ENHANCED APPROACH OF HEART DISEASE DIAGNOSIS USING MODIFIED FEATURE EXTRACTION METHOD

[1]Mr. H. Ramprasanth, [2]Dr. N. Kamalraj

[1]Research Scholar, Department of Computer Science, Park's College (Autonomous), Tirupur. ramprasanth1408@gmail.com.

[2]Research Supervisor, Associate Professor and Vice Principal, Park's College (Autonomous), Tirupur. tpkamal@gmail.com.

DOI: 10.63001/tbs.2025.v20.i03.S.I(3).pp1146-1151

KEYWORDS
Heart Disease, Heart Disease, Feature Extraction, PCA, Chi-squire, Relief, Principal Components Analysis.

Received on:

04-08-2025

Accepted on:

06-09-2025

Published on:

04-10-2025

ABSTRACT

Heart diseases are becoming one of the leading causes of death in the arena. Consequently, the medical community has shown a great deal of interest in cardiac disease prediction. In order to assist doctors in the design of clinical procedures, a number of studies have created machine learning algorithms for the early prediction of coronary heart diseases. The function set that is chosen has a significant impact on how well such structures perform. This will be more difficult if the schooling dataset has missing values for the various capacities. It is well known that Principal Component Analysis (PCA) can be used to address the issue of missing attribute data. This study offers a method for identifying heart disease by using scientific testing data as input, identifying coronary heart disease by extracting a low dimensional characteristic subset. The suggested approach uses Modified Principal Component Analysis (M-PCA) to extract improved depth characteristics from fresh projections. PCA reduces the size of the function by assisting in the extraction of projection vectors that significantly contribute to the maximum covariance. Three datasets are analyzed to determine the impact, accuracy, sensitivity, and specificity of the suggested approach. The results obtained from the use of the suggested M-PCA technique are compared to earlier studies in order to demonstrate its relevance. The dataset generated by the suggested M-PCA technique was incredibly accurate.

INTRODUCTION

The heart works like an engine or motor and is responsible for blood circulation throughout the body. Diagnosing heart disease is seen as a crucial activity that needs to be finished accurately and promptly. Early detection, which is made possible by a number of medical tests, a thorough medical record, and the patient's daily routine, can help monitor the high number. Until a healthcare expert is present to analyze the data, it is unlikely that data alone will be adequate.

Finding the optimal answer to the issues of precise diagnosis and therapeutic delivery is the primary objective of a physical evaluation. However, there is still a significant research gap when it comes to analyzing a lot of data to forecast cardiac disease. There haven't been enough contributions to incorporate or extract features to establish the class labels because the processed data has a well defined feature set to try to define the heart disease.

Similarly, machine learning would be quite helpful. Heart disease prediction is the world's most challenging undertaking. To extract features, the majority of academics have focused on data or signal gathering. Even so, the processed data's designated features are not very good at differentiating between the class labels. Supervised learning is a prerequisite for this unique function. Furthermore, supervised learning relies on the efficacy of the training method, which necessitates adequate infrastructure to handle the nonlinearities of both extracted features. Therefore, a better heart disease prediction system is needed, and people would gain from its precise prediction process.

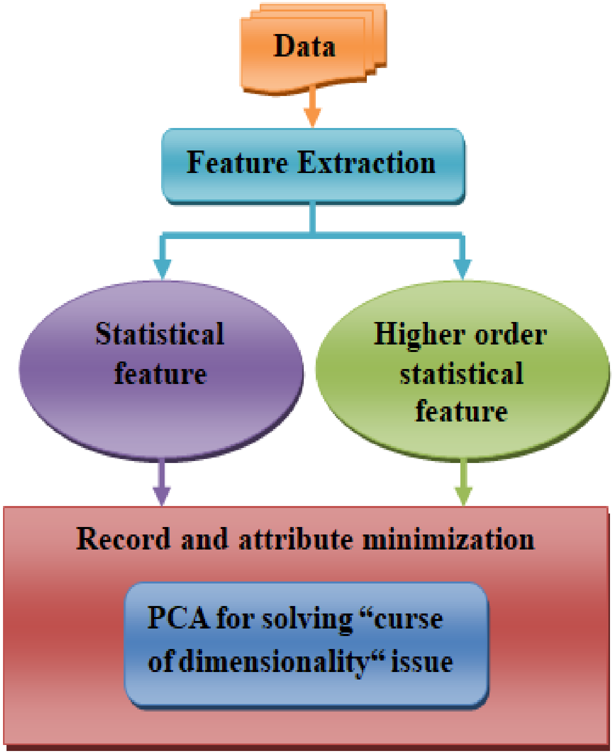

Data
Feature Extraction
Statistical feature
Higher order statistical feature
Record and attribute minimization
PCA for solving "curse of dimensionality" issue

Figure 1: - Feature Extraction

One way to start analyzing high-dimensional data that is easy to understand by just looking at the vast quantity of information is to use Principal Component Analysis (PCA), a type of design recognition approach. Prior to charting and understanding the data, the increasing component of the information must be lowered to a low measurement for information analysis. The score plot and stacking plot are two basic charts that employ PCA to include significant data. Breaking down a lot of information in the research area of investigation is difficult. The association between the highly correlated informative indexes is ascertained using the PCA calculation. With variables as sections and perceptions or tests as lines, PCA's mathematical analysis in linear algebra provides a clear explanation of the link between the data. Reducing the incredibly high number of connections between different factors to a tolerable level is the aim of the PCA computation. Principal components are the terms used to describe these interrelated elements. The basic idea is to build a network with the most important information from the first two sections, and then use Python programming to arrange the data into a 2-dimensional plot.

The development of an effective prediction mechanism is necessary to lower the risk of heart disease. This research suggests a Modified Principal Component Analysis (M-PCA)-based feature reduction method. The suggested method presents a compelling argument for dimensionality reduction.

This article's remaining content is organized as follows. The many ways the researchers approached the issue statement are examined in Section 2. Section 3 examines current approaches that offer different approaches or solutions for CD prediction. The system method utilized in the suggested plan to address the problem statement is described in Section 4. The performance of the suggested M-PCA approach is contrasted with that of current techniques in Section 4. A summary of the suggested work is provided at the end of Section 5.

2. RELATED WORK

The performance characteristics and accessibility of heart disease are affected by a number of feature extraction approaches. Research on data mining algorithms and coronary disease risk prediction has been published; it is challenging to assess whether algorithm performs better than others because each one has unique advantages and disadvantages [5]. The suggested study focuses on medical diagnosis, but additional work is needed to determine the method.

Table 1: Literature Review and Study

Study	Proposed	Techniques and Tools	Findings
Amin, M. S., et al., (2019)	To use and created a model for predicting heart disease.	The vote algorithm is a hybrid of logistic regression and NB.	The feature selection algorithm does not exist.
Saqlain, S. M., et al., (2019)	The feature selection algorithms are proposed for a cardiac disease diagnosis system.	SVM for binary classification.	The feature selection algorithm requires improvement.
Purnomo, A., et al., (2020)	Various FS & NB.	NB + (Backward Elimination / Optimization / Forward Selection).	The classifier algorithm must be improved.
Vivekanandan, T., & Iyengar, N. C. S. N. (2017)	The effectiveness of DE set of rules for prediction of heart disease.	AHP and artificial neural network (ANN).	The classification accuracy must be improved.

Several adaptive methods for identifying cardiovascular disease with modified principal component analysis have been developed, according to a review of the literature. The transformed matrix used in PCA is likewise derived from the entire image since the covariance matrix is usually calculated using the complete data set. Every potential cover class in the field of study is considered, including those that have no bearing on a particular application. A Modified Principal Component Analysis (M-PCA) is suggested in this study, where data from specific classes are the only ones used to construct the transformed matrix.

3 THE PROPOSED MODEL

The search space and information storage will be expanded due to the duplicated and inconsistent statistics in the pre-processed dataset. In order to achieve classification accuracy, we must eliminate all redundant and superfluous records. To reduce dimensional records, superfluous dimensional records are compressed using the dimensionality reduction approach, subject to certain limitations. The most pertinent feature, the important function, is extracted using characteristic extraction using primary factor analysis.

Modified M-PCA, also known as Principal Component Analysis, is a variation of the PCA dimensionality-reduction technique. It is frequently used to reduce the dimensionality of large data sets by reducing a large number of variables to a smaller one while maintaining the majority of the records in the large set.

Accuracy is sacrificed for reduced dimensionality; nonetheless, the dimensionality discount approach involves sacrificing some precision in favor of simplicity. Because there are fewer variables to process, smaller data sets are easier to study and illustrate, and they also make data analysis much easier and faster for gadget mastering algorithms.

Suppose we have a random vector population A, where

$$A = (a1, a2, \dots an)^P \qquad (equ.1)$$

And the mean of that population is denoted by,

$$\mu_a = P(A) \qquad (equ.2)$$

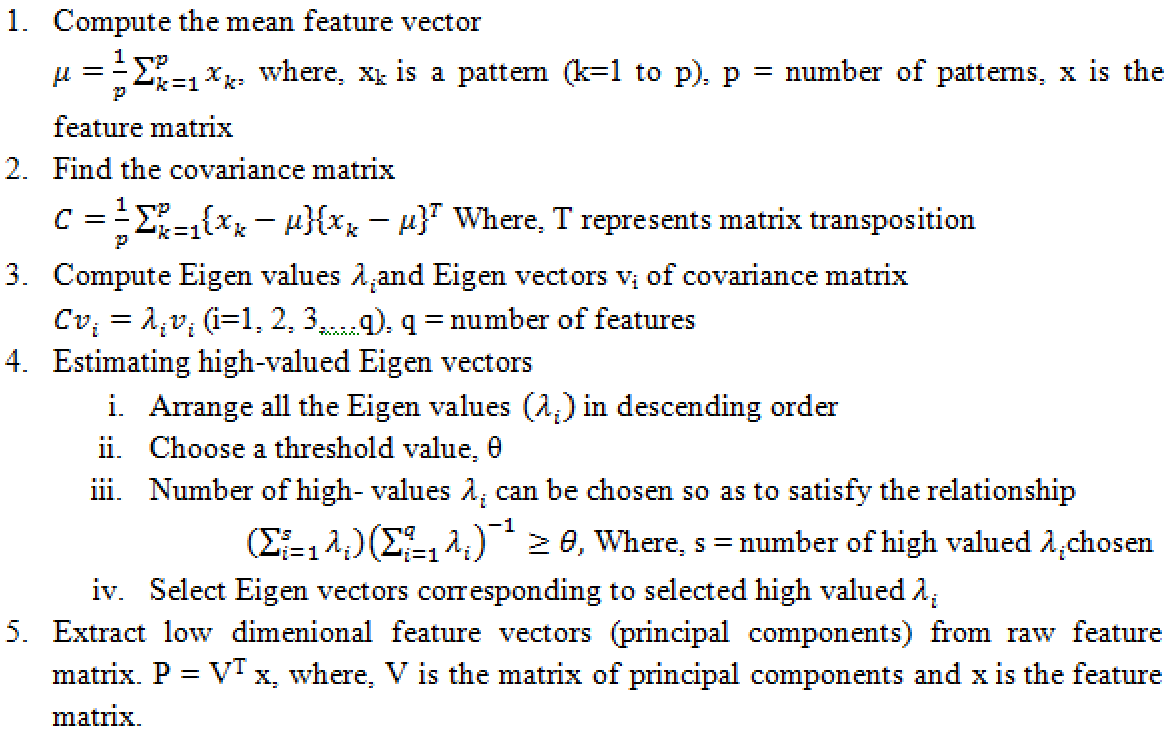
And the covariance matrix of the same data set is

$$Z_A = P\{(A - \mu_a)(A - \mu_a)^P\} \qquad (equ.3)$$

By identifying the eigenvalues and eigenvectors of a symmetric covariance matrix, it is possible to degree an orthogonal foundation from it. By arranging the eigenvectors in descending order of eigenvalues (biggest first), an ordered orthogonal basis can be produced. The first eigenvector will have the records' path of finest variance.

This enables us to determine which instructions have the greatest significant parts of power in the statistics set. A statistical set of 500 information sets is used to identify patterns in coronary heart disease. When diagnosing coronary heart disease, a number of factors are taken into account. Nonetheless, there are 15 qualities that stand out as being crucial.

Algorithm



```
1. Compute the mean feature vector
   $\mu = \frac{1}{p}\sum_{k=1}^{p} x_k$, where, $x_k$ is a pattern (k=1 to p), p = number of patterns, x is the feature matrix
2. Find the covariance matrix
   $C = \frac{1}{p}\sum_{k=1}^{p}\{x_k - \mu\}\{x_k - \mu\}^T$ Where, T represents matrix transposition
3. Compute Eigen values $\lambda_i$and Eigen vectors $v_i$ of covariance matrix
   $Cv_i = \lambda_i v_i$ (i=1, 2, 3,.....q), q = number of features
4. Estimating high-valued Eigen vectors
   i. Arrange all the Eigen values $(\lambda_i)$ in descending order
   ii. Choose a threshold value, $\theta$
   iii. Number of high- values $\lambda_i$ can be chosen so as to satisfy the relationship
        $(\sum_{i=1}^{s}\lambda_i)(\sum_{i=1}^{q}\lambda_i)^{-1} \geq \theta$, Where, s = number of high valued $\lambda_i$chosen
   iv. Select Eigen vectors corresponding to selected high valued $\lambda_i$
5. Extract low dimenional feature vectors (principal components) from raw feature matrix. $P = V^T x$, where, V is the matrix of principal components and x is the feature matrix.
```

4. EXPERIMENTAL SETUP

In this stage, the suggested method is assessed using publicly available datasets, and its overall performance is then contrasted with that of other recent methods. Additionally, the suggested method will be contrasted with methods that implement M-PCA and those that do not.

4.1. Performance Evaluation

i) Accuracy

An Intel Xeon processor with two 2.20-GHz cores and eight gigabytes of random-access memory is used for all calculations. Additionally, the studies are conducted using the scikit-study Python programming language package.

Table 2: Comparison table of Accuracy

Dataset	Chi square	ReliefF	Proposed M-PCA
100	73	65	79
200	89	85	93
300	85	81	91
400	82	76	89
500	72	69	80

The accuracy of the proposed and present techniques is stated in table 2. It shows that the general overall performance of the proposed technique is better than the present methods.

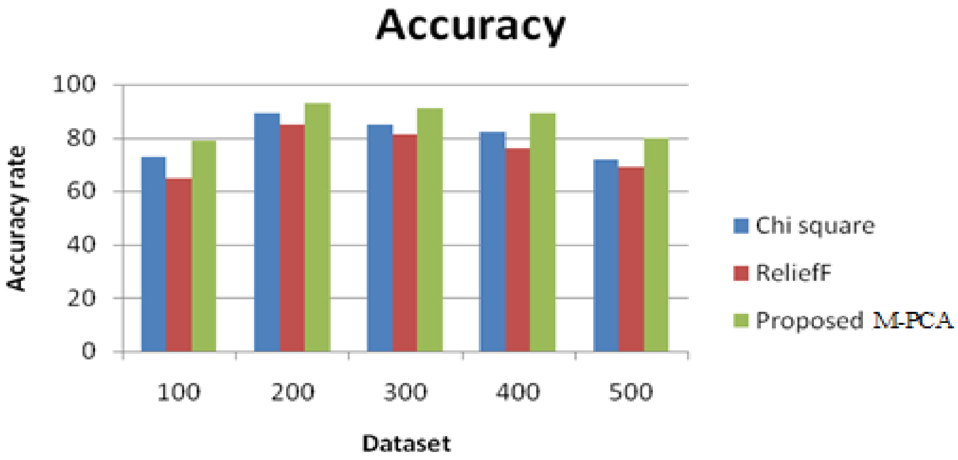


Figure 2: Comparison chart of Accuracy

An accuracy comparison chart comparing the suggested (M-PCA) and current (Chi square, ReliefF) methods is shown in Figure 2. The suggested M-PCA settings perform better than the current method. For the suggested M-PCA values, the existing algorithm values range from 73 to 72, 65 to 69, and 79 to 80, indicating a 5% increase in performance.

ii) Sensitivity

Table 3: Comparison table of Sensitivity

Dataset	Chi square	ReliefF	Proposed M-PCA
100	83.40	81.23	86.82
200	84.74	83.52	88.74
300	88.21	84.01	90.55
400	90.48	87.35	92.46
500	93.51	90.60	96.91

The sensitivity values represent the ability of a test to correctly identify those with the disease. The above table 3 shows the overall performance of the existing and proposed approach.

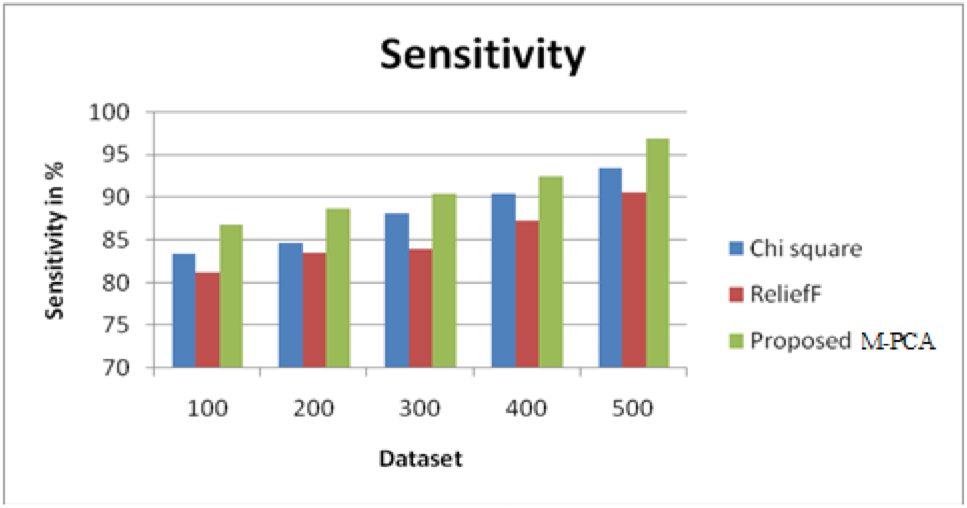


Figure 3 Comparison chart of Sensitivity

A sensitivity comparison chart showing the percentage of true positives that the model correctly predicts is shown in Figure 3. When compared to the current methods, the suggested M-PCA values perform better. The performance of the suggested strategy will be 5% better than that of the current ReliefF approach and 2% better than that of the current chi-square.

Specificity

Table 4: Comparison table of Specificity

Dataset	Chi square	ReliefF	Proposed M-PCA
100	78.48	80.22	85.87
200	80.74	82.52	88.74
300	82.21	84.01	90.55
400	85.48	87.35	92.46
500	88.66	91.65	97.10

The Specificity values represent the capability of the take a look at to correctly discover the ones without the sickness. Table 4 analyses the performance of the proposed and existing procedures.

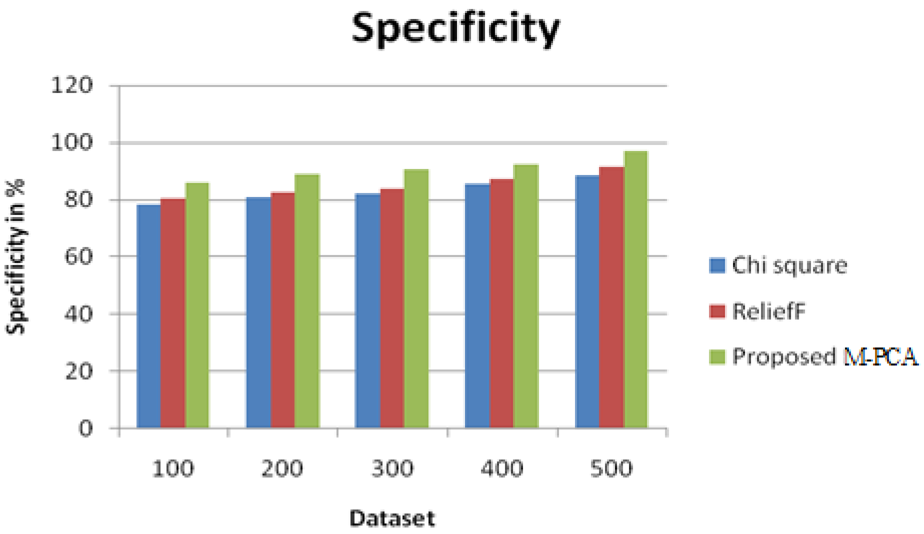


Figure 4: Comparison chart of Specificity

The percentage of true negatives that the model accurately predicts is seen in Figure 4. The performance of the suggested (M-PCA) and current (Chi square, ReliefF) models is depicted in the comparison chart above. The suggested M-PCA values vary from 85.87 to 97.10, while the values of the current algorithm fall between 78.48 and 88.66, 80.22 and 91.65. It demonstrates that the suggested method produces 5% greater specificity than the current method.

CONCLUSION

This work presents research methodologies for coronary heart disease analysis that use principle element analysis (PCA) to generate a feature subset as the first step. Parallel analysis completes the selection of the key components. The Cleveland, Hungarian, and Swiss UCI datasets are used. For Cleveland, Hungary, and Switzerland, the suggested method based entirely on modified major thing evaluation (M-PCA) generated characteristic subsets with dimensions lowered by 70%, 62%, and 70%, respectively. The application of M-PCA for extraction results in individuals with coronary heart disease and frequent issue training for suspected cases of heart disease. Metrics for assessment include specificity, sensitivity, and accuracy. The suggested M-PCA methodology performed well on all three measures when compared to the current method. Experiment outcomes also are offered, and their facts extended our confidence inside the proposed approach.

REFERENCES

- [1] Chugh, A. (2018). ML: chi-square test for feature selection.
- [2] Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53(1), 23-69.
- [3] Kavitha, R., & Kannan, E. (2016). An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. In 2016 international conference on emerging trends in engineering, technology and science (icetets) (pp. 1-5). IEEE.
- [4] Thomas, G. S., Budhkar, S. S., Cheulkar, S. K., Choudhary, A. B., & Rohan, S. (2015). Heart disease diagnosis system using apriori algorithm. International Journal of Advanced Research in Computer Science and Software Engineering, 5(2), 430-432.
- [5] Wang, Z. H., Wang, C. M., & Jong, G. J. (2018). Feature Extraction of the VSD Heart Disease based on Audicor Device Measurement. In 2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII) (pp. 138-141). IEEE.
- [6] Gárate-Escamila, A. K., El Hassani, A. H., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19, 100330.
- [7] Ghongade, R. (2007). A brief performance evaluation of ECG feature extraction techniques for artificial neural network based classification. In TENCON 2007-2007 IEEE Region 10 Conference (pp. 1-4). IEEE.
- [8] Sun, S. (2021). Segmentation-based adaptive feature extraction combined with Mahalanobis distance classification criterion for heart sound diagnostic system. IEEE Sensors Journal, 21(9), 11009-11022.
- [9] Gupta, A., Arora, H. S., Kumar, R., & Raman, B. (2021). DMHZ: a decision support system based on machine computational design for heart disease diagnosis using z-alizadeh sani dataset. In 2021 International Conference on Information Networking (ICOIN) (pp. 818-823). IEEE.
- [10] Putra, L. S. A., Isnanto, R. R., Triwiyatno, A., & Gunawan, V. A. (2018). Identification of Heart Disease With Iridology Using Backpropagation Neural Network. In 2018 2nd Borneo International Conference on Applied Mathematics and Engineering (BICAME) (pp. 138-142). IEEE.
- [11] Sun, S., Wang, H., Cheng, C., Chang, Z., & Huang, D. (2017). PCA-based heart sound feature generation for a ventricular septal defect discrimination. In 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) (pp. 128-133). IEEE.
- [12] Suseendran, G., Zaman, N., Thyagaraj, M., & Bathla, R. K. (2019). Heart Disease Prediction and Analysis using PCO, LBP and Neural Networks. In 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 457-460). IEEE.
- [13] Kumar, P. R., Ravichandran, S., & Narayana, S. (2021). Parametric Analysis on Heart Disease Prediction using Ensemble based Classification. In 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-13). IEEE.

- [14] Sonawane, R., & Patil, H. D. (2022). Prediction of Heart Disease by Optimized Distance and Density-Based Clustering. In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 1001-1008). IEEE.
- [15] Ambesange, S., Vijayalaxmi, A., Sridevi, S., & Yashoda, B. S. (2020). Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques. In 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4) (pp. 827-832). IEEE.
- [16] Chandra, R., Kapil, M., & Sharma, A. (2021). Comparative Analysis of Machine Learning Techniques with Principal Component Analysis on Kidney and Heart Disease. In 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1965-1973). IEEE.
- [17] Shah, S. M. S., Batool, S., Khan, I., Ashraf, M. U., Abbas, S. H., & Hussain, S. A. (2017). Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. Physica A: Statistical Mechanics and its Applications, 482, 796-807.
- [18] Rehman, A., Khan, A., Ali, M. A., Khan, M. U., Khan, S. U., & Ali, L. (2020). Performance analysis of pca, sparse pca, kernel pca and incremental pca algorithms for heart failure prediction. In 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE) (pp. 1-5). IEEE.
- [19] Ziasabounchi, N., & Askerzade, I. N. (2014). A comparative study of heart disease prediction based on principal component analysis and clustering methods. Turkish Journal of Mathematics and Computer Science (TJMCS), 16, 18.