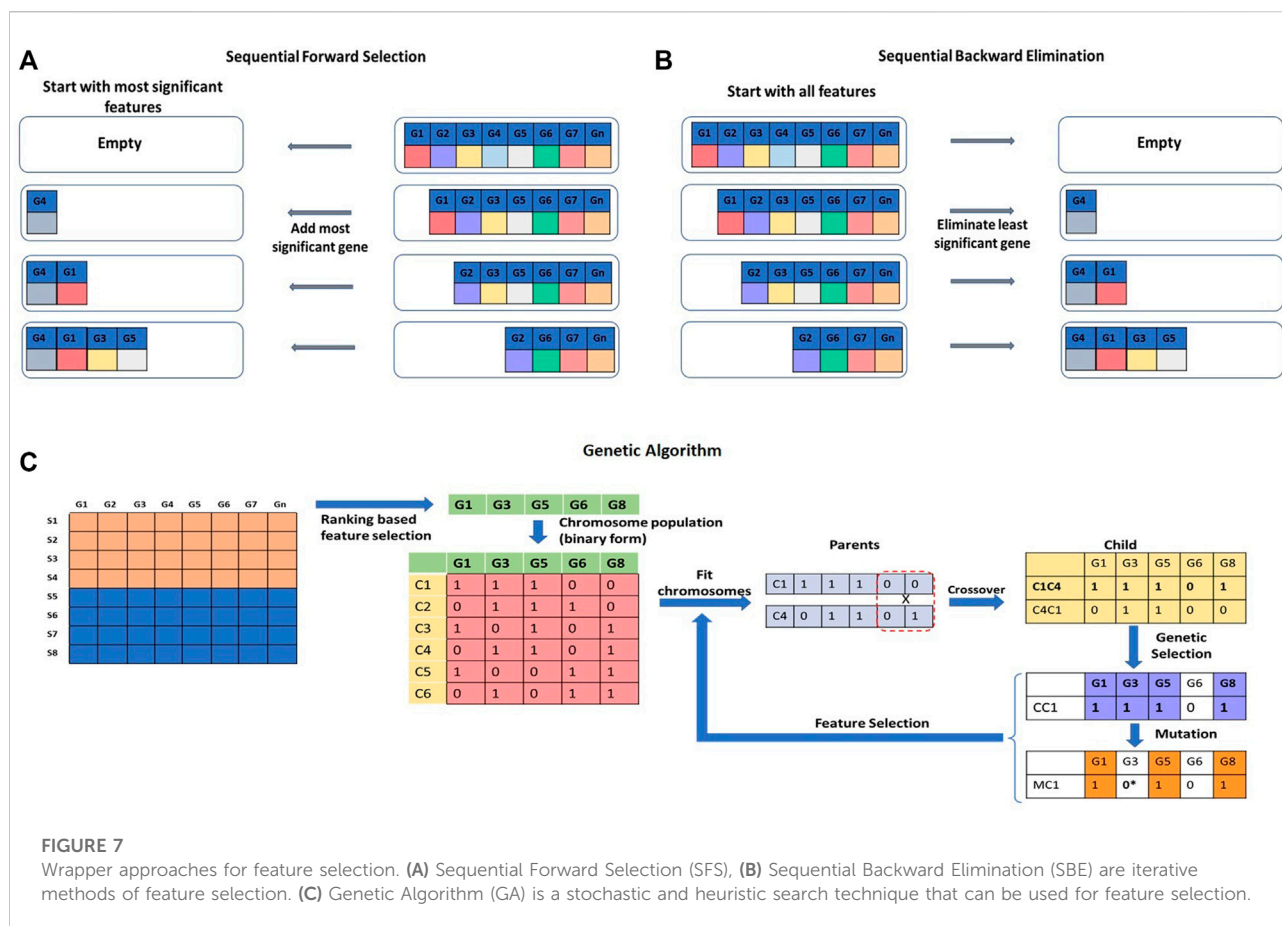
The performance evaluation for classification analysis using classification techniques can be achieved by error rate or accuracy parameters. Root Mean Squared Error (RMSE) or Root Relative Squared Error (RRSE) are examples of error-rate-based evaluation. The accuracy metric is the most common performance evaluation parameter utilized to find the accuracy of classification. However, accuracy alone is not enough for performance evaluation (McNee, Riedl and Konstan, 2006; Sturm, 2013) and therefore, a confusion matrix is computed. A set of predictions is compared with actual targets to compute the confusion matrix. The confusion matrix represents true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP, TN, FP and FN are utilized to calculate more concise metrics such as precision, recall (sensitivity), specificity, Matthew's correlation coefficient (MCC), etc. ROC (Receiver Operating Characteristic) curve and Precision-Recall curve are other standard tools used by binary classifiers as performance measures. ROC and MCC are more robust measures as compared to accuracy since accuracy is affected by class imbalance (Chicco and Jurman, 2020).

The problem of classification of expression data is both biologically important and computationally challenging. From a

computational perspective one of the major challenges in analyzing microarray gene expression data is a small sample size. Error estimation is greatly affected by the small sample size, and the possibility of overfitting of data is very high (Hambali, Oladele and Adewole, 2020). Another important issue in gene expression array data analysis is class imbalance for the classification tasks. In clinical research on rare diseases, generally, the number of case samples is very less as compared to healthy controls which may lead to biased results. With decreasing costs of microarray profiling and high-throughput sequencing, this challenge can be expected to be resolved in the near future.

## 5.2 Class discovery

The third type of microarray analysis is class discovery which involves the analysis of a set of gene expression profiles for the discovery of novel gene regulatory networks or sample types. Hierarchical Clustering Analysis (HCA) is a simple process of sorting instances into groups of similar features and is very commonly used for the analysis of expression array data (Eisen et al., 1998). Hierarchical clustering produces a dendrogram which is a binary tree structure and represents the distance relationships between clusters. HCA is a highly structured approach and the most widely used technique for expression analysis (Bouguettaya et al., 2015). However, the graphical representation of hierarchy is very



complex in HCA. The lack of robustness and inversion problems complicate the interpretation of the hierarchy. HCA is also sensitive to small data variations. Self-Organizing Maps (SOM) is another clustering technique used for the identification of prevalent gene expression patterns and simple visualization of specific genes or pathways (Tamayo et al., 1999). SOM can perform non-linear mapping of data with a two-dimensional map grid. Unlike HCA, SOM is less sensitive to small data variations (Nikkila et al., 2002).

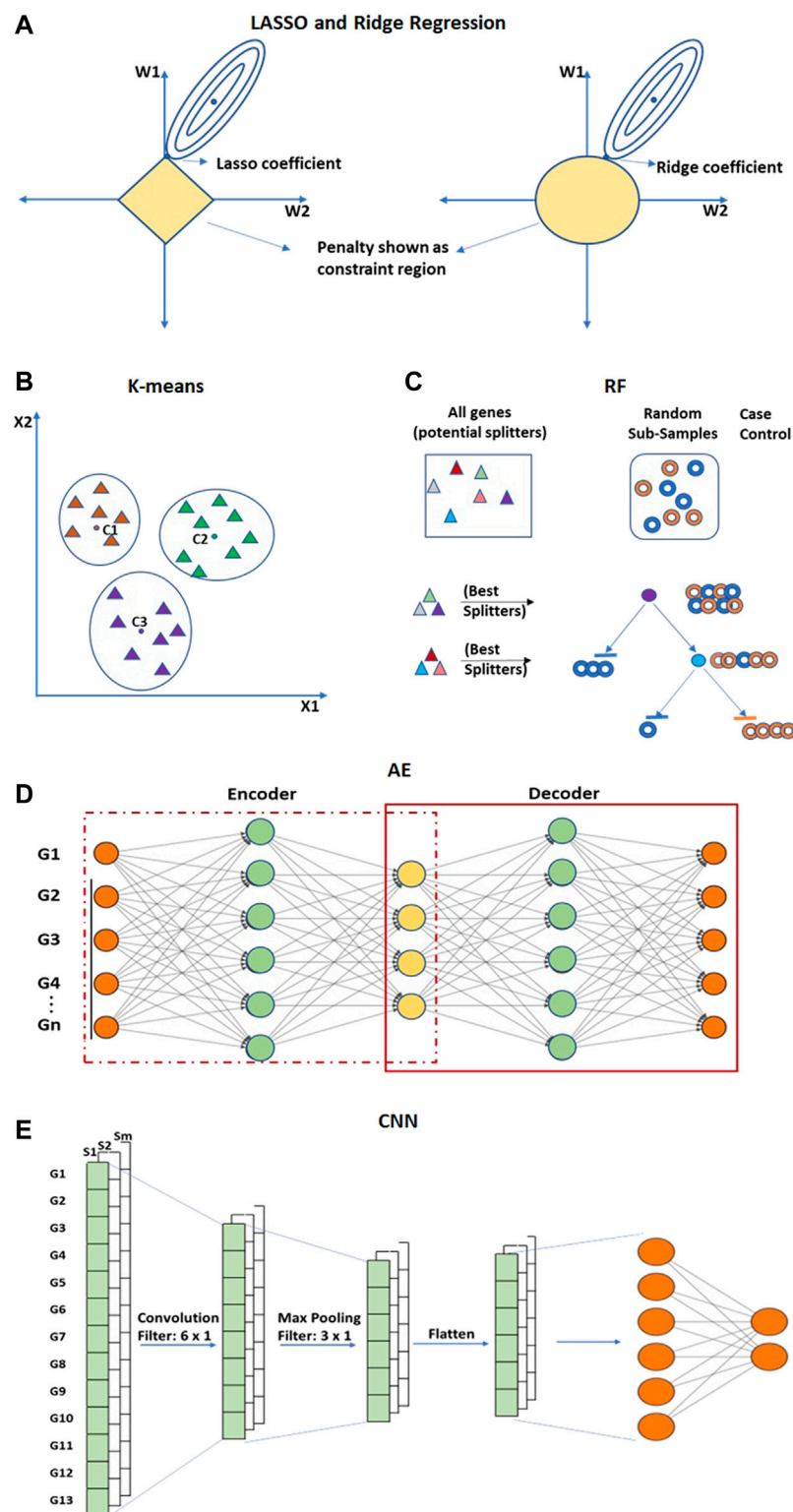
K-means is an iterative technique that minimizes the overall within-cluster dispersion. K-means algorithm has been utilized to discover transcriptional regulatory sub-networks of yeast without any prior assumptions of their structure (Tavazoie et al., 1999). The advantage of K-means over other clustering techniques is that it can deal with entirely unstructured input data (Gentleman and Carey, 2008). However, the K-means technique easily gets caught with the local optimum if the initial center points are selected randomly. Therefore various modified versions of K-means are applied for converging to the global optimum (Lu et al., 2004; Nidheesh, Abdul Nazeer and Ameer, 2017; Jothi, Mohanty and Ojha, 2019).

Another technique for class discovery analysis is the Bayesian probabilistic framework which uses Bayes

theorem (Friedman et al., 2000; Baldi and Long, 2001). This technique is a good fit for small sample sizes of microarray studies; however, it is computationally exhaustive for a dataset with a very high number of samples and features. Nonnegative Matrix Factorization (NMF) is also a clustering technique utilized for pattern analysis of gene expression data (Kim and Tidor, 2003; Brunet et al., 2004). NMF involves factorization into matrices with nonnegative entries and recognizes the similarity between sub-portions of the data corresponding to localized features in expression space (Kim and Park, 2007; Devarajan and Ebrahimi, 2008).

Evaluation measures for clustering algorithms utilized for class discovery can be of three different types, viz. internal validation index, relative validation index, and external validation index (Dalton, Ballarin and Brun, 2009). The internal validation index method calculates properties of the resulting clusters based on internal properties of clusters such as compactness, separation, and roundness. Dunn's Index and Silhouette Index are examples of internal validation indices. The relative validation indexing method compares clusters generated by algorithms with different parameters or subsets of the data. It can measure the stability



**FIGURE 8**

Embedded approaches performs feature selection and extraction. (A) LASSO and Ridge are regularized versions of multiple linear regression used for feature selection. (B) K-means clustering is an unsupervised method for dimensionality reduction that selects feature genes allocated to the nearest centroid. (C) Random Forest (RF) is an ensemble of decision trees. (D) Convolutional Neural network (CNN) and (E) Autoencoders (AE) are deep learning-based methods of feature reduction. Hollow circles in (C) represent samples, and solid triangles in (B,C) represent genes.



TABLE 5 Evaluation Parameters for analysis of microarray gene expression data.

Evaluation metric	Specifics	References
<b>Prediction performance evaluation parameters</b>		
Root Mean Squared Error (RMSE)	RMSE is a square root of mean of the difference between predicted values and actual values for each sample	Vihinen, 2012, Parikh et al., 2008a, Parikh et al., 2008b, Goffinet and Wallach, 1989
Root Relative Squared Error (RRSE)	RRSE is a normalized RMSE which enables the comparison between datasets or models with different scales. Standard deviation is used for normalization	
Accuracy	The accuracy of a test is its ability to differentiate the cases and controls correctly	
Precision/Positive Prediction Value	The Precision of a test is its ability to determine cases that are true cases	
Sensitivity/Recall/True Positive Rate	The sensitivity of a test is its ability to determine the cases (positive for disease) correctly	
Specificity/True negative Rate	The specificity of a test is its ability to determine the healthy cases correctly	
F1-score	F1-score of a test is its ability to determine harmonic mean of precision and recall	
MCC	MCC of a test is a correlation coefficient between the true and predicted values	Chicco and Jurman, 2020, Matthews, 1975
ROC curve	ROC curve is a graph where each point on a curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. Area Under the ROC curve is a measure of how well a parameter can distinguish between cases and controls. ROC curves should be used when there are roughly equal numbers of instances for each class	Fawcett, 2006, Davis and Goadrich, 2006
Precision-Recall Curve	A precision-recall (PR) curve is a graph where each point on a curve represents a precision/sensitivity pair corresponding to a particular threshold. PR curves should be used when there is moderate to high class imbalance	Buckland and Gey, (1994)
<b>Clustering performance evaluation parameters</b>		
Dunn's Index	Dunn's index is a ratio between the minimum distance between two clusters and the size of largest cluster. Larger the index better the clustering	Dunn, 1974, Dalton, Ballarin and Brun, 2009
Silhouette Index	Silhouette Index of a cluster is defined as the average Silhouette width of its points. Silhouette width of a given point defines its proximity to its own cluster relative to its proximity to other clusters	Rousseeuw, 1987, Dalton, Ballarin and Brun, 2009
Figure of Merit Index	The FOM of a feature gene is computed by clustering the samples after removing that feature and by measuring the average distance between all samples and their cluster's centroids. The FOM for a clustering technique is the sum of FOM over each feature gene at a time	Smith and Snyder, 1979, Dalton, Ballarin and Brun, 2009
Instability Index	Instability index is disagreement between labels obtained over data points to parts of a dataset, averaged over repeated random partitions of the data points. Clustering method is applied to a part of dataset, and the labels obtained on that part of the dataset are utilized to train a classifier that partitions the whole space	Guruprasad, Reddy and Pandit, 1990, Dalton, Ballarin and Brun, 2009
Hubert's Correlation, Rand Statistics, Jaccard Coefficient, Folke's and Mallow's index	All these measures analyse the relationship between pairs of points using the co-occurrence matrices for the expected partition and the one generated by the clustering algorithm	Dalton, Ballarin and Brun, 2009, Brun et al., 2007

of the technique against variations in the data, or consistency of the results in the case of redundancy. The figure of merit index and instability index are examples of relative validation indices. External validation index method compares the groups generated by the clustering technique to the actual cluster of the data. Generally, external methods are considered to be better correlated to

the actual error as compared to internal and relative indexing methods. Hubert's Correlation, Rand Statistics, Jaccard Coefficient, and Folke's and Mallow's index are a few examples of external evaluation parameters. Table 5 describes all the evaluation parameters discussed above.

While dealing with a very large number of gene features in expression arrays, multiple gene feature selection

techniques are available to deal with dimensionality problem. However, an elaborate study is required to identify optimum methods for downstream analysis that can be combined with specific dimensionality reduction techniques.

## 6 Conclusion and future directions

In this paper, we have attempted to describe the complete pipeline for the analysis of expression arrays. Conventional ML methods for missing value imputation, dimensionality reduction, and classification analysis have achieved success. However, with an increase in data complexity, deep learning techniques may find increasing usage. The current applications of genomics in clinical research may benefit from the data coming from different modalities. For gene expression data analysis of complex diseases, data sparsity or class imbalance is a real concern. This issue can be addressed with the recent technology of data augmentation, for example, Generative Adversarial Networks (GANs) (Chaudhari, Agrawal and Kotecha, 2020). The aim of any class prediction algorithm for diagnostic applications in a clinical research is not only to predict but also to disclose the reasons behind the predictions made. This understanding of the undercover mechanism with some evidence makes the model interpretable. Therefore, it is important to develop interpretable models which help to understand the problem and the situation where the model may fail (Holzinger et al., 2017). Interpretation models such as perturbation-based, derivative-based, local and global surrogate-based should get attention to solve these problems (Ribeiro, Singh and Guestrin, 2016; Zou et al., 2019).

## References

- Abapihi, B., Mukhsar, Adhi Wibawa, G. N., Baharuddin, Lumbanraja, F. R., Faisal, M. R., et al. (2021). Parameter estimation for high dimensional classification model on colon cancer microarray dataset. *J. Phys. Conf. Ser.* 1899 (1), 012113. doi:10.1088/1742-6596/1899/1/012113
- Abberton, M., Batley, J., Bentley, A., Bryant, J., Cai, H., Cockram, J., et al. (2016). Global agricultural intensification during climate change: A role for genomics. *Plant Biotechnol. J.* 14 (4), 1095–1098. doi:10.1111/pbi.12467
- Abdi, M. J., Hosseini, S. M., and Rezghi, M. (2012). A novel weighted support vector machine based on particle swarm optimization for gene selection and tumor classification. *Comput. Math. Methods Med.*, 320698. doi:10.1155/2012/320698
- Aboudi, N. El, and Benhlila, L. (2016). “Review on wrapper feature selection approaches,” in Proceedings - 2016 International Conference on Engineering and MIS, ICMIS 2016 (IEEE). doi:10.1109/ICMIS.2016.7745366
- Adiwijaya, A., Wisesty, U., Kusumo, D., and Aditsania, A. (2018). Dimensionality reduction using Principal Component Analysis for cancer detection based on microarray data classification. *J. Comput. Sci.* 14 (11), 1521–1530. doi:10.3844/jcsp.2018.1521.1530
- Aghdam, R., Baghfalaki, T., Khosravi, P., and Saberi Ansari, E. (2017). The ability of different imputation methods to preserve the significant genes and pathways in cancer. *Genomics Proteomics Bioinforma.* 15 (6), 396–404. doi:10.1016/j.gpb.2017.08.003
- Agrahari, R., Foroushani, A., Docking, T. R., Chang, L., Duns, G., Hudoba, M., et al. (2018). *Applications of Bayesian network models in predicting types of hematological malignancies*. Scientific Reports. United States: Springer 8 (1), 1–12. doi:10.1038/s41598-018-24758-5
- Aittokallio, T. (2009). Dealing with missing values in large-scale studies: Microarray data imputation and beyond. *Brief. Bioinform.* 11 (2), 253–264. doi:10.1093/bib/bbp059
- Al-Batah, M., Zaqaibeh, B. M., Alomari, S. A., and Alzboon, M. S. (2019). Gene Microarray Cancer classification using correlation based feature selection algorithm and rules classifiers. *Int. J. Onl. Eng.* 15 (8), 62–73. doi:10.3991/ijoe.v15i08.10617
- Algarni, Z. Y., and Lee, M. H. (2015). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Syst. Appl.* 42 (23), 9326–9332. doi:10.1016/j.eswa.2015.08.016
- Alloul, A., Spanoghe, J., Machado, D., and Vlaeminck, S. E. (2022). Unlocking the genomic potential of aerobes and phototrophs for the production of nutritious and palatable microbial food without arable land or fossil fuels. *Microb. Biotechnol.* 15 (1), 6–12. doi:10.1111/1751-7915.13747
- Almugren, N., and Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access* 7, 78533–78548. doi:10.1109/ACCESS.2019.2922987
- Alshamlan, H. M., Badr, G. H., and Alohal, Y. A. (2016). ABC-SVM: Artificial bee colony and SVM method for microarray gene selection and Multi class cancer classification. *Int. J. Mach. Learn. Comput.* 6 (3), 184–190. doi:10.18178/ijmlc.2016.6.3.596
- Alshamlan, H. M. (2018). Co-ABC: Correlation artificial bee colony algorithm for biomarker gene discovery using gene expression profile. *Saudi J. Biol. Sci.* 25 (5), 895–903. doi:10.1016/j.sjbs.2017.12.012

## Author contributions

NB and SK wrote the manuscript. SK, RW, and KK outlined the manuscript. RW and KK reviewed the manuscript and inspired the overall work.

## Funding

This work has been supported by the Scheme for Promotion of Academic and Research Collaboration (SPARC) 2018–19, MHRD (project no. P104). NB was supported by the Junior Research Fellowship Award 2018 by Symbiosis International Deemed University, India. Satyajeet Khare is also a beneficiary of a DST SERB SRG grant (SRG/2020/001414).

## 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.

- Arbitrio, M., Scionti, F., Di Martino, M. T., Caracciolo, D., Pensabene, L., Tassone, P., et al. (2021). Pharmacogenomics biomarker discovery and validation for translation in clinical practice. *Clin. Transl. Sci.* 14 (1), 113–119. doi:10.1111/cts.12869
- Ayadadenta, H., and Adiwijaya (2018). A clustering approach for feature selection in microarray data classification using random forest. *J. Inf. Process. Syst.* 14 (5), 1167–1175. doi:10.3745/JIPS.04.0087
- Aydilek, I. B., and Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Inf. Sci.* 233, 25–35. doi:10.1016/j.ins.2013.01.021
- Ayyad, S. M., Saleh, A. I., and Labib, L. M. (2019). Gene expression cancer classification using modified K-Nearest Neighbors technique. *Biosystems.* 176 (12), 41–51. doi:10.1016/j.biosystems.2018.12.009
- Aziz, R., Verma, C., Jha, M., and Srivastava, N. (2017). Artificial neural network classification of microarray data using new hybrid gene selection method. *Int. J. Data Min. Bioinform.* 17 (1), 42. doi:10.1504/ijdm.2017.084026
- Baans, O. S., Hashim, U., and Yusof, N. (2017). Performance comparison of image normalisation method for DNA microarray data. *Pertanika J. Sci. Technol.* 25 (S), 59–68.
- Baldi, P., and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17 (6), 509–519. doi:10.1093/bioinformatics/17.6.509
- Baltes, N. J., and Voytas, D. F. (2015). Enabling plant synthetic biology through genome engineering. *Trends Biotechnol.* 33 (2), 120–131. doi:10.1016/j.tibtech.2014.11.008
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.* 41 (1), 991–995. doi:10.1093/nar/gks1193
- Batista, G. E., and Monard, M. C. (2002). *A study of k-nearest neighbour as an imputation method*, 1–12.
- Begum, S., Chakraborty, D., and Sarkar, R. (2015). “Data classification using feature selection and kNN machine learning approach,” in 2015 International Conference on Computational Intelligence and Communication Networks (CICN) (IEEE), 6–9. doi:10.1109/CICN.2015.165
- Behzadi, P., Behzadi, E., and Ranjbar, R. (2014). The application of microarray in medicine. *ORL* 24, 36–38.
- Ben Hur, A. (2001). Support vector clustering. *J. Mach. Learn. Res.* 2, 125–137.
- Bengio, Y., and Gingras, F. (1995). Recurrent neural networks for missing or asynchronous data. *Adv. neural Inf. Process. Syst.* 8.
- Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi:10.1038/nature07517
- Bhandari, N., Khare, S., Walambe, R., and Kotecha, K. (2021). Comparison of machine learning and deep learning techniques in promoter prediction across diverse species. *PeerJ. Comput. Sci.* 7, 3655–e417. doi:10.7717/peerj-cs.365
- Blanchard, A. P., Kaiser, R. J., and Hood, L. E. (1996). High-density oligonucleotide arrays. *Biosens. Bioelectron.* 11 (6/7), 687–690. doi:10.1016/0956-5663(96)83302-1
- Bo, T. H., Dysvik, B., and Jonassen, I. (2004). LSImpute: Accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res.* 32 (3), e34–e38. doi:10.1093/nar/gnh026
- Bolón-Canedo, V., Sanchez-Marono, N., Alonso-Betanzos, A., Benitez, J., and Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Inf. Sci.* 282, 111–135. doi:10.1016/j.ins.2014.05.042
- Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19 (2), 185–193. doi:10.1093/bioinformatics/19.2.185
- Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., and Song, A. (2015). Efficient agglomerative hierarchical clustering. *Expert Syst. Appl.* 42 (5), 2785–2797. doi:10.1016/j.eswa.2014.09.054
- Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., et al. (2003). ArrayExpress - a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 31 (1), 68–71. doi:10.1093/nar/gkg091
- Breiman, L., and So, K. (2001). Random forests. *Mach. Learn.* 45 (1), 117–127. doi:10.1007/978-3-662-56776-0\_10
- Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., et al. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U. S. A.* 97 (1), 262–267. doi:10.1073/pnas.97.1.262
- Brown, M. P. S., Slonim, D., and Zhu, Q. (1999). *Support vector machine classification of microarray gene expression data*. Santa Cruz: University of California, 25–28. Technical Report UCSC-CRL-99-09.
- Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., et al. (2007). Model-based evaluation of clustering validation measures. *Pattern Recognit. DAGM.* 40, 807–824. doi:10.1016/j.patcog.2006.06.026
- Brunet, J. P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.* 101 (12), 4164–4169. doi:10.1073/pnas.0308531101
- Buckland, M., and Gey, F. (1994). The relationship between recall and precision. *J. Am. Soc. Inf. Sci.* 45 (1), 12–19. doi:10.1002/(sici)1097-4571(199401)45:1<12:aid-asi2>3.0.co;2-I
- Bullard, J. H., Purdom, E., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinforma.* 11 (94), 1–13. doi:10.1186/1471-2105-11-94
- Carvalho, B. S., and Irizarry, R. A. (2010). “A framework for oligonucleotide microarray preprocessing,” 2363–2367. doi:10.1093/bioinformatics/btq431
- Chandrasekhar, T., Thangave, K., and Sathishkumar, E. N. (2013). “Unsupervised gene expression data using enhanced clustering method,” in 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology, ICE-CCN 2013 (IEEE), 518–522. doi:10.1109/ICE-CCN.2013.6528554
- Chandrasekhar, T., Thangavel, K., and Elayaraja, E. (2011). Effective clustering algorithms for gene expression data. *Int. J. Comput. Appl.* 32 (4), 25–29.
- Chaudhari, P., Agrawal, H., and Kotecha, K. (2020). Data augmentation using MG-GAN for improved cancer classification on gene expression data. *Soft Comput.* 24 (15), 11381–11391. doi:10.1007/s00500-019-04602-2
- Cheadle, C., Vawter, M. P., Freed, W. J., and Becker, K. G. (2003). Analysis of microarray data using Z score transformation. *J. Mol. Diagn.* 5 (2), 73–81. doi:10.1016/S1525-1578(10)60455-2
- Chen, J. J., Wang, S. J., Tsai, C. A., and Lin, C. J. (2007). Selection of differentially expressed genes in microarray data analysis. *Pharmacogenomics J.* 7, 212–220. doi:10.1038/sj.tpj.6500412
- Chen, K. H., Wang, K. J., Tsai, M. L., Wang, K. M., Adrian, A. M., Cheng, W. C., et al. (2014). Gene selection for cancer identification: A decision tree model empowered by particle swarm optimization algorithm. *BMC Bioinforma.* 15 (1), 49–9. doi:10.1186/1471-2105-15-49
- Chen, Y., Li, Y., Narayan, R., Subramanian, A., and Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics* 32 (12), 1832–1839. doi:10.1093/bioinformatics/btw074
- Chen, Z., Dodig-Crnkovic, T., Schwenk, J. M., and Tao, S. C. (2018). Current applications of antibody microarrays. *Clinical Proteomics. Clin. Proteomics* 15 (1), 7–15. doi:10.1186/s12014-018-9184-2
- Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21 (1), 6–13. doi:10.1186/s12864-019-6413-7
- Collobert, R., and Weston, J. (2008). ‘A unified architecture for natural language processing: Deep neural networks with multitask learning’, in Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 160–167.
- Curtis, R. K., Orešić, M., and Vidal-Puig, A. (2005). Pathways to the analysis of microarray data. *Trends Biotechnol.* 23 (8), 429–435. doi:10.1016/j.tibtech.2005.05.011
- Dallora, A. L., Eivazzadeh, S., Mendes, E., Berglund, J., and Anderberg, P. (2017). Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review. *PLoS ONE* 12 (6), e0179804–e0179823. doi:10.1371/journal.pone.0179804
- Dalton, L., Ballarin, V., and Brun, M. (2009). Clustering algorithms: On learning, validation, performance, and applications to genomics. *Curr. Genomics* 10 (6), 430–445. doi:10.2174/138920209789177601
- Danaee, P., Ghaeini, R., and Hendrix, D. A. (2017). “A deep learning approach for cancer detection and relevant gene identification,” in Pacific Symposium on Biocomputing 2017 Biocomputing, 219–229. doi:10.1142/9789813207813\_0022
- Davis, J., and Goadrich, M. (2006). ‘The relationship between precision-recall and ROC curves’, In Proceedings of the 23rd international conference on Machine learning, 233–240.
- Dayan, P. (1996). *Unsupervised learning*. The MIT Encyclopedia of the Cognitive Sciences.

- De Guia, J. M., Devaraj, M., and Vea, L. A. (2019). "Cancer classification of gene expression data using machine learning models," in 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control (Environment and Management, HNICEM 2018. IEEE). doi:10.1109/HNICEM.2018.8666435
- Deegalla, S., and Bostr, H. (2007). "Classification of microarrays with kNN : Comparison of dimensionality reduction," in International Conference on Intelligent Data Engineering and Automated Learning (Springer-Verlag), 800–809.
- Deng, L., and Yu, D. (2014). "Deep learning: Methods and applications," in *Foundations and Trends® in signal processing*, 198–349.
- Devarajan, K., and Ebrahimi, N. (2008). Class discovery via nonnegative matrix factorization. *Am. J. Math. Manag. Sci.* 28 (3–4), 457–467. doi:10.1080/01966324.2008.10737738
- Dhote, Y., Agrawal, S., and Deen, A. J. (2015). "A survey on feature selection techniques for internet traffic classification," in Proceedings - 2015 International Conference on Computational Intelligence and Communication Networks (CICN 2015. IEEE), 1375–1380. doi:10.1109/CICN.2015.267
- Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinforma.* 7 (3), 3–13. doi:10.1186/1471-2105-7-3
- Dick, S. (2019). Artificial intelligence. *Harv. Data Sci. Rev.* 1 (1), 1–7. doi:10.4324/9780203772294-10
- Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* 3 (2), 185–205. doi:10.1142/s0219720005001004
- Dittman, D. J., Wald, R., and Hulse, J. (2010). "Comparative analysis of DNA microarray data through the use of feature selection techniques," in Proceedings - 9th International Conference on Machine Learning and Applications (ICMLA 2010. IEEE), 147–152. doi:10.1109/ICMLA.2010.29
- Doran, M., Raicu, D. S., Furst, J. D., Settini, R., SchipMaM.and Chandler, D. P. (2007). Oligonucleotide microarray identification of Bacillus anthracis strains using support vector machines. *Bioinformatics* 23 (4), 487–492. doi:10.1093/bioinformatics/btl626
- Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: A pipeline for processing Illumina microarray. *Bioinformatics* 24 (13), 1547–1548. doi:10.1093/bioinformatics/btm224
- Dubey, A., and Rasool, A. (2021). Efficient technique of microarray missing data imputation using clustering and weighted nearest neighbour', *Scientific Reports. Sci. Rep.* 11 (1), 24297–24312. doi:10.1038/s41598-021-03438-x
- Dudoit, S., and Fridlyannnd, J. (2005). "Classification in microarray experiments," in *A practical approach to microarray data analysis*, 132–149. doi:10.1007/0-306-47815-3\_7
- Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* 4 (1), 95–104. doi:10.1080/01969727408546059
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. doi:10.1126/science.1162986
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868. doi:10.1073/pnas.95.25.14863
- Eisenstein, M. (2012). Oxford Nanopore announcement sets sequencing sector abuzz'. *Nat. Biotechnol.* 30 (4), 295–296. doi:10.1038/nbt0412-295
- Fan, L., Poh, K. L., and Zhou, P. (2009). 'A sequential feature extraction approach for naive bayes classification of microarray data'. *Expert Syst. Appl.* 36, 9919–9923. doi:10.1016/j.eswa.2009.01.075
- Farswan, A., Gupta, A., Gupta, R., and Kaur, G. (2020). Imputation of gene expression data in blood cancer and its significance in inferring biological pathways. *Front. Oncol.* 9, 1442–1514. doi:10.3389/fonc.2019.01442
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 861–874. doi:10.1016/j.patrec.2005.10.010
- Fernandez-Castillo, E., Barbosa-Santillan, L. I., Falcon-Morales, L., and Sanchez-Escobar, J. J. (2022). Deep splicer: A CNN model for splice site prediction in genetic sequences. *Genes* 13 (5), 907. doi:10.3390/genes13050907
- Fernández-Delgado, M., Sirsat, M. S., Cernadas, E., Alawadi, S., Barro, S., and Febrero-BandeM. (2019). An extensive experimental survey of regression methods. *Neural Netw.* 111, 11–34. doi:10.1016/j.neunet.2018.12.010
- Franks, J. M., Cai, G., and Whitfield, M. L. (2018). Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data. *Bioinformatics* 34 (11), 1868–1874. doi:10.1093/bioinformatics/bty026
- Freyhult, E., Landfors, M., Onskog, J., Hvidsten, T. R., and Ryden, P. (2010). Challenges in microarray class discovery: A comprehensive examination of normalization, gene selection and clustering. *BMC Bioinforma.* 11 (1), 503–514. doi:10.1186/1471-2105-11-503
- Friedman, N., LiniM.Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7 (3–4), 601–620. doi:10.1089/106652700750050961
- Frommlet, F., Szulc, P., König, F., and Bogdan, M. (2022). Selecting predictive biomarkers from genomic data. *Plos One* 17 (6), e0269369. doi:10.1371/journal.pone.0269369
- Furey, T. S., CristianNiNiN.DuffyN.Bednarski, D. W., SchuMMerM.and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16 (10), 906–914. doi:10.1093/bioinformatics/16.10.906
- Gan, X., Liew, A. W. C., and Yan, H. (2006). Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Res.* 34 (5), 1608–1619. doi:10.1093/nar/gkl047
- García-Laencina, P. J., Sancho-Gómez, J. L., and Figueiras-Vidal, A. R. (2008). "Machine learning techniques for solving classification problems with missing input data," in Proceedings of the 12th World Multi-Conference on Systems, Cybernetics and Informatics, 1–6.
- Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). Affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20 (3), 307–315. doi:10.1093/bioinformatics/btg405
- Gentleman, R., and Carey, V. J. (2008). "Unsupervised machine learning", in *Bioconductor case studies* (New York: Springer), 137–157. doi:10.1007/978-0-387-77240-0\_7
- Goffinet, B., and Wallach, D. (1989). Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecol. Model.* 44, 299–306. doi:10.1016/0304-3800(89)90035-5
- Guo, Y., and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* 8 (1), 86–100. doi:10.1093/biostatistics/kxj035
- Guruprasad, K., Reddy, B. V. B., and Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting *in vivo* stability of a protein from its primary sequence. *Protein Eng.* 4 (2), 155–161. doi:10.1093/protein/4.2.155
- Guyon, I., Matin, N., and Vapnik, V. (1996). *Discovering informative patterns and data cleaning*, 145–156.
- Guyon, I., Weston, J., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* (46), 62–72. doi:10.1007/978-3-540-88192-6-8
- Hambali, M. A., Oladele, T. O., and Adewole, K. S. (2020). Microarray cancer feature selection: Review, challenges and research directions. *Int. J. Cognitive Comput. Eng.* 1 (11), 78–97. doi:10.1016/j.ijcce.2020.11.001
- Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 13 (2), 204–216. doi:10.1093/biostatistics/kxr054
- Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., et al. (2008). Single-molecule DNA sequencing of a viral genome. *Science* 320 (5872), 106–109. doi:10.1126/science.1150427
- Hijikata, A., Kitamura, H., Kimura, Y., Yokoyama, R., Aiba, Y., Bao, Y., et al. (2007). Construction of an open-access database that integrates cross-reference information from the transcriptome and proteome of immune cells. *Bioinformatics* 23 (21), 2934–2941. doi:10.1093/bioinformatics/btm430
- Hoffmann, R. (2007). Text mining in genomics and proteomics. *Fundam. Data Min. Genomics Proteomics* 9780387475, 251–274. doi:10.1007/978-0-387-47509-7\_12
- Holzinger, A., Biemann, C., and Kell, D. (2017). *What do we need to build explainable AI systems for the medical domain?* 1–28. *arXiv preprint arXiv:1712.09923*.
- Hu, J., Li, H., Waterman, M. S., and Zhou, X. J. (2006). Integrative missing value estimation for microarray data. *BMC Bioinforma.* 7, 449–514. doi:10.1186/1471-2105-7-449
- Huang, C., Clayton, E. A., Matyunina, L. V., McDonald, L. D., Benigno, B. B., Vannberg, F., et al. (2018). Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. *Sci. Rep.* 8 (1), 16444–16449. doi:10.1038/s41598-018-34753-5
- Huang, H. J., Campana, R., Akinfenwa, O., Curin, M., Sarzsinszky, E., Karsonova, A., et al. (2021). Microarray-based allergy diagnosis: Quo vadis? *Front. Immunol.* 11, 594978–595015. doi:10.3389/fimmu.2020.594978
- Hyvärinen, A. (2013). Independent component analysis: Recent advances. *Philos. Trans. A Math. Phys. Eng. Sci.* 371, 20110534. doi:10.1098/rsta.2011.0534



- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi:10.1093/biostatistics/4.2.249
- Jagga, Z., and Gupta, D. (2015). Machine learning for biomarker identification in cancer research - developments toward its clinical application. *Per. Med.* 12 (6), 371–387. doi:10.2217/pme.15.5
- Jenike, M. A., and Albert, M. S. (1984). The dexamethasone suppression test in patients with presenile and senile dementia of the Alzheimer's type. *J. Am. Geriatr. Soc.* 32 (6), 441–444. doi:10.1111/j.1532-5415.1984.tb02220.x
- Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer.
- Jörnsten, R., Wang, H. Y., Welsh, W. J., and Ouyang, M. (2005). DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics* 21 (22), 4155–4161. doi:10.1093/bioinformatics/bti638
- Jothi, R., Mohanty, S. K., and Ojha, A. (2019). DK-Means: A deterministic K-means clustering algorithm for gene expression analysis. *Pattern Anal. Appl.* 22 (2), 649–667. doi:10.1007/s10044-017-0673-0
- Kang, M., and Jameson, N. J. (2018). 'Machine learning: Fundamentals'. *Prognostics Health Manag. Electron.*, 85–109. doi:10.1002/9781119515326.ch4
- Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., and Wu, A. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (7), 881–892. doi:10.1109/tpami.2002.1017616
- Karthik, S., and Sudha, M. (2018). A survey on machine learning approaches in gene expression classification in modelling computational diagnostic system for complex diseases. *Int. J. Eng. Adv. Technol.* 8 (2), 182–191.
- Karthik, S., and Sudha, M. (2021). Predicting bipolar disorder and schizophrenia based on non-overlapping genetic phenotypes using deep neural network. *Evol. Intell.* 14 (2), 619–634. doi:10.1007/s12065-019-00346-y
- Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.* 8 (2), e1002375. doi:10.1371/journal.pcbi.1002375
- Kia, D. A., Zhang, D., Gueffi, S., Manzoni, C., Hubbard, L., Reynolds, R. H., et al. (2021). Identification of candidate Parkinson disease genes by integrating genome-wide association study, expression, and epigenetic data sets. *JAMA Neurol.* 78 (4), 464–472. doi:10.1001/jamaneurol.2020.5257
- Kim, H., and Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* 23 (12), 1495–1502. doi:10.1093/bioinformatics/btm134
- Kim, P., and Tidor, B. (2003). Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.* 13 (7), 1706–1718. doi:10.1101/gr.903503
- Kira, K., and Rendell, L. A. (1992). "A practical approach to feature selection, machine learning," in Proceedings of the Ninth International Workshop (ML92) (Burlington, Massachusetts: Morgan Kaufmann Publishers, Inc). doi:10.1016/B978-1-55860-247-2.50037-1
- Kodama, Y., Mashima, J., Kosuge, T., and Ogasawara, O. (2019). DDBJ update: The Genomic Expression Archive (GEA) for functional genomics data. *Nucleic Acids Res.* 47 (1), D69–D73. doi:10.1093/nar/gky1002
- Kong, W., Mou, X., and Hu, X. (2011). Exploring matrix factorization techniques for significant genes identification of Alzheimer's disease microarray gene expression data. *BMC Bioinforma.* 12 (5), 7–9. doi:10.1186/1471-2105-12-S5-S7
- Kong, W., Vanderburg, C. R., Gunshin, H., Rogers, J. T., and Huang, X. (2008). A review of independent component analysis application to microarray gene expression data. *BioTechniques* 45 (5), 501–520. doi:10.2144/000112950
- Kotsiantis, S., and Kanellopoulos, D. (2006). Association rules mining: A recent overview. *Science* 32 (1), 71–82.
- Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. *Informatica* 31, 249–268. doi:10.1007/s10751-016-1232-6
- Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* 37 (2), 233–243. doi:10.1002/aic.690370209
- Krętowski, M., and Grześ, M. (2007). Decision tree approach to microarray data analysis. *Biocybern. Biomed. Eng.* 27 (3), 29–42.
- Kumar, M., Rath, N. K., Swain, A., and Rath, S. K. (2015). Feature selection and classification of microarray data using MapReduce based ANOVA and K-nearest neighbor. *Procedia Comput. Sci.* 54, 301–310. doi:10.1016/j.procs.2015.06.035
- Lai, Y. H., Chen, W. N., Hsu, T. C., Lin, C., Tsao, Y., and Wu, S. (2020). Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Sci. Rep.* 10 (1), 4679–4711. doi:10.1038/s41598-020-61588-w
- Lakiotaki, K., Vorniotakis, N., Tsagris, M., Georgakopoulos, G., and Tsamardinos, I. (2018). BioDataome: A collection of uniformly preprocessed and automatically annotated datasets for data-driven biology. *Database (Oxford)*. 2018, 1–14. doi:10.1093/database/bay011
- Land, W. H., Qiao, X., Margolis, D. E., Ford, W. S., Paquette, C. T., Perez-Rogers, J. F., et al. (2011). Kernelized partial least squares for feature reduction and classification of gene microarray data. *BMC Syst. Biol.* 5, S13. doi:10.1186/1752-0509-5-S3-S13
- Langfelder, P., and Horvath, S. (2008). Wgcna: An R package for weighted correlation network analysis. *BMC Bioinforma.* 9, 559. doi:10.1186/1471-2105-9-559
- Larsen, M. J., Thomassen, M., Tan, Q., Sorensen, K. P., and Kruse, T. A. (2014). Microarray-based RNA profiling of breast cancer: Batch effect removal improves cross-platform consistency. *Biomed. Res. Int.* 2014, 651751. doi:10.1155/2014/651751
- Lazar, C., Gatto, L., Ferro, M., Bruley, C., and Burger, T. (2016). Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *J. Proteome Res.* 15 (4), 1116–1125. doi:10.1021/acs.jproteome.5b00981
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 13 (1), 436–444. doi:10.1038/nature14539
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86 (11), 2278–2324. doi:10.1109/5.726791
- Lee, S., and Batzoglu, S. (2003). Application of independent component analysis to microarrays. *Genome Biol.* 4 (11), R76–R21. doi:10.1186/gb-2003-4-11-r76
- Li, E., Luo, T., and Wang, Y. (2019). Identification of diagnostic biomarkers in patients with gestational diabetes mellitus based on transcriptome gene expression and methylation correlation analysis. *Reproductive Biology and Endocrinology. Reprod. Biol. Endocrinol.* 17 (1), 112–12. doi:10.1186/s12958-019-0556-x
- Li, H., Zhao, C., Shao, F., Li, G. Z., and Wang, X. (2015). A hybrid imputation approach for microarray missing value estimation. *BMC Genomics* 16 (9), 1–11. doi:10.1186/1471-2164-16-S9-S1
- Li, W., Suh, Y. J., and Zhang, J. (2006). "Does logarithm transformation of microarray data affect ranking order of differentially expressed genes?," in Conf. Proc. IEEE Eng. Med. Biol. Soc., 6593–6596. doi:10.1109/IEMBS.2006.260896
- Li, X., Li, M., and Yin, M. (2016). Multiobjective ranking binary artificial bee colony for gene selection problems using microarray datasets. *IEEE/CAA J. Autom. Sin.*, 1–16. doi:10.1109/JAS.2016.7510034
- Li, Z., Xie, W., and Liu, T. (2018). Efficient feature selection and classification for microarray data. *PLoS ONE* 13 (8), 02021677–e202221. doi:10.1371/journal.pone.0202167
- Liew, A. W. C., Law, N. F., and Yan, H. (2011). Missing value imputation for gene expression data: Computational techniques to recover missing data from available information. *Brief. Bioinform.* 12 (5), 498–513. doi:10.1093/bib/bbq080
- Liu, Y.-C., Cheng, C.-P., and Tseng, V. S. (2011). Discovering relational-based association rules with multiple minimum supports on microarray datasets. *Bioinformatics* 27 (22), 3142–3148. doi:10.1093/bioinformatics/btr526
- Liu, Y. (2008). Detect key gene information in classification of microarray data. *EURASIP J. Adv. Signal Process.*, 612397. doi:10.1155/2008/612397
- Liu, Y. (2009). Prominent feature selection of microarray data. *Prog. Nat. Sci.* 19 (10), 1365–1371. doi:10.1016/j.pnsc.2009.01.014
- Liu, Z., Sokka, T., Maas, K., Olsen, N. J., and Aune, T. M. (2009). Prediction of disease severity in patients with early rheumatoid arthritis by gene expression profiling. *Hum. Genomics Proteomics*. 1 (1), 484351. doi:10.4061/2009/484351
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550–621. doi:10.1186/s13059-014-0550-8
- Lu, H., Xie, R. D., Lin, R., Zhang, C., Xiao, X. J., Li, L. J., et al. (2017). Vitamin D-deficiency induces eosinophil spontaneous activation. *Cell. Immunol.* 256, 56–63. doi:10.1016/j.cellimm.2017.10.003
- Lu, Y., Lu, S., and Deng, Y. (2004). Fgka: A fast genetic K-means clustering algorithm. *Proc. ACM Symposium Appl. Comput.* 1, 622–623. doi:10.1145/967900.968029
- Ma, S., Song, X., and Huang, J. (2007). Supervised group Lasso with applications to microarray data analysis. *BMC Bioinforma.* 8, 60–17. doi:10.1186/1471-2105-8-60



- Mack, C., Su, Z., and Westreich, D. (2018). Managing missing data in patient registries: Addendum to registries for evaluating patient outcomes. *A User's Guide*.
- MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations," in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 281–297. doi:10.1007/s11665-016-2173-6
- Manikandan, G., and Abirami, S. (2018). "A survey on feature selection and extraction techniques for high-dimensional microarray datasets," in *Knowledge computing and its applications* (Springer Singapore), 311–333.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437 (7057), 376–380. doi:10.1038/nature03959
- Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405 (2), 442–451. doi:10.1016/0005-2795(75)90109-9
- McNee, S. M., Riedl, J., and Konstan, J. A. (2006). "Being accurate is not enough: How accuracy metrics have hurt recommender systems," in Conference on Human Factors in Computing Systems - Proceedings, 1097–1101. doi:10.1145/1125451.1125659
- McNicholas, P. D., and Murphy, T. B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics* 26 (21), 2705–2712. doi:10.1093/bioinformatics/btq498
- Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70 (1), 53–71. doi:10.1111/j.1467-9868.2007.00627.x
- Micheuz, P. (2020). "Approaches to artificial intelligence as a subject in school education," in Open Conference on Computers in Education (Cham.: Springer), 3–13.
- Moorthy, K., Jaber, A. N., Ismail, M. A., Ernawan, F., Mohamad, M. S., and Deris, S. (2019). Missing-values imputation algorithms for microarray gene expression data. *Methods Mol. Biol.*, 255–266. doi:10.1007/978-1-4939-9442-7\_12
- Moorthy, K., and Mohamad, M. S. (2012). Random forest for gene selection and microarray data classification. *Bioinformation* 7 (3), 142–146. doi:10.6026/97320630007142
- Morais-Rodrigues, F., Silv Erio-Machado, R., Kato, R. B., Rodrigues, D. L. N., Valdez-Baez, J., Fonseca, V., et al. (2020). Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. *Gene* 726, 144168–8. doi:10.1016/j.gene.2019.144168
- 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 (8), 246–254. doi:10.1016/j.imu.2017.10.004
- Neubauer, C. (1998). Evaluation of convolutional neural networks for visual recognition. *IEEE Trans. Neural Netw.* 9 (4), 685–696. doi:10.1109/72.701181
- Nguyen, N. G., Tran, V. A., Ngo, D. L., Phan, D., Lumbanraja, F. R., Faisal, M. R., et al. (2016). DNA sequence classification by convolutional neural network. *J. Biomed. Sci. Eng.* 09 (05), 280–286. doi:10.4236/jbise.2016.95021
- Nidheesh, N., Abdul Nazeer, K. A., and Ameer, P. M. (2017). An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data. *Comput. Biol. Med.* 91, 213–221. doi:10.1016/j.compbimed.2017.10.014
- Nikkila, J., Toronen, P., Kaski, S., Venna, J., Castren, E., and Wong, G. (2002). Analysis and visualization of gene expression data using Self-Organizing Maps. *Neural Netw.* 15, 953–966. doi:10.1016/s0893-6080(02)00070-9
- Nikumbh, S., Ghosh, S., and Jayaraman, V. K. (2012). "Biogeography-based informative gene selection and cancer classification using SVM and Random Forests," in 2012 IEEE Congress on Evolutionary Computation (Brisbane, QLD: CEC 2012), 1–6. doi:10.1109/CEC.2012.6256127
- Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K. i., and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 19 (16), 2088–2096. doi:10.1093/bioinformatics/btg287
- O'Connell, M. (2003). Differential expression, class discovery and class prediction using S-PLUS and S+ArrayAnalyzer. *SIGKDD Explor. Newsl.* 5 (2), 38–47. doi:10.1145/980972.980979
- Oladejo, A. K., Oladele, T. O., and Saheed, Y. K. (2018). Comparative evaluation of linear support vector machine and K-nearest neighbour algorithm using microarray data on leukemia cancer dataset. *Afr. J. Comput. ICT* 11 (2), 1–10.
- Önskog, J., Freyhult, E., Landfors, M., Ryden, P., and Hvidsten, T. R. (2011). Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. *BMC Bioinforma.* 12, 390. doi:10.1186/1471-2105-12-390
- O'Shea, K., and Nash, R. (2015). *An introduction to convolutional neural networks*, 1–11. arXiv preprint, arXiv:1511.
- Ouyang, M., Welsh, W. J., and Georgopoulos, P. (2004). Gaussian mixture clustering and imputation of microarray data. *Bioinformatics* 20 (6), 917–923. doi:10.1093/bioinformatics/bth007
- Pan, H., Zhu, J., and Han, D. (2003). Genetic algorithms applied to multi-class clustering for gene ex- pression data partitioned clustering techniques'. *Genomics Proteomics Bioinforma.* 1 (4), 279–287. doi:10.1016/S1672-0229(03)01033-7
- Pan, W., Lin, J., and Le, C. T. (2002). Model-based cluster analysis of microarray gene-expression data. *Genome Biol.* 3 (2), RESEARCH0009–8. doi:10.1186/gb-2002-3-2-research0009
- Pan, X., Tian, Y., Huang, Y., and Shen, H. B. (2011). Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach', *Genomics. Genomics* 97 (5), 257–264. doi:10.1016/j.ygeno.2011.03.001
- Pan, X., and Yan, J. (2017) 'Attention based convolutional neural network for predicting RNA-protein binding sites', arXiv preprint, arXiv:1712, pp. 8–11.
- Parihar, A., Mondal, S., and Singh, R. (2022). "Introduction, scope, and applications of biotechnology and genomics for sustainable agricultural production," in *Plant genomics for sustainable agriculture*. Editor R. Lakhan (Springer), 1–14. doi:10.1007/978-981-16-6974-3
- Parikh, R., Andjelković Apostolović, M., and Stojanović, D. (2008a). Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.* 56 (1), 341–350. doi:10.4103/0301-4738.41424
- Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., and Thomas, R. (2008b). Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.* 56 (1), 45–50. doi:10.4103/0301-4738.37595
- Park, C., Ha, J., and Park, S. (2020). Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. *Expert Syst. Appl.* 140, 112873. doi:10.1016/j.eswa.2019.112873
- Park, H.-S., Yoo, S.-H., and Cho, S.-B. (2007). Forward selection method with regression analysis for optimal gene selection in cancer classification. *Int. J. Comput. Math.* 84 (5), 653–667. doi:10.1080/00207160701294384
- Pease, A. C., Solas, D., and Sullivan, E. J. (1994). "Light-generated oligonucleotide arrays for rapid DNA sequence analysis," in Proceedings of the National Academy of Sciences of the United States of America, 5022–5026. doi:10.1073/pnas.91.11.5022
- Peng, J., Guan, J., and Shang, X. (2019). Predicting Parkinson's disease genes based on node2vec and autoencoder. *Front. Genet.* 10, 226–6. doi:10.3389/fgene.2019.00226
- Peng, Y., Li, W., and Liu, Y. (2006). A hybrid approach for biomarker discovery from microarray gene expression data for cancer classification. *Cancer Inf.* 2, 117693510600200–117693510600311. doi:10.1177/117693510600200024
- Peterson, L. E., and Coleman, M. A. (2008). Machine learning-based receiver operating characteristic (ROC) curves for crisp and fuzzy classification of DNA microarrays in cancer research. *Int. J. Approx. Reason.* 47 (1), 17–36. doi:10.1016/j.ijar.2007.03.006
- Pirooznia, M., Yang, J. Y., Yang, M. Q., and Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics* 9 (1), S13–S13. doi:10.1186/1471-2164-9-S1-S13
- Pochet, N., De Smet, F., Suykens, J. A. K., and De Moor, B. L. R. (2004). Systematic benchmarking of microarray data classification: Assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* 20 (17), 3185–3195. doi:10.1093/bioinformatics/bth383
- Prasanna, K., Seetha, M., and Kumar, A. P. S. (2014). "CApriori: Conviction based Apriori algorithm for discovering frequent determinant patterns from high dimensional datasets," in 2014 International Conference on Science Engineering and Management Research, ICSEMR 2014 (IEEE). doi:10.1109/ICSEMR.2014.7043622
- Qiu, Y. L., Zheng, H., and Gevaert, O. (2018). *A deep learning framework for imputing missing values in genomic data*. BioRxiv, 406066.
- Qiu, Y. L., Zheng, H., and Gevaert, O. (2020). Genomic data imputation with variational auto-encoders. *Gigascience*, 9, gaa082–12. doi:10.1093/gigascience/giaa082
- Quackenbush, J. (2001). Computational analysis of microarray data. *Nat. Rev. Genet.* 2, 418–427. doi:10.1038/35076576
- Radovic, M., Ghalwash, M., Filipovic, N., and Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data'. *BMC Bioinforma.* 18 (1), 9–14. doi:10.1186/s12859-016-1423-9
- Ram, P. K., and Kuila, P. (2019). Feature selection from microarray data : Genetic algorithm based approach. *J. Inf. Optim. Sci.* 40 (8), 1599–1610. doi:10.1080/02522667.2019.1703260
- Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. *Encycl. Database Syst.* 5, 532–538. doi:10.1007/978-0-387-39940-9\_565

- Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinforma.* 13 (5), 278–289. doi:10.1016/j.gpb.2015.08.002
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “Why should I trust you?” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 1135–1144. doi:10.1145/2939672.2939778
- Ringnér, M. (2008). What is principal component analysis. *Nat. Biotechnol.* 26 (3), 303–304. doi:10.1038/nbt0308-303
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43 (7), e47. doi:10.1093/nar/gkv007
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26 (1), 139–140. doi:10.1093/bioinformatics/btp616
- Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475 (7356), 348–352. doi:10.1038/nature10242
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi:10.1016/0377-0427(87)90125-7
- Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63 (3), 581–592. doi:10.1093/biomet/63.3.581
- Ryan, C., Greene, D., Cagney, G., and Cunningham, P. (2010). Missing value imputation for epistatic MAPs. *BMC Bioinforma.* 11, 197. doi:10.1186/1471-2105-11-197
- Saey, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23 (19), 2507–2517. doi:10.1093/bioinformatics/btm344
- Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man. Cybern.* 21 (3), 660–674. doi:10.1109/21.97458
- Saha, S., Ghost, A., and Dey, K. (2017). “An ensemble based missing value estimation in DNA microarray using artificial neural network,” in Proceedings - 2016 2nd IEEE International Conference on Research in Computational Intelligence and Communication Networks, February 2019 (Kolkata, India: ICRCIN 2016), 279–284. doi:10.1109/ICRCIN.2016.7813671
- Sahu, B., and Mishra, D. (2012). A novel feature selection algorithm using particle swarm optimization for cancer microarray data. *Procedia Eng.* 38, 27–31. doi:10.1016/j.proeng.2012.06.005
- Sahu, M. A., Swarnkar, M. T., and Das, M. K. (2011). Estimation methods for microarray data with missing values: A review. *Int. J. Comput. Sci. Inf. Technol.* 2 (2), 614–620.
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74 (12), 5463–5467. doi:10.1073/pnas.74.12.5463
- Sayed, S., Nassef, M., Badr, A., and Farag, I. (2019). A Nested Genetic Algorithm for feature selection in high-dimensional cancer Microarray datasets. *Expert Syst. Appl.* 121 (C), 233–243. doi:10.1016/j.eswa.2018.12.022
- Schafer, J. L., and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychol. Methods* 7 (2), 147–177. doi:10.1037/1082-989X.7.2.147
- Schmidt, L. J., Murillo, H., and Tindall, D. J. (2004). Gene expression in prostate cancer cells treated with the dual 5 alpha-reductase inhibitor dutasteride. *J. Androl.* 25 (6), 944–953. doi:10.1002/j.1939-4640.2004.tb03166.x
- Segundo-Val, I. S., and Sanz-Lozano, C. S. (2016). Introduction to the gene expression analysis. *Methods Mol. Biol.* 1434, 29–43. doi:10.1007/978-1-4939-3652-6\_3
- Sharma, A., Paliwal, K. K., Imoto, S., and Miyano, S. (2014). A feature selection method using improved regularized linear discriminant analysis. *Mach. Vis. Appl.* 25, 775–786. doi:10.1007/s00138-013-0577-y
- Sharma, A., and Rani, R. (2021). ‘A systematic review of applications of machine learning in cancer prediction and diagnosis’. *Arch. Comput. Methods Eng.* 28, 4875–4896. doi:10.1007/s11831-021-09556-z
- Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., et al. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309, 1728–1732. doi:10.1126/science.1117389
- Smith, G. S., and Snyder, R. L. (1979).  $F_{\langle D \rangle N \langle I \rangle S}$ : A criterion for rating powder diffraction patterns and evaluating the reliability of powder-pattern indexing. *J. Appl. Crystallogr.* 12, 60–65. doi:10.1107/s002188987901178x
- Smyth, G. K., and Speed, T. (2003). Normalization of cDNA microarray data. *Methods* 31 (4), 265–273. doi:10.1016/s1046-2023(03)00155-5
- Smyth, G. K. (2005). ‘limma: Linear models for microarray data’. *Bioinforma. Comput. Biol. Solutions Using R Bioconductor* 11, 397–420. doi:10.1007/0-387-29362-0\_23
- Souto, M. C. P. D., Jaskowiak, P. A., and Costa, I. G. (2015). Impact of missing data imputation methods on gene expression clustering and classification. *BMC Bioinforma.* 16, 64–69. doi:10.1186/s12859-015-0494-3
- Statnikov, A., Wang, L., and Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinforma.* 9, 319–410. doi:10.1186/1471-2105-9-319
- Storey, J., and Tibshirani, R. (2003). “Statistical methods for identifying differentially expressed genes in DNA microarrays,” in *Methods in molecular biology* (Totowa, NJ: Humana Press), 149–157.
- Sturm, B. L. (2013). Classification accuracy is not enough: On the evaluation of music genre recognition systems. *J. Intell. Inf. Syst.* 41, 371–406. doi:10.1007/s10844-013-0250-y
- Subashini, P., and Krishnaveni, M. (2011). “Imputation of missing data using bayesian principal component analysis on tec ionospheric satellite dataset,” in Canadian Conference on Electrical and Computer Engineering (IEEE), 001540–001543. doi:10.1109/CCECE.2011.6030724
- Tabares-Soto, R., Orozco-Arias, S., Romero-Cano, V., Segovia Bucheli, V., Rodriguez-Sotelo, J. L., and Jimenez-Varon, C. F. (2020). A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ. Comput. Sci.* 6 (207), 2700–e322. doi:10.7717/peerj-cs.270
- Tamayo, P., Slonim, D., and Zhu, Q. (1999). “Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation,” in Proceedings of the National Academy of Sciences of the United States of America, 2907–2912. doi:10.1073/pnas.96.6.2907
- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat. Genet.* 22 (3), 281–285. doi:10.1038/10343
- Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.* 18 (1), 104–117. doi:10.1214/ss/1056397488
- Tibshirani, B. R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58 (1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x
- Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* 1, A68–A77. doi:10.5114/wo.2014.47136
- Toro-Domínguez, D., Lopez-Dominguez, R., Garcia Moreno, A., Villatoro-Garcia, J. A., Martorell-Marugan, J., Goldman, D., et al. (2019). Differential treatments based on drug-induced gene expression signatures and longitudinal systemic lupus erythematosus stratification. *Sci. Rep.* 9 (1), 15502–15509. doi:10.1038/s41598-019-51616-9
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* 17 (6), 520–525. doi:10.1093/bioinformatics/17.6.520
- Tuikkala, J., Elo, L. L., Nevalainen, O. S., and Aittokallio, T. (2008). Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC Bioinforma.* 9, 202–214. doi:10.1186/1471-2105-9-202
- Tuikkala, J., Elo, L., Nevalainen, O. S., and Aittokallio, T. (2006). Improving missing value estimation in microarray data with gene ontology. *Bioinformatics* 22 (5), 566–572. doi:10.1093/bioinformatics/btk019
- Turgut, S., Dagtekin, M., and Ensari, T. (2018). “Microarray breast cancer data classification using machine learning methods,” in 2018 Electric Electronics, Computer Science, Biomedical Engineering Meeting, EBBT 2018 (IEEE), 1–3. doi:10.1109/EBBT.2018.8391468
- Tyagi, V., and Mishra, A. (2013). A survey on different feature selection methods for microarray data analysis. *Int. J. Comput. Appl.* 67 (16), 36–40. doi:10.5120/11482-7181
- Uhl, M., Tran, V. D., Heyl, F., and Backofen, R. (2021). RNAProt: An efficient and feature-rich RNA binding protein binding site predictor. *Gigascience*, 10. GigaScience, giab054–13. doi:10.1093/gigascience/giab054
- Umarov, R. K., and Solovyev, V. V. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS ONE* 12 (2), e0171410–e0171412. doi:10.1371/journal.pone.0171410
- Valouev, A., Ichikawa, J., Thontat, T., Stuart, J., Ranade, S., Peckham, H., et al. (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18 (7), 1051–1063. doi:10.1101/gr.076463.108
- Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC genomics* 13, S2–S10. doi:10.1186/1471-2164-13-S4-S2

- Vincent, P., Larochelle, H., and Lajoie, I. (2010). Stacked denoising Autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* 11, 3371–3408.
- Vincent, P., and Larochelle, H. (2008). “Extracting and composing robust features with denoising,” in Proceedings of the 25th international conference on Machine learning, 1096–1103.
- Vo, A. H., Van Vleet, T. R., Gupta, R. R., Liguori, M. J., and Rao, M. S. (2020). An overview of machine learning and big data for drug toxicity evaluation. *Chem. Res. Toxicol.* 33 (1), 20–37. doi:10.1021/acs.chemrestox.9b00227
- Wang, A., Chen, Y., An, N., Yang, J., Li, L., and Jiang, L. (2019). Microarray missing value imputation: A regularized local learning method. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16 (3), 980–993. doi:10.1109/TCBB.2018.2810205
- Wang, X., Li, A., Jiang, Z., and Feng, H. (2006). Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. *BMC Bioinforma.* 7, 32–10. doi:10.1186/1471-2105-7-32
- Winston, P. H. (1992). *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc. ACM digital library.
- Xiang, Q., Dai, X., Deng, Y., He, C., Wang, J., Feng, J., et al. (2008). Missing value imputation for microarray gene expression data using histone acetylation information. *BMC Bioinforma.* 9, 1–17. doi:10.1186/1471-2105-9-252
- Yang, Y., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., et al. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30 (4), e15–e10. doi:10.1093/nar/30.4.e15
- Yip, W., Amin, S. B., and Li, C. (2011). “A survey of classification techniques for microarray data analysis,” in *Handbook of statistical bioinformatics springer* (Berlin, Heidelberg: Springer Berlin Heidelberg), 193–223. doi:10.1007/978-3-642-16345-610
- Yu, L., and Liu, H. (2003). “Feature selection for high-dimensional data: A fast correlation-based filter solution,” in Proceedings, Twentieth International Conference on Machine Learning, 856–863.
- Yuxi, L., Schukat, M., and Howley, E. (2018) ‘Deep reinforcement learning: An overview’, arXiv preprint arXiv:1701.07274, 16, pp. 426–440. doi: doi:10.1007/978-3-319-56991-8\_32
- Zeebaree, D. Q., Haron, H., and Abdulazeez, A. M. (2018). “Gene selection and classification of microarray data using convolutional neural network,” in International Conference on Advanced Science and Engineering (ICOASE) (IEEE), 145–150. doi:10.1109/ICOASE.2018.8548836
- Zhang, X., Jonassen, I., and Goksøyr, A. (2021). Machine learning approaches for biomarker discovery using gene expression data. *Bioinformatics*, 53–64.
- Zhang, Y., Yang, Y., Wang, C., Wan, S., Yao, Z., and Zhang, Y. (2020). Identification of diagnostic biomarkers of osteoarthritis based on multi-chip integrated analysis and machine learning. *DNA Cell Biol.* 39, 2245–2256. doi:10.1089/dna.2020.5552
- Zheng, C. H., Huang, D. S., and Shang, L. (2006). Feature selection in independent component subspace for microarray data classification. *Neurocomputing* 69, 2407–2410. doi:10.1016/j.neucom.2006.02.006
- Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. *Nat. Genet.* 51 (1), 12–18. doi:10.1038/s41588-018-0295-5

# 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](https://frontiersin.org)

### Contact us

+41 (0)21 510 17 00  
[frontiersin.org/about/contact](https://frontiersin.org/about/contact)

