Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 6691096, 7 pages
https://doi.org/10.1155/2021/6691096

Research Article

Integrated Learning: Screening Optimal Biomarkers for Identifying Preeclampsia in Placental mRNA Samples

Rong Guo,[1] Zhixia Teng,[1] Yiding Wang,[1] Xin Zhou,[1] Heze Xu,[2] and Dan Liu[1]

[1] *Information and Computer Engineering College, Northeast Forestry University, Harbin 150040, China*
[2] *Department of Gynecology and Obstetrics, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China*

Correspondence should be addressed to Dan Liu; nefuliudan@nefu.edu.cn

Received 8 November 2020; Revised 17 January 2021; Accepted 27 January 2021; Published 23 February 2021

Academic Editor: Waqas Haider Bangyal

Preeclampsia (PE) is a maternal disease that causes maternal and child death. Treatment and preventive measures are not sound enough. The problem of PE screening has attracted much attention. The purpose of this study is to screen placental mRNA to obtain the best PE biomarkers for identifying patients with PE. We use Limma in the R language to screen out the 48 differentially expressed genes with the largest differences and used correlation-based feature selection algorithms to reduce the dimensionality and avoid attribute redundancy arising from too many mRNA samples participating in the classification. After reducing the mRNA attributes, the mRNA samples are sorted from large to small according to information gain. In this study, a classifier model is designed to identify whether samples had PE through mRNA in the placenta. To improve the accuracy of classification and avoid overfitting, three classifiers, including C4.5, AdaBoost, and multilayer perceptron, are used. We use the majority voting strategy integrated with the differentially expressed genes and the genes filtered by the best subset method as comparison methods to train the classifier. The results show that the classification accuracy rate has increased from 79% to 82.2%, and the number of mRNA features has decreased from 48 to 13. This study provides clues for the main PE biomarkers of mRNA in the placenta and provides ideas for the treatment and screening of PE.

1. Introduction

Preeclampsia (PE) is a pregnancy-specific syndrome that affects 3-5% of pregnant women and is characterized by edema, hypertension, and proteinuria [1]. PE is a multifactor and multigene disease with a family genetic predisposition: assuming a mother had PE, the incidence of PE in her daughters is 20-40%. If a twin is a patient with PE, the incidence of PE in the other twin is 22-47%. PE makes women more susceptible to cardiovascular disease later in life and may affect brain function. However, so far, the genetic pattern is unclear. It remains a major factor in maternal and newborn morbidity and mortality [2]. The only treatment is the termination of pregnancy and delivery of newborns and the placenta [3]. Although the causes of PE are still being discussed, clinical and pathological studies have shown that the core of the pathogenesis of this syndrome is the placenta [4]. The placenta is an important organ shared by the mother and the fetus. It has important biological functions such as substance exchange, metabolism, and barrier function. Abnormal placental function can lead to pregnancy diseases such as PE. Many physiological and biochemical processes related to placental function are coordinated by proteins that form complex networks in the placenta, and the production of proteins requires the participation of RNA.

In the study of the biomarkers of PE, genetic factors were found to be the cause of the disease [5]. In naive Bayesian predictive modeling and path analysis, the quantitative real-time polymerase chain reaction (qRT-PCR) concluded that genes in the placenta are related to PE through genetic testing [6]. The genetic markers of normal fluff in the placenta can express pathology and PE [7]. Clinically relevant subcategories are identified by the gene expression profile of the placenta [8].

However, research on mRNA in the placenta that can be used as a biomarker for PE is still insufficient. mRNA is a direct template for directing protein synthesis, that is, transferring genetic information from DNA to RNA. mRNA is used as a template for protein synthesis to determine the amino acid sequence of the peptide chain produced by gene expression proteins, and proteins such as VEGF, sFlt1, PlGF, SENG, PAPP-A, PP13, HSP70, and HbF have shown certain value in the prediction or diagnosis of PE [9] (review). This helps in the understanding of the pathogenesis of PE. In these studies, it can be inferred that an abnormal expression of mRNA in the placenta is related to the occurrence of PE, and there are mRNA samples that can be used to identify PE.

Traditional medical methods generally can find only a few biomarkers, and the accuracy of disease recognition is difficult to guarantee. Machine learning performs well in feature filtering and processing large amounts of data [10] (review) and performs well in RNA research [11, 12]. In this study, classification algorithms can be used to detect the mRNA biomarkers of PE in the placenta and to screen out the mRNA samples that can be used as biomarkers in the dataset to avoid the omission of biomarkers. By fusing multiple classifiers, the accuracy of the algorithm is improved, and overfitting is reduced.

2. Materials and Methods

2.1. Data Source. The GSE75010 dataset used in the experiment is the placental microarray dataset released on May 16, 2016, for the analysis of placental gene expression profiles. It is a large dataset containing data from seven published studies ($N = 330$) [13]. This dataset was downloaded from the GEO database and contains gene expression data from 157 placentas with PE and 173 placentas without PE. For convenience, we choose 157 highly annotated samples as the experimental dataset to model and test the classification effect.

2.2. Identification of Differentially Expressed Genes. To standardize the microarray data [14], the GSE75010 dataset downloaded from the GEO database was converted to log2 with the Limma package in R3.4.1. The dataset is divided into two categories according to disease status: placenta with PE and healthy control placenta. The two sets of gene expression matrices were compared by the Limma package [15], and the thresholds were set to $|\text{logFC}| > 2$ and $P < 0.01$.

2.3. Feature Selection. Correlation-based feature selection (CFS) [16] is a heuristic algorithm based on filter patterns. It can improve the classification effect by evaluating the correlation among features as well as the correlation between features and categories. It finds the optimal subset of features to avoid redundancy among the features. Differentially expressed mRNA samples do not necessarily belong to the mRNA samples related to PE. The purpose of CFS is to exclude irrelevant differential genes while avoiding too many mRNA samples participating in the classification, thereby improving the accuracy of classification.

$$\text{Merit}_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}. \quad (1)$$

k represents the number of features in the subset, $\overline{r_{cf}}$ represents the average correlation between features and categories, and $\overline{r_{ff}}$ represents the average correlation among features. The Pearson correlation coefficient [17] is used to calculate $\overline{r_{cf}}$ and $\overline{r_{ff}}$ and can be used to measure the correlation between two variables and screen out mRNA samples related to PE.

$$r(X, Y) = \frac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_1^n (X_i - \bar{X})^2 (Y_i - \bar{Y})^2}}. \quad (2)$$

To improve the classification efficiency of the optimal subset, the information gain ratio algorithm is used to sort the mRNA samples in the optimal subset from large to small [18]. It enables the classifier to classify according to the amount of mRNA information, thereby improving the accuracy of classification. This is an algorithm developed by information entropy. The formula is as follows:

$$\text{Ent}(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k. \quad (3)$$

$|y|$ represents the number of categories p_k, which is the proportion of each category feature in the set. The result of this formula represents the entropy of the information carried by mRNA [19]. The smaller the information entropy is, the purer the dataset.

The value of information gain [20] can determine whether to use this mRNA attribute a to divide dataset D. If the information gain is relatively large, this attribute is a better attribute for dividing dataset D.

$$\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \text{Ent}(D^v). \quad (4)$$

Information gain is biased towards selecting features with more values in the same category, but according to the entropy formula, the more features there are, the greater the entropy is. To change the adverse effects of such poor preferences, this study uses the information gain ratio [21] as a method to judge the division of attributes.

$$\text{Gain\_ratio}(D, a) = \frac{\text{Gain}(D, a)}{-\sum_{v=1}^{V} |D^v|/|D| \log_2 |D^v|/|D|} \quad (5)$$

Finally, according to the results of mRNA information gain, we sorted the mRNA samples in the optimal subset from largest to smallest to train the classifier. On the one hand, the information gain ratio can be a measure of the importance of mRNA, and on the other hand, it can be used as a node selection criterion for the C4.5 classifier.

2.4. Classification Algorithm Design. After filtering mRNA as a feature through the above algorithm, to facilitate sample classification, we designed a suitable model. The trained model can use mRNA samples as attributes to identify whether pregnant women have PE. The disadvantage of small sample datasets is that they are easy to overfit during classification, and ensemble learning is one of the basic methods to alleviate this situation to some extent. We chose three different classifiers as subclassifiers.

Subclassifier I is a C4.5 [22] decision tree that selects attributes according to the information gain ratio and has a good classification effect on small sample datasets. To improve operation efficiency, this C4.5 decision tree is generated in the form of a binary tree.

Subclassifier II is a multilayer perceptron [23]. The multilayer perceptron continuously updates the weights through the backpropagation (BP) algorithm. A learning rate that is too low can greatly increase the training time of the model, and a learning rate that is too large can cause underfitting, so the learning rate is set to 0.3. When the standard BP algorithm corrects the weights, a momentum factor is added to each weight change to prevent the multilayer perceptron from falling into a local minimum, and the momentum factor is set to 0.2. The momentum factor value is the opposite of the value of the last weight change, thus affecting the new weight change based on the BP method. The number of mRNA attributes is set to the number of nodes in the hidden layer, the number of training iterations is initialized to 500, and the network is reset at a lower learning rate. If the network deviates from the answer, it will automatically reset and retrain at a lower learning rate.

$$S(x) = \frac{1}{1 + e^{-x}}. \tag{6}$$

The network maps each data point to an interval (0,1) or (-1,0) to achieve the effect of classification.

Subclassifier III is a decision stump [24] integrated by AdaBoost [25]. The AdaBoost algorithm modifies the classifier and sample weights by continuously iterating the training dataset and integrating many weak classifiers into a strong classifier, as shown in the following formula:

$$F(x) = \sum_{t=1}^{T} \alpha_t f_t(x). \tag{7}$$

T represents the number of weak classifiers, α_t represents the weight of the t-th weak classifier, and $f_t(x)$ represents the prediction result of the t-th weak classifier. The final classification decision rules are as follows:

$$\operatorname{sgn}\,(F(x)) = \begin{cases} 1, x > 0, \\ -1, x < 0. \end{cases} \tag{8}$$

Finally, we use majority voting to integrate subclassifiers. This is an ensemble method that uses most of the output results of the subclassifiers as the final classification result. All models and algorithms are built into Weka 3.8.4 [26].

2.5. Evaluation Criteria. Cross validation (CV) [27, 28], sometimes called rotation estimation, is a statistical method proposed by Seymour Geisser to cut data samples into smaller subsets. For small sample data, CV can avoid overfitting to a certain extent, making the training model more versatile, robust, and accurate. This experiment uses a 10-fold CV method to train the model, that is, the dataset is divided into ten parts, each of which uses 9 different parts for training and one to verify the model to ensure that all the datasets are tested. The idea of tenfold CV is shown in Figure 1.

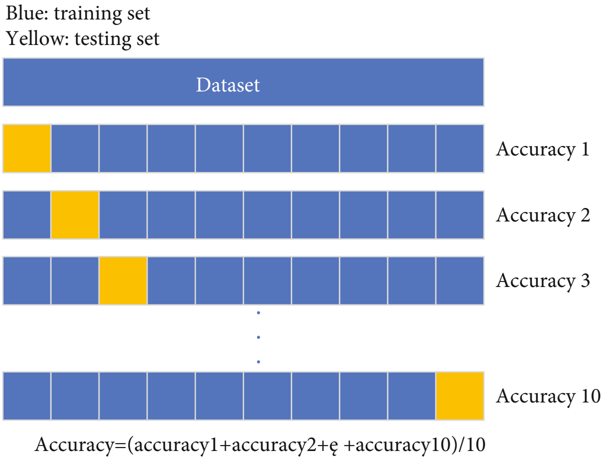

Figure 1: 10-fold cross-validation method.

To facilitate the discussion below, we set the placentas with PE as positive samples and the healthy placentas as negative samples. We use the following indicators as the criteria for evaluating the classifier [29–35]:

$$\begin{aligned} \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\ \text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}}. \end{aligned} \tag{9}$$

Specificity and sensitivity are both indicators of successful model classification. Specificity is an indicator that measures the probability of diagnosis, and sensitivity is an indicator that measures the recognition ability of a classification model. We also introduced the area under the curve (AUC) as an indicator to measure the effectiveness of the model. The AUC is the area between the receiver operating curve (ROC) and the coordinate axis. Its value is in the interval (0.5,1). The closer the AUC value is to 1, the better the classifier is.

3. Results and Discussion

The results of the differential gene analysis are represented by a volcano plot in Figure 2(green indicates a relatively low

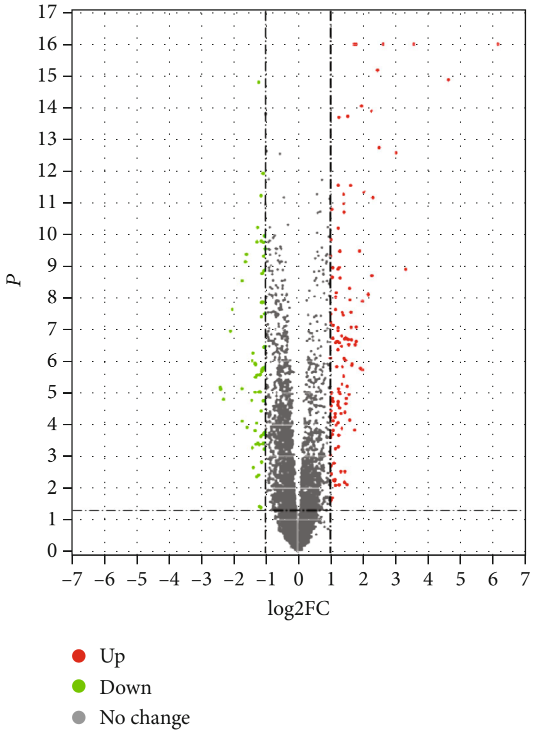

Figure 2: Volcano plot of GSE75010.

mRNA expression, red indicates relatively high mRNA expression, and gray indicated undifferentiated mRNA samples).

Figure 3 below shows the relative expression levels of the 48 differential genes screened by Limma in the preeclamptic placenta, denoted in the figure as PE, and the control healthy placenta, denoted in the figure as control (green indicates relatively low mRNA expression and red indicates relatively high mRNA expression).

After screening out 48 differential genes, CFS was used to filter mRNA to remove irrelevant mRNA and redundant mRNA. The optimal subset consisting of 13 mRNA attributes was obtained (HTRA4, PROCR, MYCN, ERO1A, EAF1, PPP1R16B, CRH, FLNB, PIK3CB, PLAAT3, FBN2, RFLNB, and TKT). The results, sorted from largest to smallest by the information gain ratio, were PIK3CB, HTRA4, ERO1A, PPP1R16B, PROCR, CRH, FLNB, PLAAT3, FBN2, EAF1, TKT, RFLNB, and MYCN. In the table, PE represents a sample of patients with PE, control represents a sample of healthy pregnant women as a control group, and average represents the average of the two sets of data. The classification results are derived from mRNA samples in the best subset training model (see Table 1).

Next, we used 48 differentially expressed mRNA samples that were not processed by CFS (the original differentially expressed mRNA samples) to train the model and test the classification effect. The results are as follows (see Table 2):

Table 2 shows that the likelihood of a sample being correctly classified as PE is 0.763, and the accuracy of the overall classification results of the model is 0.790.

By comparing the results from the two sets of experiments, it can be seen that the accuracy of the optimal subset of mRNA is 0.822, in which the correctly classified PE samples have increased from 61 to 63, and the correctly classified control samples have increased from 63 to 66.

However, it is not comprehensive to select the best biomarkers based on classification accuracy alone. We also use recall, precision, and AUC as classification criteria to obtain more comprehensive results.

Notably, recall, precision, and AUC reached the maximum values in Table 1; the specificity increased from 0.818 to 0.857, and the sensitivity increased from 0.763 to 0.788 (see Table 3), which can be considered as high specificity and sensitivity.

In the study of PE, the positive accuracy of the biomarkers discovered by Zeisler et al. was no more than 50% [36], and the positive accuracy of the mRNA PE biomarkers we screened for reached 0.788.

The experiments show that the use of CFS filtering attributes is also applicable to mRNA. After reducing the mRNA dimension, all indicators that have a positive significance for the classification effect are improved.

In the research of Mehmood et al., the voting integration method and the CFS algorithm were also used to achieve ideal results. Although similar algorithms on different datasets may have very different results, for this experiment, the CFS algorithm can exclude irrelevant differential genes, can avoid the redundancy of related mRNA samples, and can maintain the maximum independence among attributes, which is necessary for our research [37].

In this study, the use of the information gain ratio to analyze differential genes can allow for the measurement of the amount of information that mRNA carries for disease outcomes. A variety of classification algorithms are used to test the association between mRNA and PE, and we used placental mRNA to identify PE. Analyses and comparisons were also conducted.

We analyzed mRNA with the highest information gain ratio in the best subset through the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, and the results showed that the expression of mRNA (PIK3CB and OCRL) is related to the inositol phosphate metabolism and the phosphatidylinositol signaling systems. This may be related to the biochemical process of PE.

4. Conclusions

In recent years, in the classification and screening of genes, it has often only been possible to obtain a single result through differential gene expression analysis. Although this result might be related to PE, the diagnostic effect may not be good enough. However, with the development of machine learning, the use of feature engineering can better improve the classification efficiency, and the use of an appropriate classification model can intuitively reflect the classification effect and indirectly reflect the advantages and disadvantages of attribute selection. Therefore, after analyzing the differentially expressed genes, to improve the classification effect, we used a feature selection algorithm based on correlations as the standard for dimensionality reduction. To obtain better results, we designed a classification model based on a

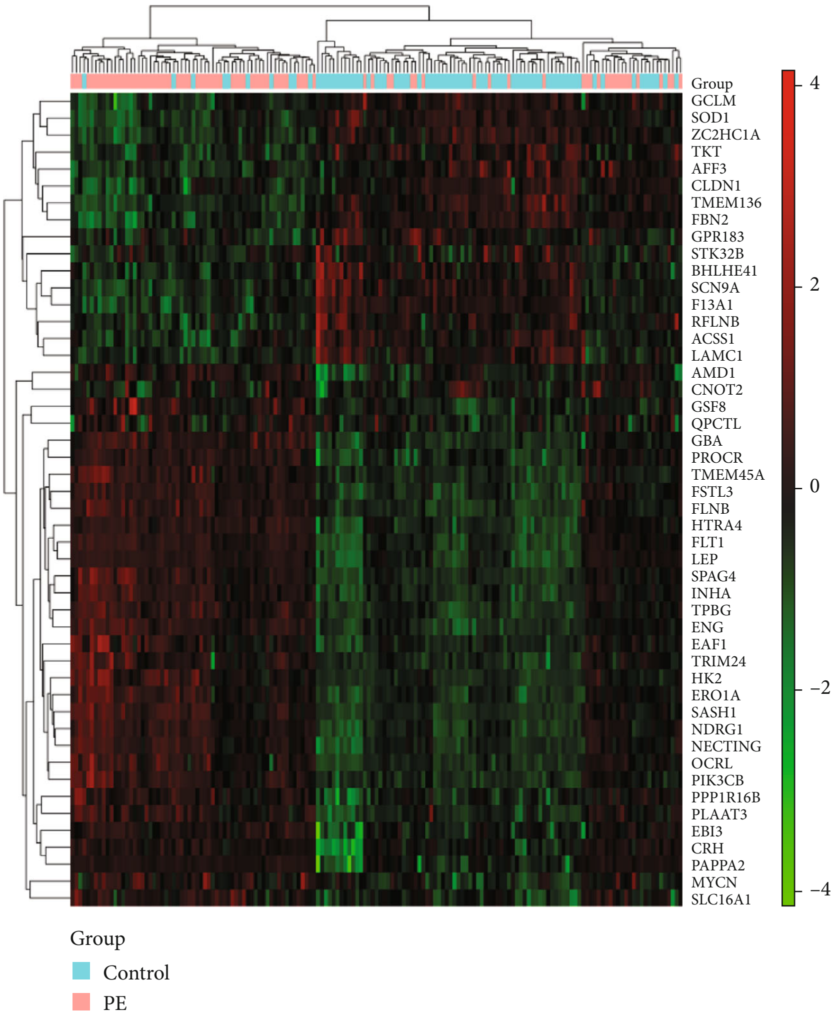

Figure 3: Heatmap of GSE75010.

Table 1: Results and indicators of attribute classification (the optimal subset).

	TP/TN	FP/FN	Precision	Recall	AUC
PE	0.788	0.143	0.851	0.788	0.822
Control	0.857	0.213	0.795	0.857	0.822
Average	0.822	0.177	0.824	0.822	0.822

Table 2: Results and indicators of attribute classification (the original differentially expressed mRNA samples).

	TP/TN	FP/FN	Precision	Recall	AUC
PE	0.763	0.182	0.813	0.763	0.790
Control	0.818	0.238	0.768	0.818	0.790
Average	0.790	0.209	0.791	0.790	0.790

voting mechanism to address the features of small sample datasets that are prone to overfitting. Finally, we used 13 mRNA samples as attributes to obtain satisfactory results. When training the model, we used 10-fold CV to enhance the robustness of the model.

Table 3: Specificity and sensitivity of the original differentially expressed mRNA samples (original) and optimal subset (optimal).

	Specificity	Sensitivity
Original	0.818	0.763
Optimal	0.857	0.788

The results show that an accuracy of 82.2% is achieved by the 13 mRNA samples screened out, and the specificity and sensitivity reach 0.857 and 0.788, respectively. The recall of the model is 0.822, the precision is 0.824, and the AUC value reaches 0.822. These indicators reflect that the model has good robustness and a certain generalization ability.

Through the KEGG analysis of the genes in Table 1, PIK3CB and OCRL were found to be involved in the phosphoinositide metabolism and the silanol phosphatidylinositol signaling systems. This can explain the cause of PE from one angle and provide clues for the future treatment of PE.

The optimal subset selected by the CFS algorithm is evaluated by the Pearson coefficient, but the Pearson coefficient cannot screen out important mRNA samples that are nonlinearly related to PE. How to select the optimal subset of mRNA samples that contain nonlinear correlations is still under discussion.

Data Availability

The PE dataset in this study can be downloaded from https://www.ncbi.nlm.nih.gov/geo/.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

Dan Liu and Zhixia Teng conceived and designed this project. Rong Guo analyzed the data and adjusted the models, Heze Xu collected the information, and Yiding Wang and Xin Zhou wrote the paper. Rong Guo and Dan Liu contributed equally to this work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (grant number 61901103).

References

[1] A. Filipek and E. Jurewicz, "Preeclampsia - a disease of pregnant women," *Postepy Biochemii*, vol. 64, no. 4, pp. 229–232, 2018.

[2] R. Landau and O. Irion, "Recent data on the physiopathology of preeclampsia and recommendations for treatment," *Revue Médicale Suisse*, vol. 1, no. 4, pp. 290–292, 2005.

[3] A. Bokslag, M. van Weissenbruch, B. W. Mol, and C. J. M. de Groot, "Preeclampsia: short and long-term consequences for mother and neonate," *Early Human Development*, vol. 102, pp. 47–50, 2016.

[4] S. Rana, E. Lemoine, J. P. Granger, and S. A. Karumanchi, "Preeclampsia: pathophysiology, challenges, and perspectives," *Circulation Research*, vol. 124, no. 7, pp. 1094–1112, 2019.

[5] P. J. B. PridjianG, "Preeclampsia.Part2: experimental and genetic considerations," *Obstetrical & Gynecological Survey*, vol. 57, no. 9, pp. 619–640, 2002.

[6] S. A. Founds, Y. P. Conley, J. F. Lyons Weiler, A. Jeyabalan, W. A. Hogge, and K. P. Conrad, "Altered global gene expression in first trimester placentas of women destined to develop preeclampsia," *Placenta*, vol. 30, no. 1, pp. 15–24, 2009.

[7] K. Leavey, S. J. Benton, D. Grynspan, S. A. Bainbridge, E. K. Morgen, and B. J. Cox, "Genemarkers of normal villous maturation and their expression in placentas with maturational pathology," *Placenta*, vol. 58, pp. 52–59, 2017.

[8] K. Leavey, S. J. Benton, D. Grynspan, J. C. Kingdom, S. A. Bainbridge, and B. J. Cox, "Unsupervised placental gene expression profiling identifies clinically relevant subclasses of human preeclampsia," *Hypertension*, vol. 68, no. 1, pp. 137–147, 2016.

[9] A. He, Y. Zhou, Y. Wei, and R. Li, "Potential protein biomarkers for preeclampsia," *Cureus*, vol. 12, no. 6, article e8925, 2020.

[10] B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, and P. Ping, "Machine learning and integrative analysis of biomedical big data," *Genes*, vol. 10, no. 2, p. 87, 2019.

[11] A. Sagar and B. Xue, "Recent advances in machine learning based prediction of RNA-protein interactions," *Protein and Peptide Letters*, vol. 26, no. 8, pp. 601–619, 2019.

[12] D. Goksuluk, G. Zararsiz, S. Korkmaz et al., "LSeq: machine learning interface for RNA-sequencing data," *Computer Methods and Programs in Biomedicine*, vol. 175, pp. 223–231, 2019.

[13] K. Leavey, S. L. Wilson, S. A. Bainbridge, W. P. Robinson, and B. J. Cox, "Epigenetic regulation of placental gene expression in transcriptional subtypes of preeclampsia," *Clinical Epigenetics*, vol. 10, no. 1, p. 28, 2018.

[14] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor, Statistics for Biology and Health*, pp. 397–420, Springer, New York, NY, USA, 2005.

[15] M. E. Ritchie, B. Phipson, D. Wu et al., "Limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, article e47, 2015.

[16] B. Ranjan, W. Sun, J. Park et al., *DUBStepR: correlation-based feature selection for clustering single-cell RNA sequencing data*, bioRxiv, 2020.

[17] J. Song, Y. D. Liu, J. Su, D. Yuan, F. Sun, and J. Zhu, "Systematic analysis of alternative splicing signature unveils prognostic predictor for kidney renal clear cell carcinoma," *Journal of Cellular Physiology*, vol. 234, no. 12, pp. 22753–22764, 2019.

[18] J. Dai and Q. Xu, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification," *Applied Soft Computing*, vol. 13, no. 1, pp. 211–221, 2013.

[19] A. Monaco, E. Pantaleo, N. Amoroso et al., "Identifying potential gene biomarkers for Parkinson's disease through an information entropy-based approach," *Physical Biology*, vol. 18, article 016003, 2021.

[20] C. Batur and B. Diri, "Sequence classification based on information gain feature groups using ensemble gene selection framework," in *8th International Advanced Technologies Symposium (IATS'17)*, Elazığ, Turkey, October 2017.

[21] F. Ghasemi, B. S. Neysiani, and N. Nematbakhsh, "Feature selection in pre-diagnosis heart coronary artery disease detection: a heuristic approach for feature selection based on information gain ratio and Gini index," in *2020 6th International Conference on Web Research (ICWR)*, Tehran, Iran, April 2020.

[22] Z.-H. Zhou and Y. Jiang, "Medical diagnosis with C4. 5 rule preceded by artificial neural network ensemble," *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, no. 1, pp. 37–42, 2003.

[23] J. Yun, J. E. Park, H. Lee, S. Ham, N. Kim, and H. S. Kim, "Radiomic features and multilayer perceptron network classifier: a robust MRI classification strategy for distinguishing glioblastoma from primary central nervous system lymphoma," *Scientific Reports*, vol. 9, no. 1, article 5746, pp. 1–10, 2019.

[24] M. S. Dawngliani, N. Chandrasekaran, R. Lalmawipuii, and H. Thangkhanhau, "Comparison of Decision Tree-Based Learning Algorithms Using Breast Cancer Data," in *Proceeding of the International Conference on Computer Networks,*

Big Data and IoT (ICCBI - 2019). ICCBI 2019. Lecture Notes on Data Engineering and Communications Technologies, vol 49, A. Pandian, R. Palanisamy, and K. Ntalianis, Eds., Springer, Cham, 2019.

[25] Q. Huang, Y. Chen, L. Liu, D. Tao, and X. Li, "On combining biclustering mining and AdaBoost for breast tumor classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 4, pp. 728–738, 2019.

[26] A. Maheshwari and P. Chakrabarti, "Machine learning classifier and neural modelling perspective of echocardiography for thalassemia patients in context to pediatric age group," *Journal of Critical Reviews*, vol. 7, no. 9, pp. 1144–1149, 2020.

[27] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, no. 2, article 022022, 2019.

[28] R. A. Peterson and J. E. Cavanaugh, "Ordered quantile normalization: a semiparametric transformation built for the cross-validation era," *Journal of Applied Statistics*, vol. 47, no. 13-15, pp. 2312–2327, 2020.

[29] X. Min, M. Li, D. Dong et al., "Multi-parametric MRI-based radiomics signature for discriminating between clinically significant and insignificant prostate cancer: cross-validation of a machine learning method," *European Journal of Radiology*, vol. 115, no. 2019, pp. 16–21, 2019.

[30] C. d'Abramo, L. D'Adamio, and L. Giliberto, "Significance of blood and cerebrospinal fluid biomarkers for Alzheimer's disease: sensitivity, specificity and potential for clinical use," *Journal of Personalized Medicine*, vol. 10, no. 3, p. 116, 2020.

[31] Y. Deng, X. Xu, Y. Qiu, J. Xia, W. Zhang, and S. Liu, "A multimodal deep learning framework for predicting drug-drug interaction events," *Bioinformatics*, vol. 36, no. 15, pp. 4316–4322, 2020.

[32] W. Zhang, Z. Li, W. Guo, W. Yang, and F. Huang, "A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1, 2019.

[33] W. Zhang, K. Jing, F. Huang et al., "SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions," *Information Sciences*, vol. 497, pp. 189–201, 2019.

[34] Y. Zhao, F. Wang, S. Chen, J. Wan, and G. Wang, "Methods of microRNA promoter prediction and transcription factor mediated regulatory network," *BioMed Research International*, vol. 2017, Article ID 7049406, 8 pages, 2017.

[35] L. Cheng, P. Wang, R. Tian et al., "LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse," *Nucleic Acids Research*, vol. 47, no. D1, pp. D140–D144, 2019.

[36] H. Zeisler, E. Llurba, F. Chantraine et al., "Predictive value of the sFlt-1:PlGF ratio in women with suspected preeclampsia," *The New England Journal of Medicine*, vol. 374, no. 1, pp. 13–22, 2016.

[37] Y. Mehmood, M. Ishtiaq, M. Tariq, and M. A. Jaffar, "Classifier ensemble optimization for gender classification using genetic algorithm," in *2010 International Conference on Information and Emerging Technologies*, pp. 1–5, Karachi, Pakistan, June 2010.