A Machine Learning Framework for Cancer Prediction Using Gene Selection and Hyperband-Optimized XGBoost

P. ANILKUMAR,
Assistant Professor, Sir C.R.Reddy College of Engineering, Eluru, Andhra Pradesh, India

Abstract - Gene selection plays a significant role in the medical field, as it has strong potential for the early diagnosis of diseases. However, existing methods often suffer from limitations such as data imbalance and lower performance in feature selection. In this research, a Hyperband Optimization-based XGBoost classifier is proposed to improve classification performance in gene-based disease prediction. The NCBI gene dataset is used to evaluate the effectiveness of the proposed method for gene selection and disease diagnosis. Initially, a normalization technique is applied to scale the input data and minimize variations among the features. Subsequently, Principal Component Analysis (PCA) is employed to extract the most relevant features from the high-dimensional gene data, thereby reducing dimensionality, although it may reduce interpretability of the independent variables. The selected features are then used as input to the XGBoost classifier to perform disease classification. To further enhance the model performance, Hyperband optimization is utilized to efficiently search for the optimal hyperparameter settings of the XGBoost model. This method performs a distributed and resource-efficient search strategy that improves exploration of parameter configurations and enhances classifier performance. Experimental results show that the proposed XGB-PCA-HO model achieves an accuracy of 97.06%, which significantly outperforms the conventional XGBoost model (88.24%) and Random Forest model (85.29%). The results demonstrate that the proposed approach is effective for gene selection and early disease diagnosis.Keywords: Hyperband Optimization, NCBI gene dataset, Normalization, Principal Component Analysis, and XGBoost Classifier.

I. INTRODUCTION

An early and accurate prognosis of cancer facilitates the proper line of treatment, and DNA microarray technology has shown great potential in diagnosis of cancer and its classification. The cancer datasets produced by microarray technology typically have thousands of gene expressions obtained from each biological sample [1]. For the DNA microarray datasets, tumor classification based on gene expression profiles has drawn great attention, and gene selection plays a significant role in improving the classification performance of microarray data [2]. Gene selection as an important data preprocessing technique for cancer classification is one of the most challenging issues in the field of microarray data analysis. It aims to select the most representative gene subset with a high resolution by eliminating redundant and unimportant genes [3]. Gene expression data reduction involves two aspects: relevant and redundant. The relevancy between genes and the label information is measured with respect to the class labels, which is related to the importance of a gene for the classification task [4]. Gene selection techniques in general are grouped into three, namely: (1) Filter approach, (2) Wrapper approach and (3) Hybrid approach. The filter-based approaches select genes based on the general characteristics of the data while considering each gene separately. On the other hand, the wrapper approaches take into consideration the gene-to-gene dependencies and also use a classification model to evaluate the various gene subsets before selecting the most promising gene subset [5].

Machine learning has been an extremely powerful tool for biological data analysis. It has had several applications in various fields of bio-logical sciences mostly in two recent decades. Designing the prediction models is one of the most interesting applications of machine learning [6]. The common issues of high-dimensional gene expression data are that many of the genes may not be relevant, and there exists a high correlation among genes. Gene selection has been proven to be an effective way to improve the results of many classification methods [7]. Many existing methods were involving in applying the gene selection for disease classification. Existing methods have limitations of imbalance dataset, small sample size, and overfitting problem [8, 9, 10].

This paper is organized as literature review in section 2, the proposed method explanation is given in section 3, results is given in section 4 and conclusion is given in section 5.

II. LITERATURE REVIEW

Recently, the gene selection method is used for cancer classification for early diagnosis and this is one of active research topic. Many researches were carried out for the

cancer classification based on gene selection and some of the notable methods were discussed in this section.

Huang, *et al*. [11] integrated Cancer Linker Degree (CLD), weighted Domain Frequency Score (DFS), Domain-Domain Interaction (DDI), Protein-Protein Interaction (PPI) data were integrated for gene classification for cancer prediction. Three types of process were carried out like individual methods, combined methods and combination of same types of methods for prediction. The developed machine learning with voting method has higher performance compared to existing method in prediction. The weighted DFS method adaptively measure the propensity of domain occurrence in non-cancer and cancer proteins. The feature selection method performance is low and this creates overfitting problem in the machine learning method.

Azzawi, *et al*. [12] applied Gene Expression Programming (GEP) based model to predict the lung cancer from micro-array data. Two gene selection method is proposed to extract significant gene of lung cancer and proposes different prediction models based on gene selection method. Three machine learning based models such as Radial Basis Function neural network, Multi-layer perceptron, and support vector machine were applied for prediction. The developed method has higher performance than existing methods in prediction of gene. The developed method has lower performance in missing dataset and affects the performance of prediction.

Wu, *et al*. [13] proposed L_1 logistic regression model for the gene prediction in high dimensional cancer classification for estimation of gene coefficient. The L_1 based gene selection is performed and doesn't have oracle property. The DNA based microarray dataset was used to test the performance of the L_1 logistic regression method in prediction of gene. The gene selection method is applied for cancer classification to overcome the overfitting problem, small sample size, and high-dimensional data. The datasets such as GSE10072, GSE19084, and GSE4115 were used to test the performance of developed method. The developed method has higher performance compared to existing methods in gene selection. The developed method has lower performance in imbalance dataset in prediction.

Li and Liu, [14] applied regularized logistic regression method for gene selection in gene expressed data for disease classification. The seven penalty was applied in regularized logistic regression method to effectively select the gene for classification. The multiple datasets were used to test the performance of developed method in classification. The SCAD, $L_{1/2}$, lasso and elastic net were applied as penalty to improve the performance of the classification. A functional enrichment analysis is carried out on gene selection and developed logistic regression to improve the performance of the model. The developed method has higher performance compared to existing method in gene selection. The developed method has limitation of overfitting problem in disease classification.

Shukla, *et al*. [15] proposed hybrid wrapper method that combines Gravitational Search Algorithm (GSA) and Teaching Learning Based Optimization (TLBO) for gene classification. A new encoding method is applied for continuous search space. The feature selection of minimum redundancy maximum relevance was used for feature selection in gene expression dataset. The gravitational search method is applied in teaching phase to improve search capability in evolution process. The Naïve Bayes model is used for fitness function to select the gene in cancer classification. The biological datasets were used to test the performance of the developed method and compared with existing methods. The developed method has lower convergence and easily trap into local optima.

III. PROPOSED METHOD

In this research, the hyperband optimization method in XGBoost classifier to improve the performance of classifier in gene selection. The NCBI dataset of genes expression were applied to test the performance of classification. Normalization method is applied to scale the data and reduce the difference in the data instance. The PCA method selects the relevant features for gene classification and independence variable become less interpretable. The block diagram of the proposed XGB-PCA-HO method is shown in Figure 1.

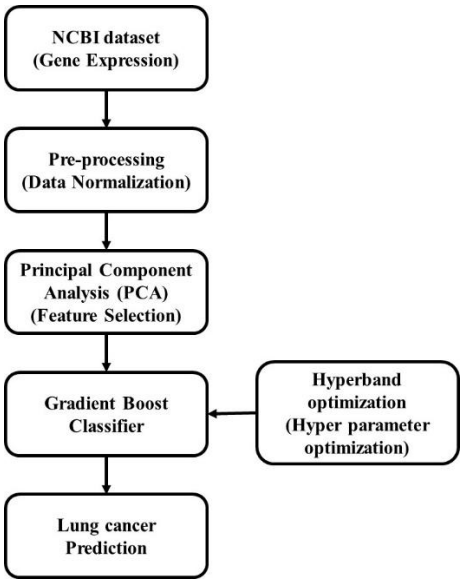

Figure 1. The proposed XGB-PCA-HO method in gene classification

a. Principal Component Analysis (PCA)

The PCA is a dimension reduction technique for data [16 - 18]. Since it is simple and easy to understand and does not have limitations of parameters, the PCA has been widely applied in all kinds of fields. The main idea of the PCA is to map n-dimensional features to k-dimensional features ($k \leq n$). The k-dimensional features are new orthogonal features, called principal components, which are reconstructed from the original n-dimensional features. The essence of the PCA is to reduce the redundancy of data under the premise of losing information as little as possible, so as to achieve the purpose of dimension reduction.

The steps of the PCA are described in detail as follows:

Step 1: Calculate the sample mean of the n-dimensional data set X, where $X = \{x_1, x_2, \dots, x_m\}$.

$$\alpha = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1}$$

where m is total number of samples, $i = 1, \dots, m$, α is the obtained sample mean.

Step 2: Use the generated sample mean to calculate the covariance matrix of the sample set.

$$C = \frac{1}{m}\sum_{i=1}^{m}(x_i - \alpha)(x_i - \alpha)^T \tag{2}$$

where C is covariance matrix of the sample set.

Step 3: Calculate the feature values and feature vectors of the sample covariance matrix.

$$C = Q.\sum.Q^T \tag{3}$$

$$\sum = diag(\lambda_1, \lambda_2, \dots, \lambda_n)\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_2 \geq 0 \tag{4}$$

$$Q = [q_1, q_2, \dots, q_n] \tag{5}$$

where $\sum$ is the arranged diagonal matrix of n feature values of the covariance matrix in descending order, λ_i is the corresponding feature values of covariance matrix, and Q is the feature matrix composed of the corresponding feature vector qi of the feature value λ_i , $i = 1, \dots, n$.

Step 4: Use the obtained feature values and feature vectors to calculate the cumulative variance contribution rate of the first k-row principal elements.

$$\theta = \sum_{i=1}^{k}\lambda_i \,/\, \sum_{j=1}^{m}\lambda_j \tag{6}$$

where θ is cumulative variance contribution rate of the former k-row principal elements, and the value of θ is usually greater than or equal to 0.9. In theory, the value of θ should be as large as possible. From a practical point of the view, the value of θ should be reasonably selected according to the specific solving problem. When the value of θ is reasonably selected, the information of the summarized original sample set of the k-row principal elements can be determined.

Step 5: Realize the dimension reduction using the obtained k-row feature vector.

$$P = Q_k \tag{7}$$

$$Y = P.X \tag{8}$$

where P is a feature matrix, which is composed of corresponding feature vectors of the first k-row feature values ($k \leq n$). Q_k is a feature matrix, which is composed of the first k-row feature values ($k \leq n$). And Y is the k-dimensional data. The transformation of data set X to Y also realizes the linear transformation of data from n-dimension to k-dimension in order to achieve dimension reduction.

b. Hyperband optimization

The essential idea of Hyperband is to allocate more resources to more promising hyperparameter configurations. First, it initializes a set of n trial points (each trial point corresponding to one hyperparameter configuration). Then, it uniformly allocates a budget to each trial point, and evaluates its performance (i.e., objective function) given that budget. The algorithm for hyperband optimization is given below.

Algorithm: Hyper parameter optimization

Input: Single hyper-parameter configuration R, and proportion controller η

Output: one hyper-parameter configuration

Initialization: $s_{max} = \lfloor \log_\eta(R) \rfloor$, $B = (s_{max} + 1)R$

For $s \in \{s_{max}, s_{max} - 1, \dots, 0\}$ do

$n = \left\lceil \frac{B}{R}\left(\frac{\eta^s}{s+1}\right) \right\rceil$, $r = R\eta^{-s}$

$X = get\_hyperparameter\_configuration(n)$

for $i \in 0, \dots, s$ do

$n_i = \lfloor n\eta^{-i} \rfloor$

$n_i = \lfloor n\eta^{-i} \rfloor$

$r_i = r\eta^i$

$F = \{run\_then\_return\_obj\_val(x, r_i) : x \in X\}$

$X = top\_k(X, F, \lfloor n_i/\eta \rfloor)$

return configuration with the best objective function value

c. XGBoost Algorithm

The XGBoost method is machine learning method and this consists of weak predictors sequence [19, 20]. This method is based on the gradient boosting method. Gradient boosting is iterative tree estimation, residuals obtained at each step and adaptive estimates updates. Gradient descent technique is used in Gradient boosting method and method splits the favors to reduce the point of objective function.

The XGBoost optimization is compared with gradient boosting due to regularization to avoid bias and overfitting, missing values management, tree pruning operations, parallel and distribution computing use, and its scalability.

The variables x_i is a set of values in input data and predict the variable y_i, as given in equation (9).

$$\{(x_i - y_i)\}_{i=1}^{n} \tag{9}$$

This consists of training dataset, the model predict the variable value y_i based on variable x_i to characterize multiple features. The predicted value is $\widehat{y_i} = \sum_j \theta_j x_{ij}$ is used in a linear regression problem, where weight of x_j is denoted as θ_j. The model parameters is denoted as θ in a generic problem.

The objective function measures model ability to fit training data that consists of two terms, as given in equation (10).

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{10}$$

Where regularization term is denoted as $\Omega(\theta)$ and training loss function is denoted as $L(\theta)$. The prediction is evaluated using differentiable function of loss function. The regularization term helps to control model complexity and avoid overfitting.

The loss function of Taylor expansion is used in XGBoost to design objective function, as given in equation (11).

$$Obj(\theta) = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{11}$$

Where $g_i = \partial_{\widehat{y_i}^{t-1}} L(y_i, \hat{y}_i^{t-1})$, while $h_i = \partial^2_{\hat{y}^{t-1}} L(y_i, \hat{y}_i^{t-1})$. The following quantities are defined, as given in equation (12) to (14).

$$G_j = \sum_{i \in I_j} g_i \tag{12}$$

$$H_j = \sum_{i \in I_j} h_i \tag{13}$$

$$I_j = \{i | q(x_i) = j\} \tag{14}$$

The j-th leaf optimal weight value is denoted as $\theta_j = -\left(\frac{G_j}{H_j + \lambda}\right)$ that returns the leaf index itself. The j-th leaf instance set is denoted as I_j and mapping function of data instance into tree leaf. The model optimizes based on objective function is given in equation (15).

$$Obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{15}$$

Where characterize of tree leaves is denoted as T.

The algorithm computation cost is due to all tree training in simultaneous. The split candidate evaluate based on gain function, is given as in equation (16).

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G_P^2}{H_P + \lambda} \tag{16}$$

Where left nodes (subscript L) is contributed based on first term, the right nodes (subscript R) is contributed based on second term, the parent leaf node (subscript P) is contributed by the last term. The greatest gain of split condition is selected and pruning method is used to optimize a tree level to reduce overfitting.

IV. RESULTS

In this research, the hyperband optimization is applied in XGBoost classifier to improves its classification performance in gene selection. The NCBI genes such as GSE10072, GSE19084, and GSE4115 were used to test the performance of the developed method in gene selection. The PCA method is applied to select the relevant features from input dataset and apply to XGBoost classifier. The XGBoost classifier parameter is optimized based on the Hyperband optimization method.

Table 1. Metrics of proposed method

	TN	TP	FN	FP
Logistic Regression	1	11	9	13
Naive Bayes	1	13	7	13
K-Means	1	15	5	13
SVM	7	19	1	7
Neural Network	9	18	2	5
Random forest	12	17	3	2
XGB-GS	13	17	2	2
XGB-PCA	13	18	2	1
XGB-PCA- HO	13	20	1	0

The existing and proposed methods of TN, TP, FN and FP for gene classification are measured and shown in Table 1. This shows that the developed method has higher performance compared to existing method in gene classification. The Random Forest has considerable performance than SVM and Neural Network.

Table 2. Performance analysis of proposed method

Methods	Accuracy	Sensitivity	Specificity	Precision	Recall	F-measure
Logistic Regression	35.29	55.00	7.14	45.83	55.00	50.00
Naive Bayes	41.18	65.00	7.14	50.00	65.00	56.52
K-Means	47.06	75.00	7.14	53.57	75.00	62.50
SVM	76.47	95.00	50.00	73.08	95.00	82.61
Neural Network	79.41	90.00	64.29	78.26	90.00	83.72
Random forest	85.29	85.00	85.71	89.47	85.00	87.18
XGB-GS	88.24	89.47	86.67	89.47	89.47	89.47
XGB-PCA	91.18	90.00	92.86	94.74	90.00	92.31
XGB-PCA- HO	97.06	95.24	100.00	100.00	95.24	97.56

The proposed XGB-PCA- HO method and existing methods performance were measured in gene classification, as compared in Table 2. The proposed XGB-PCA- HO method has higher performance in classification compared to existing methods in terms of Accuracy, sensitivity, specificity, precision, recall and F-measure. The hyperband optimization method selects the relevant parameter settings for XGBoost method to improve the classification performance. The search process of hyperband optimization is distribution in equal manner to improve the exploration of the process. The hyperparameter setting helps to improve the performance of the developed method in gene classification. The SVM method has the limitation of imbalance data problem in the classification. The Random forest method has limitation of overfitting when number of tree is less and instable performance when number of tree is high.

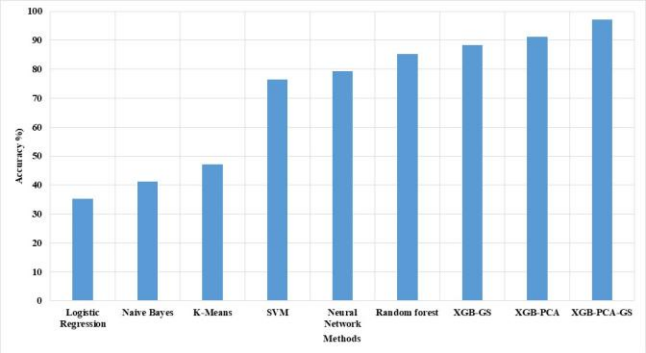

Figure 2. Accuracy of the proposed XGB-PCA- HO method

The accuracy of the proposed XGB-PCA- HO method and existing methods in gene classification, as shown in Figure 2. The proposed XGB-PCA- HO method has advantage of hyperparameter optimization based on hyperband optimization. The hyperband optimization method performs the search process in equal distributed manner that helps to improve the exploration process. The PCA method reduces the overfitting process in the training and independent variables are less interpretable. The proposed XGB-PCA- HO method has accuracy of 97.06 % and existing random forest has 85.29 % accuracy.

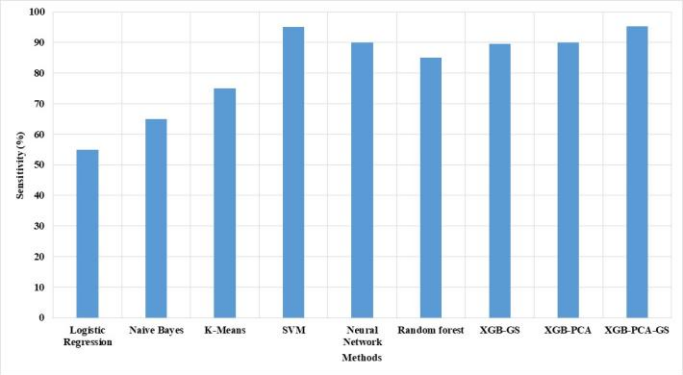

Figure 3. Sensitivity of proposed XGB-PCA-GS method

The proposed XGB-PCA-HO method and existing method sensitivity were measured, as shown in Figure 3. The sensitivity is important metrics due to its measures the classification performance related to class. The proposed method has higher sensitivity due to its selection of features and parameter setting in classification. The hyperband optimization method performs the search process in distributed manner to improve the exploration of the

parameter search process. The proposed XGB-PCA-HO method has the sensitivity of 95.24 % and existing random forest has 85.71 % sensitivity.

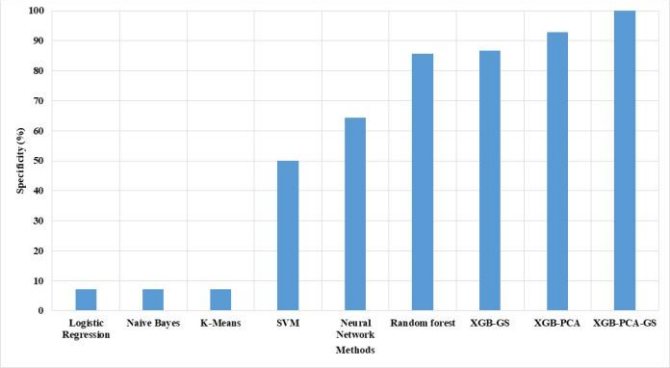

Figure 4. Specificity of proposed XGB-PCA-HO method

The specificity of the proposed XGB-PCA-HO method and existing methods were compared in Figure 4. The specificity of the proposed method is high compared to existing methods in gene selection. The proposed method has advantage of select the hyperparameter setting based on hyperband optimization. The proposed XGB-PCA-HO method has specificity of 99.99 % and existing random forest has 85.71 % specificity.

V. CONCLUSION

Gene selection method has potential to diagnosis the diseases in early stage and this is challenging task. Existing methods have limitation of imbalance data problem and lower efficiency in feature selection. In this research, the hyperband optimization method is proposed in XGBoost classifier to improve the performance of classification. The hyperband optimization method performs search in distributed manner to improve the exploration process for parameter settings. This helps to select the parameter for the classifier to avoid overfitting and overcome the imbalance problem. The proposed XGB-PCA-HO method has accuracy of 97.06 %, and random forest has 85.29 % accuracy. The future work of the proposed method involves in applying LSTM based method in large gene dataset.

VI. REFERENCES

[1] Jain, I., Jain, V.K. and Jain, R., 2018. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Applied Soft Computing, 62, pp.203-215.

[2] Sun, L., Kong, X., Xu, J., Zhai, R. and Zhang, S., 2019. A hybrid gene selection method based on ReliefF and ant colony optimization algorithm for tumor classification. Scientific reports, 9(1), pp.1-14.

[3] Sun, L., Zhang, X.Y., Qian, Y.H., Xu, J.C., Zhang, S.G. and Tian, Y., 2019. Joint neighborhood entropy-based gene selection method with fisher score for tumor classification. Applied Intelligence, 49(4), pp.1245-1259.

[4] Huang, X., Zhang, L., Wang, B., Li, F. and Zhang, Z., 2018. Feature clustering based support vector machine recursive feature elimination for gene selection. Applied Intelligence, 48(3), pp.594-607.

[5] Rani, M.J. and Devaraj, D., 2019. Two-stage hybrid gene selection using mutual information and genetic algorithm for cancer data classification. Journal of medical systems, 43(8), pp.1-11.

[6] Dashtban, M., Balafar, M. and Suravajhala, P., 2018. Gene selection for tumor classification using a novel bio-inspired multi-objective approach. Genomics, 110(1), pp.10-17.

[7] Algamal, Z.Y. and Lee, M.H., 2019. A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. Advances in data analysis and classification, 13(3), pp.753-771.

[8] Hamamoto, R., Komatsu, M., Takasawa, K., Asada, K. and Kaneko, S., 2020. Epigenetics analysis and integrated analysis of multiomics data, including epigenetic data, using artificial intelligence in the era of precision medicine. Biomolecules, 10(1), p.62.

[9] Pes, B., 2021. Learning from High-Dimensional and Class-Imbalanced Datasets Using Random Forests. Information, 12(8), p.286.

[10] Bilen, M., Işik, A.H. and Yiğit, T., 2020. A New Hybrid and Ensemble Gene Selection Approach with an Enhanced Genetic Algorithm for Classification of Microarray Gene Expression Values on Leukemia Cancer. International Journal of Computational Intelligence Systems, 13(1), pp.1554-1566.

[11] Huang, C.H., Peng, H.S. and Ng, K.L., 2015. Prediction of cancer proteins by integrating protein interaction, domain frequency, and domain interaction data using machine learning algorithms. BioMed research international, 2015.

[12] Azzawi, H., Hou, J., Xiang, Y. and Alanni, R., 2016. Lung cancer prediction from microarray data by gene expression programming. IET systems biology, 10(5), pp.168-178.

[13] Wu, S., Jiang, H., Shen, H. and Yang, Z., 2018. Gene selection in cancer classification using sparse logistic regression with L1/2 regularization. Applied Sciences, 8(9), p.1569.

[14] Li, L. and Liu, Z.P., 2020. Biomarker discovery for predicting spontaneous preterm birth from gene

expression data by regularized logistic regression. Computational and Structural Biotechnology Journal, 18, pp.3434-3446.

[15] Shukla, A.K., Singh, P. and Vardhan, M., 2020. Gene selection for cancer types classification using novel hybrid metaheuristics approach. Swarm and Evolutionary Computation, 54, p.100661.

[16] Chen, Y., Tao, J., Zhang, Q., Yang, K., Chen, X., Xiong, J., Xia, R. and Xie, J., 2020. Saliency detection via the improved hierarchical principal component analysis method. Wireless communications and mobile computing, 2020.

[17] Zhao, H., Zheng, J., Xu, J. and Deng, W., 2019. Fault diagnosis method based on principal component analysis and broad learning system. IEEE Access, 7, pp.99263-99272.

[18] Nobre, J. and Neves, R.F., 2019. Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. Expert Systems with Applications, 125, pp.181-194.

[19] Duan, J., Asteris, P.G., Nguyen, H., Bui, X.N. and Moayedi, H., 2021. A novel artificial intelligence technique to predict compressive strength of recycled aggregate concrete using ICA-XGBoost model. Engineering with Computers, 37(4), pp.3329-3346.

[20] Qiu, Y., Zhou, J., Khandelwal, M., Yang, H., Yang, P. and Li, C., 2021. Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. Engineering with Computers, pp.1-18.