BiPoP: Bipolar Disorder Optimized Preprocessing Framework for Stress Disorder Identification through Gene Expression Data using Deep Learning

M. Sarala Shobini
Department of Information Technology, School of Computer Science Engineering and Information Systems (SCORE), Vellore Institute of Technology, India
saralashobini.m2020@vitstudent.ac.in

M. Sudha
Department of Information Technology, School of Computer Science Engineering and Information Systems (SCORE), Vellore Institute of Technology, India
msudha@vit.ac.in (corresponding author)

Received: 30 November 2024 | Revised: 22 January 2025 and 5 February 2025 | Accepted: 7 February 2025

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | DOI: https://doi.org/10.48084/etasr.9776

ABSTRACT

Gene expression data are widely used in diagnosing diseases and identifying promising genes with the advancement in computational tools in biology. Gene Expression Omnibus (GEO) datasets provide the gene expression data for various diseases and disorders. For Bipolar Disorder, GSE46449 was obtained from the NCBI data repository. This study aimed to classify control (Normal) and case (Disordered) individuals from samples using Machine Learning (ML)/Deep Learning (DL) models. The preprocessing involved the removal of null values and normalization of gene expression values using R. The second step focussed on the selection of optimal features/genes from the gene expression dataset. The Pearson Correlation Coefficient (PCC) along with Principal Component Analysis (PCA) were used for feature selection. The samples were then classified using ML/DL models. A Multi-Layer Perceptron (MLP) was used to validate the optimal feature set to classify healthy and disordered individuals. The proposed Bipolar Disorder Preprocessing Framework (BiPoP) was validated for its targeted use, highlighting its multifunctional and fine-tuned approach to preprocessing and achieving a classification accuracy of 98.9%.

Keywords-gene expression data; machine learning; feature selection; bipolar disorder; multilayer perceptron

I. INTRODUCTION

Bipolar disorder is one psychiatric disorder that affects 1% of the world's population. Typically, symptoms start with depressive episodes and prolong with maniac episodes in young adults, leading to cognitive and functional disability of the human brain. Diagnosing this disorder is challenging, as symptoms overlap with other disorders such as Major Depressive Disorder (MDD) [1], and it is frequently misdiagnosed as schizophrenia [2]. Studies have shown that it can be genetically inherited. The level of heritability is estimated and be between 68 to 80%. Genetic dysfunction is also possible in this type of stress disorder for genes such as the serotonin transporter gene and DAT1, which have strong associations and also exert polymorphism [3].

The inability to diagnose the disease early leads to its progression and difficulty in treatment [1]. Computational methods can be used to predict it accurately. Gene Expression (GE) profiles analysis aims to discover promising genes (disease-causing) from microarray data but suffers from the curse of dimensionality ($p >> n$), wherein a multitude of features (p) are observed from fewer samples (n) leading to the overfitting of Machine Learning (ML) models. Therefore, the large feature/gene set has to be reduced using feature selection or extraction techniques to retain more significant features and remove redundant and irrelevant ones. The reduced feature set is then fed into an ML-based model for classification and regression [4-5]. In [6], Logistic Regression (LR) and Convolutional Neural Networks (CNN) worked well for classification, achieving 90% accuracy in a tumor dataset. In [7], Ant Bee Colony (ABC) optimization was used to select the optimal features from leukemia microarray data, and a CNN model achieved 98% accuracy. Hybrid feature selection was carried out by implementing Maximum Relevance Minimum

Redundancy (MRMR), followed by a two-tailed unpaired t-test and metaheuristics for breast cancer detection.

In [8], the XGBoost classifier achieved 97% accuracy. Alzheimer's disease is one of the psychological disorders that occurs at an age above 50 years, and its symptoms include loss of memory [9]. Differential GE data were used to determine the genes responsible for the disease and biomarkers are identified using the Rhinoceros Search Algorithm (RSA). The samples were then classified using a Multilayer Perceptron (MLP) neural network that achieved the highest accuracy (>98%) [9]. In [10], a four-stage framework was proposed to predict Alzheimer's disease using integrated GE data. Preprocessing involved normalizing data and imputing missing attributes, followed by gene selection to select the optimal feature set to train various classifiers, with SVM outperforming the others. In [11], an ML-based framework was proposed using 6 different GEO datasets. In [12], differential GE analysis was performed to identify the most significant genes. Two feature selection techniques, namely Mean Decrease in Impurity (MDI) and Boruta, were used to identify important genes, and six ML algorithms were used to validate them. Random Forest (RF) worked well and predicted disease from control samples with good accuracy. Gene enrichment analysis was also performed using the identified genes.

In [13], intelligent hybrid ensemble gene selection was used to select informative genes from autism spectrum disorder microarray data. This can be implemented using two modules: data and functional perturbation. The selected genes were evaluated using a Deep Neural Network (DNN) with good accuracy. Parkinson's disease is a neurodegenerative disorder whose symptoms include rigidity, tremor, etc. In [14], gene expression data was considered to find valid genes using Lasso and ridge regression techniques. Models such as Support Vector Machines (SVM) and Logistic Regression (LR) along with Explainable AI (XAI), namely Shapely Additive Explanations (SHAP), were used to evaluate the selected genes. In [15], hybrid feature selection (filter-based ANOVA-F statistic and wrapper-based LASSO) was used on preprocessed GE data for Alzheimer's disease. The reduced gene set was processed using KNN, MLP, and SVM, with the latter outperforming the others in one dataset and MLP achieving the highest accuracy in the remaining four datasets [15]. In [5], feature selection was performed on 22 GE datasets based on integer coding and fuzzy granulation (SIFE). This technique worked well to identify key attributes with a smaller number of instances. Multiple lung carcinoma data were used for classification in [16], merging several datasets, where ML classifiers used DGE analysis to determine the best features. The common features/genes were then fed into Kaplan-Meyer survival analysis to predict their survivability, and their functions were also analyzed. In [17], hybrid feature selection based on an artificial immune optimization technique was used to select a subset of genes from 25 benchmarked GE datasets, improving the prediction accuracy of ML classifiers.

II. PROPOSED WORK

Bipolar disorder is a multifaceted mental health problem that impacts the emotional, cognitive, and behavioral aspects of life. It is marked by fluctuations between two opposing emotional states: the high energy of mania and the low despair of depression. These mood changes, which can be unpredictable, illustrate the complex interactions of neurological, genetic, and environmental influences. Therefore, medical diagnosis and treatment are essential. This study proposes a versatile Bipolar Disorder Preprocessing and Classification Framework (BiPoCare) that encompasses a novel preprocessing solution engine referred to as BiPop to improve the predictive performance of the classifier. Various models, such as LDA, SVM [18], LR [5], RF [19], NB [19, 20], BiLSTM, and MLP [21], can be trained using normalized preprocessed target data. The primary objective is to combine various feature selection methods to attain a top-tier method and select a specific subset of features. Then, the selected feature subset from the earlier phase was used to train and test DL algorithms. Figure 1 shows the proposed BiPoCare framework.

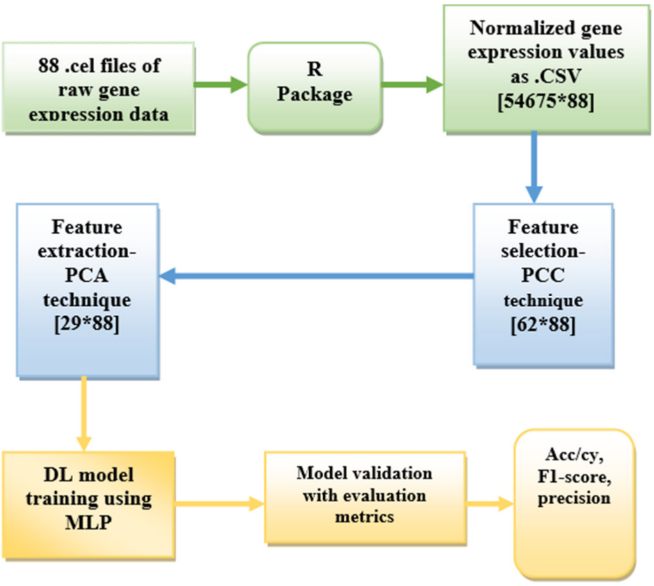
88 .cel files of raw gene expression data
R Package
Normalized gene expression values as .CSV [54675*88]
Feature selection- PCC technique [62*88]
Feature extraction- PCA technique [29*88]
DL model training using MLP
Model validation with evaluation metrics
Acc/cy, F1-score, precision

Fig. 1. Workflow of BiPoCare framework.

```
ALGORITHM: BiPoP framework
1. Start
2. Input: Raw dataset (D)
3. Phase 1: Preprocessing
3.1. Remove missing or null values from D.
3.2. Normalize the dataset to bring all
     features into a standard scale.
3.3. Encode categorical features (if any)
     into numerical format.
3.4. Split dataset into training and
     testing sets: ($D_{train}$, $D_{test}$).
4. Phase 2: Feature Selection
4.1. Calculate feature importance using a
     selection method:
     Method 1: Correlation matrix to
     identify highly correlated features –
     Pearson Correlation Coefficient[PCC]
     Method 2: Reduction of
     dimensionality using PCA
     Method 3: Model-based importance
4.2. Select top-N features (N can be
     predefined or based on a threshold).
5. Phase 3: Classification
5.1. Select a classifier (MLP)
5.2. Train the classifier on the selected
     features from $D_{train}$.
5.3. Validate the model on $D_{test}$.
5.4. Calculate performance metrics.
6. Output: Classification results and
   model performance metrics.
7. End
```

A. *Phase 1: Preprocessing*

1) *Dataset Description*

The dataset used to train and test the proposed framework was selected from GEO (Gene Expression Omnibus) from NCBI. GPL570 microarray was used to obtain gene expressions from blood. The accession number of the dataset is GSE46449 [22]. The dataset contains intensity values of probes that correspond to each gene that is expressed in this disorder, available in .cel files. Table I provides details on the dataset.

TABLE I. DATASET DESCRIPTION

Details	Source Information
Data Repository	Gene Expression Omnibus (GEO)-NCBI
Accession Number	GSE46449
Disease type	Bipolar Disorder
Platform	Affymetrix Human Genome U133 Plus 2.0 Array
Type of data	Gene Expression
Sample Count(n)	88
Feature Count(p)	54675
Data type	Numeric
Case (disease)	50 Samples
Control	38 Samples
Class	0-Control, 1-Case

2) *Preprocessing*

Preprocessing is the process of cleaning and analyzing the data. Data from the biological process contain errors and biases. Preprocessing is mandatory to perform data analysis and remove such errors. To handle outliers and missing values, normalization, standardization, and filtering [23] were performed to prepare the data for various ML/DL models. GE data provide information on the gene activity (up-regulated /down-regulated) of a particular cell. It is expressed as numerical values in continuous matrix form. These are .cel files that have to be preprocessed to convert them into an exact data format to select significant features. Data preprocessing was performed using the Robust Multi-Average (RMA) [14] technique, encompassing three steps: background correction, normalization, and summarization.

Background correction involves correcting the background signal by calculating the difference of an insignificant hybridization signal. Normalization adjusts the differences in the distribution of intensity values. Summarization calculates a unique expression value for each gene by averaging the intensities of probe values. All these processing steps were performed by invoking the RMA function defined in the affy package in R. Then normalized expression values were converted into a .csv file where features/genes were represented in rows and samples were represented in columns. This was input to the second stage to select significant features used for classification using ML/DL models. The data was split into training and validation sets with a ratio of 80:20.

B. *Phase 2: Feature Selection*

As microarray/GE data possess an intrinsic nature, with the curse of dimensionality, the number of features has to be reduced before training a model for better performance in terms of accuracy, sensitivity, specificity, etc. This reduction is achieved by feature extraction/selection methods, such as Filter, Wrapper, Hybrid, Ensemble, Embedded, and Principal Component Analysis (PCA). The Filter method uses statistical techniques to assign a score to each feature for its rank. Then, top-ranked genes are selected filtering out low-ranked, irrelevant genes. Some of the filtering approaches are Fast Correlation-Based Filter (FCBF), Pearson Correlation Coefficient (PCC), Chi-Square, Mutual Information, Relief F, mRMR, etc. Wrapper methods implement an optimization algorithm to choose the optimal feature set for a prediction model. The best subset of features is selected by assessing the model's performance. Algorithms such as Firefly, Ant colony optimization, SVM-RFE (Recursive Feature Elimination), binary bat algorithm, etc. are wrapped-based approaches. Figure 2 illustrates the complete architecture.

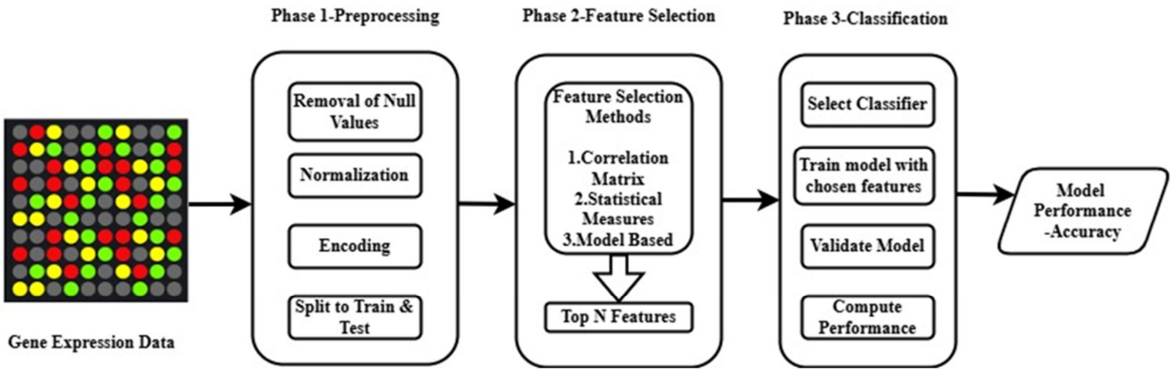

Fig. 2. Architecture of the BiPoCare framework.

C. Need for PCC and PCA as Feature Selection Approach

Feature extraction methods include PCA, where features are transformed from a high dimensional to a low dimensional space before feeding them into the model. Lack of interpretability is possible in PCA, as the transformation happens by combining the correlated features into new principal components. Since GE data depicts the characteristics of high dimensional data, where the number of features is much more compared to the samples, it is necessary to reduce the irrelevant and insignificant features and also the least correlated ones with the result. PCC is generally used for variable selection and its innate nature of associating the correlation of the input to the output variables aids in choosing the relevant and correlated features. Then the selected features are subjected to PCA to find the principal components from the selected features of PCC.

D. Classification

ML and DL play a pivotal role in the field of genomics by providing unique solutions to challenges in biomarker discovery, disease diagnosis, drug discovery, and precision medicine. Many algorithms are available for classification and regression. These algorithms are classified into three main categories: i) Supervised learning, ii) Unsupervised learning, and iii) Semi-supervised learning. Supervised learning works well if class labels are present within the data. Unsupervised learning is possible when labels are not available in the data. Semi-supervised is the best choice when less labeled and more unlabeled data are available [24]. DL has been proven the best method to find patterns in high-dimensional data, such as genomic data. Architectures such as ANN, CNN, RNN, GAN, and LSTM are used to achieve high accuracy. However, there is great demand for computing units such as GPUs or TPUs, which are high-performance hardware units to achieve parallelism [18]. In [19], leukemia classification was performed using a GE dataset with the help of increased hidden layers (5) in a DNN, achieving an accuracy of 98%. The features are selected by the multiple hidden layers used in the network. In [21], a novel feature selection technique was used to select the gene subset to predict breast cancer using ML models.

Table II details the parameters used in the model.

TABLE II. TRAINABLE PARAMETERS FOR MLP

Layer (Type)	Output shape	Param. #
Input 10 (input layer)	(None, 27)	0
Dense 36 (dense)	(None, 50)	1400
Dense 37 (dense)	(None, 50)	2550
Dense 38 (dense)	(None, 50)	2550
Dense 39 (dense)	(None, 2)	102
Total params		6602
Trainable params		6602

III. EXPERIMENTAL RESULTS

This study used the GEO GSE46449 dataset as mentioned in Table I. The dataset contains 88 samples in .cel files. R was used to perform background correction and normalization. After conversion, the data contained 54675 features/genes from 88 samples. PCC along with PCA [25] were employed to choose the dominant features from the whole dataset. Among 54675 features, 60 genes (including Differentially Expressed Genes-DEG) were selected. MLP [26] was used to classify healthy and disordered individuals. An accuracy of 98.9% was achieved in 60 epochs with Adam optimization. Relu and softmax were used as activation functions. Table III compares the accuracy of some previous algorithms with this study. Table IV details performance metrics such as precision, recall, and F1 score for the validation data. Figure 4 provides a comparison of the proposed framework with traditional methods.

TABLE III. PERFORMANCE METRICS

Algorithm	Accuracy
RF [27]	96
SVM [27]	94
BioPoP+MLP (Proposed)	98.9

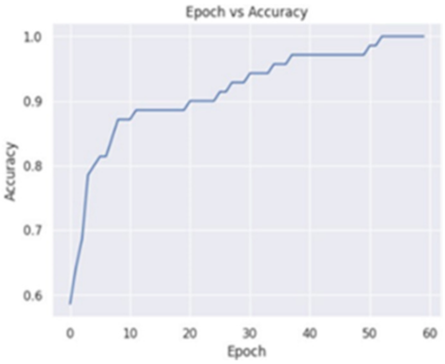

Fig. 3. Accuracy vs epochs.

TABLE IV. PERFORMANCE METRICS -VALIDATION DATA

	Precision	Recall	F1-score	Support
0	0.82	1.00	0.90	9
1	1.00	0.78	0.88	9

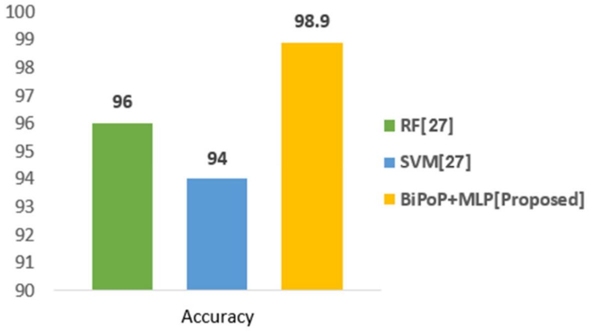

Fig. 4. Comparison of the proposed framework with other models.

IV. CONCLUSION

This study presented a novel BiPoP method for detecting bipolar disorder using microarray data. The proposed approach employed R for data preprocessing. Following this, feature selection was executed using PCC and PCA. From a total of 54,675 genes, 60 features were chosen. The MLP model was then trained on the selected gene set with specific parameters (epochs: 60, Optimizer: Adam). The performance metrics were documented, with an impressive accuracy of 98.9%. The proposed approach can also be applied to other psychological disorders in future studies.

REFERENCES

[1] T. Tekin Erguzel, C. Tas, and M. Cebi, "A wrapper-based approach for feature selection and classification of major depressive disorder–bipolar disorders," *Computers in Biology and Medicine*, vol. 64, pp. 127–137, Sep. 2015, https://doi.org/10.1016/j.compbiomed.2015.06.021.

[2] Z. Li, D. Li, and X. Chen, "Characterizing the polygenic overlaps of bipolar disorder subtypes with schizophrenia and major depressive disorder," *Journal of Affective Disorders*, vol. 309, pp. 242–251, Jul. 2022, https://doi.org/10.1016/j.jad.2022.04.097.

[3] D. Bassett, "Borderline personality disorder and bipolar affective disorder. Spectra or spectre? A review," *Australian & New Zealand Journal of Psychiatry*, vol. 46, no. 4, pp. 327–339, Apr. 2012, https://doi.org/10.1177/0004867411435289.

[4] D. Q. Zeebaree, H. Haron, and A. M. Abdulazeez, "Gene Selection and Classification of Microarray Data Using Convolutional Neural Network," in *2018 International Conference on Advanced Science and Engineering (ICOASE)*, Duhok, Iraq, Oct. 2018, pp. 145–150, https://doi.org/10.1109/ICOASE.2018.8548836.

[5] S. Osama, H. Shaban, and A. A. Ali, "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review," *Expert Systems with Applications*, vol. 213, Mar. 2023, Art. no. 118946, https://doi.org/10.1016/j.eswa.2022.118946.

[6] R. Tabares-Soto, S. Orozco-Arias, V. Romero-Cano, V. S. Bucheli, J. L. Rodríguez-Sotelo, and C. F. Jiménez-Varón, "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data," *PeerJ Computer Science*, vol. 6, Apr. 2020, Art. no. e270, https://doi.org/10.7717/peerj-cs.270.

[7] A. Wahid and M. T. Banday, "Classification of DNA microarray gene expression Leukaemia data through ABC and CNN method," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 7s, pp. 119–131, 2023.

[8] N. Alromema, A. H. Syed, and T. Khan, "A Hybrid Machine Learning Approach to Screen Optimal Predictors for the Classification of Primary Breast Tumors from Gene Expression Microarray Data," *Diagnostics*, vol. 13, no. 4, Jan. 2023, Art. no. 708, https://doi.org/10.3390/diagnostics13040708.

[9] K. Sekaran and M. Sudha, "Diagnostic Gene Biomarker Selection for Alzheimer's Classification using Machine Learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 2348–2352, Oct. 2019, https://doi.org/10.35940/ijitee.L3372.1081219.

[10] A. El-Gawady, M. A. Makhlouf, B. S. Tawfik, and H. Nassar, "Machine Learning Framework for the Prediction of Alzheimer's Disease Using Gene Expression Data Based on Efficient Gene Selection," *Symmetry*, vol. 14, no. 3, Mar. 2022, Art. no. 491, https://doi.org/10.3390/sym14030491.

[11] S. Koppad, A. Basava, K. Nash, G. V. Gkoutos, and A. Acharjee, "Machine Learning-Based Identification of Colon Cancer Candidate Diagnostics Genes," *Biology*, vol. 11, no. 3, Mar. 2022, Art. no. 365, https://doi.org/10.3390/biology11030365.

[12] N. B. Hiremath and P. Dayananda, "Differential Gene Expression Analysis of Non-Small Cell Lung Cancer Samples to Classify Candidate Genes," *Engineering, Technology & Applied Science Research*, vol. 13, no. 2, pp. 10571–10577, Apr. 2023, https://doi.org/10.48084/etasr.5770.

[13] G. Anurekha and P. Geetha, "An Intelligent Hybrid Ensemble Gene Selection Model for Autism Using DNN," *Intelligent Automation & Soft Computing*, vol. 35, no. 3, pp. 3049–3064, 2023, https://doi.org/10.32604/iasc.2023.029127.

[14] N. Bhandari, R. Walambe, K. Kotecha, and M. Kaliya, "Integrative gene expression analysis for the diagnosis of Parkinson's disease using machine learning and explainable AI," *Computers in Biology and Medicine*, vol. 163, Sep. 2023, Art. no. 107140, https://doi.org/10.1016/j.compbiomed.2023.107140.

[15] A. El-Gawady, B. S. Tawfik, and M. A. Makhlouf, "Hybrid Feature Selection Method for Predicting Alzheimer's Disease Using Gene Expression Data," *Computers, Materials & Continua*, vol. 74, no. 3, pp. 5559–5572, 2023, https://doi.org/10.32604/cmc.2023.034734.

[16] H. K. Joon, A. Thalor, and D. Gupta, "Machine learning analysis of lung squamous cell carcinoma gene expression datasets reveals novel prognostic signatures," *Computers in Biology and Medicine*, vol. 165, Oct. 2023, Art. no. 107430, https://doi.org/10.1016/j.compbiomed.2023.107430.

[17] Y. Zhu, T. Li, and W. Li, "An Efficient Hybrid Feature Selection Method Using the Artificial Immune Algorithm for High-Dimensional Data," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, 2022, Art. no. 1452301, https://doi.org/10.1155/2022/1452301.

[18] L. Koumakis, "Deep learning models in genomics; are we there yet?," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1466–1473, Jan. 2020, https://doi.org/10.1016/j.csbj.2020.06.017.

[19] P. K. Mallick, S. K. Mohapatra, G. S. Chae, and M. N. Mohanty, "Convergent learning–based model for leukemia classification from gene expression," *Personal and Ubiquitous Computing*, vol. 27, no. 3, pp. 1103–1110, Jun. 2023, https://doi.org/10.1007/s00779-020-01467-3.

[20] M. Arabfard, M. Ohadi, V. Rezaei Tabar, A. Delbari, and K. Kavousi, "Genome-wide prediction and prioritization of human aging genes by data fusion: a machine learning approach," *BMC Genomics*, vol. 20, no. 1, Nov. 2019, Art. no. 832, https://doi.org/10.1186/s12864-019-6140-0.

[21] S. Sasikala, S. A. alias Balamurugan, and S. Geetha, "A Novel Feature Selection Technique for Improved Survivability Diagnosis of Breast Cancer," *Procedia Computer Science*, vol. 50, pp. 16–23, Jan. 2015, https://doi.org/10.1016/j.procs.2015.04.005.

[22] C. L. Clelland, L. L. Read, L. J. Panek, R. H. Nadrich, C. Bancroft, and J. D. Clelland, "Utilization of Never-Medicated Bipolar Disorder Patients towards Development and Validation of a Peripheral Biomarker Profile," *PLOS ONE*, vol. 8, no. 6, 2013, Art. no. e69082, https://doi.org/10.1371/journal.pone.0069082.

[23] R. Zhou, S. K. Ng, J. J. Y. Sung, W. W. B. Goh, and S. H. Wong, "Data pre-processing for analyzing microbiome data – A mini review," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 4804–4815, Jan. 2023, https://doi.org/10.1016/j.csbj.2023.10.001.

[24] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, Jun. 2015, https://doi.org/10.1038/nrg3920.

[25] C. Jayaweera and N. Aziz, "Reliability of Principal Component Analysis and Pearson Correlation Coefficient, for Application in Artificial Neural Network Model Development, for Water Treatment Plants," *IOP Conference Series: Materials Science and Engineering*, vol. 458, no. 1, Sep. 2018, Art. no. 012076, https://doi.org/10.1088/1757-899X/458/1/012076.

[26] F. Alharbi and A. Vakanski, "Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review," *Bioengineering*, vol. 10, no. 2, Feb. 2023, Art. no. 173, https://doi.org/10.3390/bioengineering10020173.

[27] D. Kang *et al.*, "StressGenePred: a twin prediction model architecture for classifying the stress types of samples and discovering stress-related genes in arabidopsis," *BMC Genomics*, vol. 20, no. 11, Dec. 2019, Art. no. 949, https://doi.org/10.1186/s12864-019-6283-z.