Current Applied Science and Technology Vol. 23 No. 3 (May-June 2023)

Research article

Predicting Heart Disease Using FTGM-PCA Based Informative Entropy Based-Random Forest

Deepika Deenathayalan[1]* and Balaji Narayanan[2]

[1]*Department of Artificial Intelligence and Data Science, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India*
[2]*Department of Computer Science and Engineering, Velammal Institute of Technology, Chennai, India*

Received: 23 February 2022, Revised: 17 July 2022, Accepted: 23 November 2022

DOI: 10.55003/cast.2022.03.23.011

Abstract

Keywords

heart disease prediction;

FTGM-PCA;

informative entropy based-random forest;

dimensionality reduction;

cleveland heart disease dataset

In recent years, heart disease has become a reason for high mortality rate, and data mining has also gained attention in the medical domain. Predicting this disease in its initial stage helps to save lives and reduce treatment costs. Various classification models were recently introduced with expected outcomes. However, they lacked prediction accuracy. Hence, the aim of this study was to employ data mining techniques for predicting heart disease, and focused on higher accuracy. This disease was predicted by considering the Cleveland heart disease dataset, employing deep CNN models for extracting relevant features, and performing feature level fusion related to its efficient and automatic learning. FGM-PCA (Fast Track Gram Matrix-Principal Component Analysis) was proposed for dimensionality reduction and fusion to solve overfitting issues, minimise time and space complexity, eliminate redundant data, and enhance classifier performance. Further, effective classification was achieved through the newly introduced IEB-RF (Informative Entropy Based-Random Forest) because it offers high accuracy and can also handle a large amount of data flexibly. The proposed system was evaluated in terms of accuracy, sensitivity, F1-score, AUC (Area under Curve) and precision. The results revealed the superior performance of the introduced system in comparison to traditional techniques.

*Corresponding author: Tel.: (+91) 9647533289
E-mail: deepikaphd11@gmail.com

1. Introduction

According to the reports of WHO (World Health Organisation), heart disease is a major contributor to the high death rate worldwide. It is vital to predict this disease at its initial stage to afford suitable treatments in a timely manner and reduce the mortality rate [1, 2]. Recently, data mining methods were utilized to solve various issues in managing and analyzing particular data in healthcare centres [3]. Various researchers employed diverse data mining techniques to predict heart disease with better accuracy. Several ML (Machine Learning) algorithms were used including LR (Logistic Regression), SVM (Support Vector Machine), DT (Decision Tree), RF (Random Forest) and K-NN (K-Nearest Neighbour). In addition, DL (Deep Learning) was used to compare these methods, with the UCI-ML dataset that was comprised of 14 main features. The results showed the efficiency of DL with 94.2% accuracy rate [4]. Suitable feature selection has a significant role in improving classification accuracy. Dimensionality reduction also assists in enhancing the overall accuracy of prediction [5]. Employing classification methods on the disease datasets produces better outcomes as automated, intelligent and adaptive systems are used to diagnose chronic diseases like heart disease [6].

Among the classifier systems, various traditional works were used to predict heart disease through SVM, ANNs (Artificial Neural Networks), DTs and NB (Naïve Bayes). Special focus was given to parallel and classification systems as they improved success and minimised the time for decision making. However, further work must be carried out to design hybrid classification techniques for enhancing the accuracy in classification, and to perform optimisation of computational efficacy. Correspondingly, a hybridisation method was suggested whereby ANN and DT classifiers were employed to predict heart disease in a better way through WEKA software. Analysis was undertaken in terms of sensitivity, specificity, and accuracy. Accuracy was found to be 78.14%. Yet, it still had to be further enhanced [7]. Accordingly, four datasets related to heart disease were assessed through PCA (Principal Component Analysis), ReliefF, symmetrical uncertainty and chi-squared testing for creating distinct feature sets.

Subsequently, various classification methods were utilised for creating models which were later compared to find ideal feature combinations. The best model created used an integration of the chi-squared method with BayesNet. The accuracy rate was found to be 85% [8]. To accomplish a better prediction accuracy for heart disease, a floating window with adaptive size was used. Classification was performed through deep neural network and ANN. The results verified that the introduced models performed better than 18 other traditional methods that obtained accuracies within the range -50.00 to 91.83%. The ANN-based system showed 91.1% classification accuracy whereas the DNN based system showed 93.3% accuracy rate [9, 10]. In addition, NB+PSO (Particle Swarm Optimisation) was employed and compared with a traditional hybrid NB+GA (Genetic Algorithm). The suggested NB+PSO showed better accuracy at a rate of 87.91% whereas the NB+GA showed 86.29% [11].

Though previous researchers attempted to predict heart disease, their efforts were deficient in the selection of relevant features and dimensionality reduction, which ultimately affected the accuracy rate. Hence, an efficient method is needed to better predict heart disease through data mining techniques. In this study, we proposed the use of FTGM-PCA (Fast Track Gram Matrix-Principal Component Analysis) for dimensionality reduction. FTGM can be used to calculate linear-independence, and PCA can enhance ML model performance as it removes the correlated variables that do not support decision making. Furthermore, it also assists in resolving overfitting problems by minimising the features. This leads to high variance, and therefore enhanced visualisation. IEB-RF (Informative Entropy Based-Random Forest) is proposed for classification purposes. IE is associated with data compression as well as transmission, which constructs upon probability and assists ML. Further, RF affords high accuracy by cross-validation, handles missing values, maintains the rate of accuracy, and avoids overfitting issues. In this study, we propose hybrid IEB-RF because of the above-mentioned advantages that make the prediction of heart disease more effective and verifiable. The major contributions of this study are listed below.

- To extract features and carry out feature level fusion using the proposed deep CNN (deep Convolutional Neural Network) models for selecting only the relevant features. This process assisted in the attainment of minimized non-linearity and regularization operation.
- To perform faster dimensionality reduction and fusion through the introduced FTGM-PCA (Fast Track Gram Matrix-Principal Component Analysis), thus improving the performance of classifiers.
- To classify the normal and suspected patient data using the proposed IEB-RF (Informative Entropy Based-Random Forest) for accomplishing a better prediction rate.
- To analyze the efficiency of the proposed system through a comparative analysis with respect to accuracy, sensitivity, precision, F1-score and AUC (Area under Curve).

Various data mining techniques used by different conventional researchers to predict heart disease are presented and analysed. The significant and common problems identified during these analyses are also explored.

For feature extraction and selection techniques to predict heart disease, selecting appropriate and relevant features were vital to enhance the performance of suggested prediction models. Little research focused on the significant features and used different methods for feature extraction and selection to predict heart disease. However, diverse features were developed along with classification methods such as DT, NB, SVM, Vote (LR+NB), NN (Neural Network) and K-NN. Empirical outcomes revealed that "cp", "exang", "FBS", "sex", "old peak", "restecg", "thal", "ca" and "slope" were the important features for predicting heart disease. Vote was found to be an effective data mining method for predicting of heart disease with 87.4% accuracy rate [12]. Similarly, a system for diagnosing heart disease was recommended where the feature subsets had been extracted that encompass of high variance. PCA was used to select projection vectors by PA (Parallel Analysis) and SVM was used for classifying the suspected and normal data. Evaluation was undertaken via 3 performance metrics such as specificity, sensitivity, and accuracy, using 3 UCI datasets (Hungarian, Switzerland and Cleveland). The accuracy rate was found to be 85.82% for the Hungarian dataset, 82.18% for the Cleveland dataset and 91.30% for the Switzerland dataset [13, 14]. Accuracy needed to be further improved.

Feature selection has proved to be an efficient method to improve accuracy by removing noisy information and enabling clear understanding. Hence, heart disease features were weighted and then re-ordered in accordance with the weights and ranks assigned through the recommended ILFS (Infinite Latent Feature Selection) technique. Soft-margin linear SVM was employed for classifying the selected features into various heart disease classes. Experimentations were performed using a public heart disease dataset. The results revealed the efficacy of ILFS and soft-margin linear SVM with an accuracy rate of 90.65% for predicting the disease [15]. Likewise, CTDL (Cluster-based DT Learning) was used for feature selection, and RF for classifying the Cleveland dataset with an accuracy of 89.30% [16]. To further improve classification accuracy, heart disease was predicted based on ideal feature selection technique through Auto-Encoder (AE). Hybrid MLP (Multi-Layer Perceptron) and SVM were used and the accuracy rate was n found to be 91.97% [17, 18]. In addition, hybrid PSO (Particle Swarm Optimisation) and SVM were suggested to select significant features. This method identified 6 features that were significant for heart disease classification: sex, maximum heartbeat rate, major vessel counts, blood-sugar level, resting electro-cardio graphic outcomes. These features were fed into the SVM for classification and the accuracy was found to be 79.35% without feature selection. However, the classifier showed an accuracy rate of 84.36% after feature selection. Thus the significance of feature selection as well as extraction in improving the performance of the trained model in predicting the disease was revealed [19, 20].

For dimensionality reduction methods for predicting heart disease, clinical data can possess features that are hard to visualise and understand**.** Many features can also result in more memory and high computation time. Hence, using dimensionality reduction techniques are significant to develop a diagnosis system. Choosing the suitable approaches for reducing dimensionality can enhance the accuracy of a system [21]. An MLPSNM (Multi-Layer Pi-Sigma Neuron Model) was recommended to diagnose heart disease. LDA (Linear Discriminant Analysis) and PCA were utilised for dimensionality reduction. This model accomplished 94.53% as accuracy rate for classification [22]. Likewise, SOM (Self-Organising Map), Fuzzy SVM and PCA were used to impute the missing values. Fuzzy SVM and PCA were specifically utilised for incremental data learning to minimise the computation time in predicting the disease. Analysis was carried out on two real-world datasets such as Statlog and Cleveland. The outcomes showed that using fuzzy SVM could enhance classification accuracy with 0.96 for the Cleveland dataset and 0.97 for the Statlog dataset [23, 24]. Hence, dimensionality reduction helped to enhance accuracy rate. Correspondingly, dimensionality reduction was compared with respect to feature selection and extraction. LDA and PCA were used to perform feature extraction. On the other hand, DT and GA (Genetic Algorithm) were used for feature selection. GA was employed to perform two tasks encompassing attribute selector and rule selector. GA filtered efficient and relevant features, and was an attribute selector that assisted in classifying data with better accuracy. Outcomes showed the efficiency of GA with fuzzy in prediction of the disease. This algorithm was employed on three significant benchmark datasets obtained from the UCI repository, the Pima diabetes dataset, the Wisconsin (breast cancer dataset), and the heart disease dataset. Accuracy rate was found to be 88.93% for the Pima diabetes dataset, 99.5% for the Wisconsin dataset and 89.65% for the heart disease dataset [25]. Feature space, data size and class count had an impact on classifier performance. New Algorithms have to be implemented to enhance accuracy, reliability and efficiency of classifiers. DL-CNN (Deep Learning-based Convolutional Neural Network) was employed for visualising and classifying disease-oriented data. PSO and PCA techniques were used to analyse multivariate data for handling a lot of data. The efficiency of the recommended learning algorithm was explored using real-world datasets. Comparative analysis showed that DL (Deep Learning) performed better in comparison to other classifiers, with an accuracy rate of 0.9048% for the mifem dataset [26].

Classification was a powerful ML (Machine Learning) method that is generally utilised for prediction. Only a few classification algorithms perform prediction with better accuracy, and a few explore limited accuracy. ML methods employed included hybrid RF and LM (Linear Model) to determine the significant features for enhancing the accuracy rate in prediction [27]. The performance of such method was assessed and accuracy rate has been found to be 88.7% [28]. To enhance the accuracy, ensemble classification methods were applied by combining several classifiers. These algorithms include stacking, majority voting, boosting and bagging. The results of these methods were comparatively analysed, and showed that majority voting enhanced the accuracy rate by 7.26% [29]. In addition, a diagnosis system was presented that employed an optimised form of XGBoost (eXtreme Gradient Boosting) for predicting heart disease. Through this method, the accuracy rate was found to be 91.8% [30]. Furthermore, MLP and ANN model was compared on the heart disease dataset by considering other two researches utilising similar dataset. Accuracy of the MLP and ANN model was found to be 93.9%. However, the accuracy had to be further enhanced [31]. Additionally, an application system relying on optimal ANN was introduced to diagnose heart disease, which was taken as the deadliest disease worldwide. The performance of the introduced method was evaluated through the benchmark dataset taken from the UCI repository, and 95.41% was found as the accuracy rate [32]. Moreover, a comparative analysis was undertaken to find effective data mining tools and techniques for predicting heart disease. Weka, Knime, Scikit-Learn, Orange, Rapid Miner and Matlab were the utilised tools and SVM, LR, NB, RF, K-NN and ANN

(Artificial Neural Network) were the data mining techniques used. The analysis showed that Matlab's ANN model performed better with 85.86% accuracy [33]. Furthermore, classification with NN was suggested. Discretisation methods do not explore any enhancement in DT precision without or with voting. Analysis also revealed that Gini-index and NN performed better than other prediction models in the prediction of heart disease with an accuracy rate of 87.89% [34, 35]. However, this accuracy rate had to be further improved to correctly predict the occurrence of disease.

Various problems identified from the analysis of previous research are listed below.

- For selecting significant features and in order to increase the accuracy rate of heart disease prediction, various combinations of data mining techniques were chosen [12]. It was recommended to hybridise feature selection and extraction methods for improving the model performance in varied applications [28, 36].
- A few studies considered dimensionality reduction, but accuracy still had to be further enhanced. Accordingly, LDA and PCA were used and exhibited an accuracy rate of 94.53% [22].
- Accuracy was the significant parameter in the medical domain. Various researchers attempted to increase accuracy rate through several data mining methods such as vote with 87.4% accuracy [12], SVM showing 82.18% accuracy [13], soft-margin linear SVM and ILFS-90.65% accuracy rate [15], CDTL and RF exhibited 89.30% accuracy rate [16] and hybrid MLPSVM with 91.97% accuracy rate [17]. However, accuracy still had to be further improved to correctly predict the disease.

2. Materials and Methods

The research was mainly focused to predict heart disease as it was a major cause of death. Selection of relevant features, efficient and fast dimensionality reduction, and increased accuracy were the main issues identified in the existing work. This study intended to solve these issues based on data mining techniques comprised of DL algorithms for feature extraction and ML algorithms for classification. The overall flow of the proposed system is shown in Figure 1. Various processes were involved in predicting this disease. Initially, the heart disease Cleveland dataset was loaded and pre-processing was carried out. In this step, data cleaning was used to eliminate unwanted data and permitted the user to possess a dataset of highly useful information. After this, feature extraction and fusion were performed through the proposed deep CNN models. This helped to determine informative and compact feature sets for improving the classifier efficiency and reliability.

The present study uses one dimensional CNN (1D-CNN). Hence, before passing the features into CNN, the features are re-shaped and then passed into the model. Subsequently, CNN architecture is utilized where 16 hidden units exists, while it comprises 12 inputs ranging from 0-12. Before passing into the CNN, 12 input are regarded. Following this step, dimensionality reduction was performed by the proposed FTGM-PCA. This process assisted in minimising the space for data storage as the dimension counts were reduced, and this also minimised training or computation time and helped with data visualisation. This was then fed into 80% training and 20% testing. Then, classification was performed with the introduced IEB-RF classifier. This step helped in differentiating the normal and suspected patient data. Prediction was accomplished using the trained model and its efficiency was assessed by undertaking a performance analysis with the Cleveland dataset.

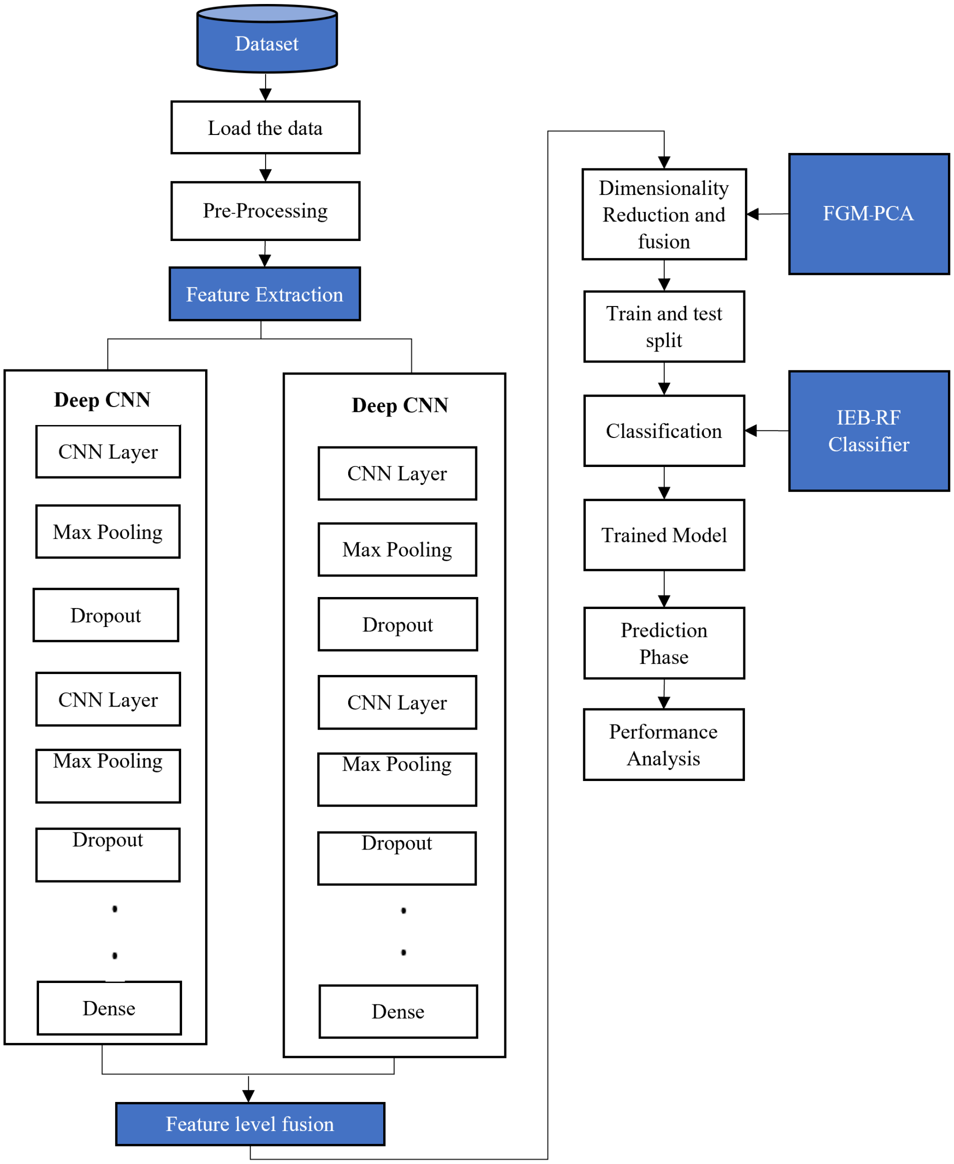

Figure 1. Overall view of the proposed system for predicting heart disease

2.1 Dataset description

The study considers Cleveland dataset for predicting the absence or presence of heart disease that consists of 14 features as described n in Table 1. This is a specific database exploited by various ML researchers until today. "Goal" field indicates that a particular patient is affected with heart disease, which is denoted by integer values like 1, 2, 3, 4 and 0, where 1, 2, 3, 4 shows the presence of disease, but 0 indicates the absence of the disease. This dataset can be accessed from UCI-ML repository that is available online. The dataset consists of 303 rows and 14 columns, and it can be accessed at https://archive.ics.uci.edu/ml/datasets/heart+disease.

Table 1. Dataset features

Features	Description
Age	Age (in years)
sex	Male is indicated as 1 Female is indicated as 0
restecg	Results of resting electrocardiographs Normal: Value 0 Having an ST-T wave abnormality (inverting T wave or elevating ST or depression rate >0.05mV): Value 1 Discovering probable or definite LVH (Left Ventricular Hypertrophy) by estes' criteria: Value 2
trestbps	RBP (Resting Blood Pressure) – mm Hg on hospital admittance
condition	No disease: 0 Disease: 1
thalach	HRR (High Heart Rate) achieved
thal	Normal: 0 Fixed defect: 1 Reversible defect and label: 2
cp	Kind of chest pain Typical-angina: Value 0 Atypical-angina: Value 1 Non-angina pain: Value 2 Asymptomatic: Value 3
chol	Serum cholesterol (mg/dl)
exang	Use prompted angina (no: 0, yes: 1)
FBS	Blood sugar >120mg/dl False: 0 True: 1
ca	Number of major vessels (0-3) colored by fluoroscopy
oldpeak	ST depression instigated based on rest
slope	Peak slope applying ST segment Up-sloping: Value 0 Flat: Value 1 Down-sloping: Value 2

2.2 Feature extraction and fusion-deep convolutional neural network

In recent years, DL has been confirmed to be a valuable tool in disease prediction as it has the ability to handle a huge amount of data. Hence, this study uses two deep CNN models for feature extraction, where the first model extracts 128 features and the second model extracts 256 features. In this case, the last dense layers of the two models extract 128 and 256 features. These models have the ability to automatically learn filters and subsequently integrate them in a hierarchical way to permit the latent concept description to recognise patterns. This model has various layers, and the filters get reduced in each of these layers, and only the relevant features get extracted with the initial layer possessing 5*5*1*64 filters that eventually gets reduced to 3*3*14*1 filters. Once the filters are applied to an input, the feature maps corresponding to CNN find the results. This means that at each of the layers, the feature map indicates the output for that particular layer. The main idea behind the visualisation of feature map for each input is to attain the detected features. CNN utilises learned filters for convoluting feature maps from preceding layers. Generally, filters are 2D weights that possess spatial association with one another. Moreover, stride indicates the number of steps moved in each convolutional stage. Its value is one by default. It can be found that output size is lower than input, as shown in Table 1. Padding is a method of adding zeroes for the input matrix in a symmetrical form and is used to maintain the output dimension as in the input. Finally, this model performs automatic detection of significant features with no human involvement. Followed this, feature level fusion is undertaken to assist in learning the overall features. Table 2 shows the five-layered CNN architecture and Table 3 shows the six-layered CNN architecture. It is observed that the features extracted are more in output in both models. This is because all the relevant features are extracted. These are further reduced through dimensionality reduction to attain a compact indication of 50 features, thus leading to better performance and low computational complexity.

Table 2. Five layered CNN architecture

Layer Name	Feature map	Filter	Stride	Pad	Output
Convolution Layer 1	16 *16 * 1	5 * 5 * 1 * 64	1	1 * 1	12 * 12 * 64
Activation function	12 * 12 *64	ReLU/P ReLU	1	0	12 * 12 * 64
Convolution Layer 2	12 * 12 * 64	1 * 1 * 64 * 44	1	1 * 1	12 * 12 * 128
Activation function	12 * 12 * 44	ReLU/P ReLU	1	0	12 *12 * 128
Convolution Layer 3	12 *12 * 64	1 * 1 * 44 * 24	1	1 * 1	12 * 12 * 256
Activation function	12 * 12 * 44	ReLU/P ReLU	1	0	12 * 12 * 256
Convolution Layer 4	12* 12 *24	1 * 1 * 24 * 14	1	1 * 1	12 * 12 *256
Activation function	12 * 12 * 24	ReLU/P ReLU	1	0	12 * 12 * 256
Convolution Layer 5	12 * 12 * 14	3 * 3 * 14 * 1	1	1 * 1	10 *10 * 128
Activation function	12 * 12 * 44	ReLU/P ReLU	1	0	12 * 12 * 128

Table 3. Six layered CNN architecture

Layer Name	Feature map	Filter	Stride	Pad	Output
Convolution Layer 1	16 *16 * 1	5 * 5 * 1 * 64	1	1 * 1	12 * 12 * 64
Activation function	12 * 12 *64	ReLU/P ReLU	1	0	12 * 12 * 64
Convolution Layer 2	12 * 12 * 64	1 * 1 * 64 * 44	1	1 * 1	12 * 12 * 128
Activation function	12 * 12 * 44	ReLU/P ReLU	1	0	12 *12 * 128
Convolution Layer 3	12 *12 * 64	1 * 1 * 44 * 24	1	1 * 1	12 * 12 * 256
Activation function	12 * 12 * 44	ReLU/P ReLU	1	0	12 * 12 * 256
Convolution Layer 4	12* 12 *24	1 * 1 * 24 * 14	1	1 * 1	12 * 12 *256
Activation function	12 * 12 * 24	ReLU/P ReLU	1	0	12 * 12 * 256
Convolution Layer 5	12 * 12 * 14	3 * 3 * 14 * 1	1	1 * 1	10 *10 * 128
Activation function	12 * 12 * 44	ReLU/P ReLU	1	0	12 * 12 * 128
Convolution Layer 6	12 * 12 * 34	3 * 3 * 14 * 1	1	1 * 1	10 *10 * 256
Activation function	12 * 12 * 14	ReLU/P ReLU	1	0	12 * 12 * 256

Various features are extracted in each layer of Deep CNN, and the extracting features get increased, which shows that all the relevant features are entirely extracted and fused. This is then fed into the proposed FTGM-PCA for dimensionality reduction and fusion.

2.3 Dimensionality reduction and Fusion-Fast Track Gram Matrix-Principal Component Analysis

In ML, kernel functions indicate the gram matrices, and this study proposed an FTGM-PCA (Fast Track Gram Matrix-Principal Component Analysis) to reduce the dimensionality of features in a fast way. The main application of GM is to calculate the linear independence as this is a significant step to remove the redundant features. The FTGM algorithm is shown in algorithm I. It takes Gram Matrix (GM) of $\mathrm{n} \times \mathrm{n}$ as input and returns $\mathrm{GM}^{\sim}{}_{\mathrm{k}} = \mathrm{CHSW}_{\mathrm{k}} + \mathrm{CHS}^{\mathrm{T}}$ as output, which is an approximate form decomposition. Through this algorithm, the gram matrix and correlation matrix are computed.

Algorithm I: Fast Track Gram Matrix

```
Input data: Gram Matrix (GM) of $\mathrm{n} \times \mathrm{n}$, $\{\mathrm{p_a}\}_{\mathrm{a}=1}^{\mathrm{n}}$ such that $\sum_{\mathrm{a}=1}^{\mathrm{n}} \mathrm{p_a} = 1, \mathrm{kc}$ and $\mathrm{c} \leq \mathrm{n}$
Output data: matrix $\mathrm{GM}^{\sim}$ of $\mathrm{n} \times \mathrm{n}$
Define $(\mathrm{sm_a} \times \mathrm{sm_b})$ matrix $\mathrm{S} = 0_{\mathrm{sm_a} \times \mathrm{sm_b}}$
Define $(\mathrm{sm_b} \times \mathrm{sm_b})$ matrix $\mathrm{RS} = 0_{\mathrm{sm_b} \times \mathrm{sm_b}}$
for $\mathrm{t} = 1, \ldots., \mathrm{sm_b}$ do
Pick $\mathrm{a_t} \in [\mathrm{n}]$ where $\Pr(\mathrm{a_t} = \mathrm{a}) = \mathrm{p_a}$
$\mathrm{R_{hett}} \in (\mathrm{sm_b pa_t})^{\frac{1}{2}}$
$\mathrm{S_{a_t t}} = 1$
```

```
end
Let $\mathrm{CHS} = \mathrm{GSRS}$ and $\mathrm{W} = \mathrm{RSS^TGSD}$, where, G-gram matrix, S-sampling matrix and D-diagonal matrix and RS=D
Calculate $\mathrm{W_k}$, best k-rank approximation to W
Return $\mathrm{GM^{\sim}}_{\mathrm{k}} = \mathrm{CHSW_k} + \mathrm{CHS^T}$
$\cap\,(\mathrm{CM}) = \min\{|(\mathrm{sm_b} + \mathrm{sm_a}) - \mathrm{X}|\}$: X is a Correlation Matrix
```

Correlation matrices are usually dense matrices with more dimensions. Hence, an approximating correlation matrix with a low rank is essential to reduce the feature counts and is computed using and FTGM algorithm. After computing gram and correlation matrices, they are fed into a PCA, and our study uses PCA for dimensionality reduction. This is a technique to reduce dimensionality in which data of high dimension is converted into low dimension by enhancing the low dimension variance. Following this, the eigenvectors, its values and CEV (Cumulative Explained Variance) are computed. If CEV is greater than or meets the threshold value, then the projection matrix is computed. Input is converted by projection matrix to attain the reduced dimensions, the eigenvectors having high eigenvalues are subsidized to new features after the dimensionality reduction. Hence, the proposed FTGM-PCA is used to perform dimensionality reduction as per Algorithm II.

Algorithm II: Fast Track Gram Matrix-Principal Component Analysis

```
Def$_{func}$: PCA
Input:
$X$-Feature set with $d$ dimension
Calculate gram and correlation matrix
While $(a \leq d)$ $do$ // $a$-iteration
While $(b \leq d)$ $do$// $b$-iteration
$sm_a$ ← sample feature mean $a$
$sm_b$ ← sample feature mean b
$\sigma_{ab} = \frac{1}{n}\sum_{k=1}^{m}(t_a^k - sm_a)(t_b^k - sm_b) * GM^{\sim}{}_k * \cap(CM)$ //
$\sigma_{ab}$-full PCA decomposition
$b = b + 1$
End while
$a = a + 1$
End while
decompose $\propto$ into eigenvectors and values
// $\propto$-magnitude of Eigenvalues and vectors
Compute CEV (Cumulative Explained Variance)
$if\ (CEV \geq threshold)$
```

```
Construct $W$     // $W$ is the projection matrix
End if
Convert input $T$ through $W$
Attain feature sub-space $T^{\sim}$ of k-dimension
Return $T^{\sim}$
```

The process involved in dimensionality reduction through the proposed FTGM-PCA is shown in Figure 2. Initially, the original matrix data is taken as input and mean is determined. Then, normalisation is performed to eliminate or minimise the redundant data. After this, matrix computation is performed where the gram and correlation matrix are calculated. Following this, Eigenvector and values are calculated and the values are set in descending order to compute the score matrix. Based on this, only relevant and reduced features are attained which improve the classification performance.

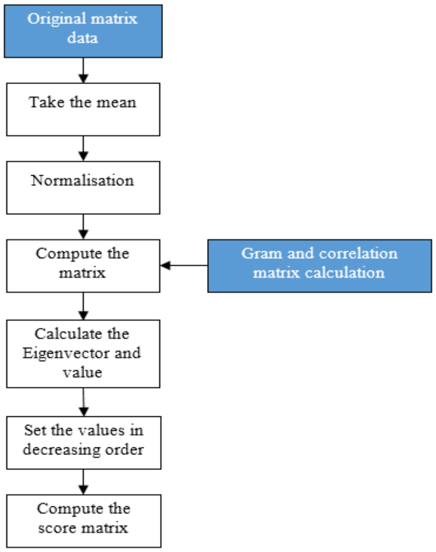
Original matrix data
Take the mean
Normalisation
Compute the matrix
Gram and correlation matrix calculation
Calculate the Eigenvector and value
Set the values in decreasing order
Compute the score matrix

Figure 2. Overall process of dimensionality reduction by FTGM-PCA

2.4 Classification-Informative Entropy Based-Random Forest

In this study, IEB-RF (Informative Entropy-Based Random Forest) was proposed for classification, where information entropy was used to determine the average information amount transferred through an event by taking into account all the probable results. RF produces effective predictions and can also be easily understood. It has the capability to deal with huge datasets flexibly and offers a high accuracy level in predicting the results. Hence, this study proposed IEB-RF to classify the normal and affected patient data. The stepwise process for this classification is presented in Figure 3. Initially, the tree nodes were created. Subsequently, a training data subset was chosen to build the next split. In this process, the variable subset was chosen. For each of the chosen variables, Gini index and info gain entropy were computed to return the best split, which was then fed to build the next split. After this, the prediction error was computed and the predicted value was returned.

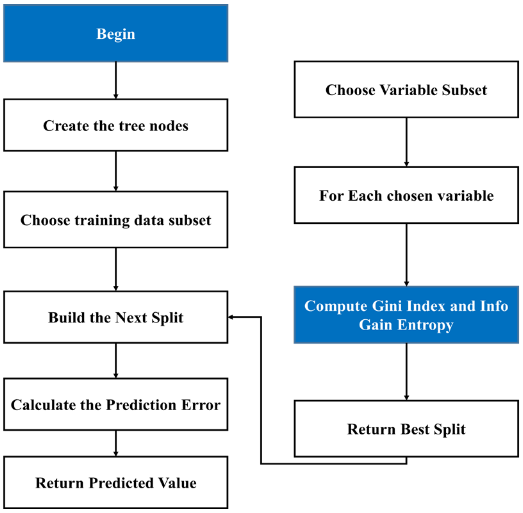

Figure 3. Overall process involved in classification by Informative Entropy Based-Random Forest

The overall process to classify the disease affected and normal patient data for heart disease prediction is shown in Figure.3. The stepwise process involved in classification is discussed below. A sample dataset SD in the feature space X having M dimension and count of good trees E was chosen from RF. Uncorrelated good trees $UCorr$ were chosen from E tress. The technique used to construct an enhanced RF from X along with $UCorr$ followed the following six main steps.

Step 1: Data sampling: utilise bagging technique to produce K of subsets of bag data $\{BSB_1, BSB_2, \dots, BSB_k\}$ through SD with substitution.

Step 2: Tree classifier construction: employ individual of data subsets of bag BSB_i for constructing trees and subsequently afford assessment of tree value. Repeat this step till all the trees are processed and generated.

Step 3: Tree ordering: categorise all the K trees in the descending order of AUC.

Step 4: Calculate Gini index and information gain for subsets of variable. Sort information gain into a vector form to perform faster processing. Gini index is given by equation 1.

$$\mathrm{GI(K)} = \frac{1}{2} - \left[1 - \sum\nolimits_{j} \mathrm{p}\,(\mathrm{r_f})^2\right] \qquad (1)$$

Where $\mathrm{p(r_f)}$ denotes the relative frequency for class $\mathrm{r_f}$ at K node. In addition, the information gain is given by equation 2 and equation 3.

$$\mathrm{Gain(CLC)} = -\sum\nolimits_{\mathrm{I \in P}} \mathrm{r\,(I)\log_2 r(I)} \qquad (2)$$

Where $\mathrm{r(I)}$ is the ratio of CLC pertaining to class I.

$$\mathrm{Gain(CLC, A)} = -\mathrm{Entropy(CLC)} - \sum\nolimits_{\mathrm{v \in A}} \left(\frac{|\mathrm{CLC_V}|}{\mathrm{CLC}}\right). -\mathrm{Entropy(CLC)} \qquad (3)$$

Where CLC_v indicates the subset corresponding to CLC and feature A possesses value v, $|CLC_v|$ indicates the count of classified instances and $|CLC|$ represents the amount of classifier instances.

Step 5: Choose effectively performing trees: Choose top trees E with high values of AUC.

Step 6: Improved RF construction: Correlation amongst the predicted possibilities of E trees is perceived and is given by equation 4.

$$\mathrm{t} = \begin{bmatrix} 1 & \mathrm{t}(1{,}2) & \mathrm{t}(1{,}3) \\ \vdots & \vdots & \vdots \\ \mathrm{t}(\mathrm{t},1) & \mathrm{t}(\mathrm{t},2) & 1 \end{bmatrix} * \mathrm{Gain}(\mathrm{CLC},\mathrm{A}) \tag{4}$$

In equation 4, $\mathrm{t}(1{,}2)$, $\mathrm{t}\,(1,3)$, $\mathrm{t}(\mathrm{t},2)$ are utilized as inputs in variable matrix. Correlations amongst predicted probabilities of such trees (t) are indicated in equation 4. The perceived t is utilised as an input in the variable clustering method to obtain the Uncorrelated tree clusters $Corr$. In each of the clusters, categorise trees in descending order of AUC are shown. Choosing a tree from individual clusters that have high values of AUC affords high performing uncorrelated trees $UCorr$ and ensemble the trees into an improved RF. Majority vote for the trees was used to make decisions for ensemble classification. Additionally, during the iterating matrix, last row and column are considered as an input in variable clustered features. Moreover, this t matrix imitates top working trees from uncorrelated high working trees and these are chosen for forming an improvised Random Forest. The improved RF technique possesses the count of high performing trees E from which the high performing uncorrelated trees $UCorr$ are chosen to get the improved RF. Finally, classification is performed to predict the disease accurately and is then verified.

3. Results and Discussion

3.1 Performance metrics

The proposed system was assessed in terms of accuracy, AUC (Area under Curve), F1-score, precision, and sensitivity to find each value of these metrics.

- **Accuracy:** This represents the correct classification achieved. This study considers this metric as it is important in the healthcare sector to find the effectiveness of the introduced system to predict heart disease and is given in equation 5.

$$\mathrm{Accuracy} = \frac{\text{True Positive+True Negative}}{\text{True Positive+True Negative+False Positive+False Negative}} \tag{5}$$

- ***AUC (Area under Curve):*** It is the correct integral curve that reveals the alterations in classification. In this study, AUC is taken into account to determine the degree to which alterations occur during the classification process and is given in equation 6.

$$AUC = \frac{1}{2}\left(\left(\frac{\text{True Positive}}{\text{True Positive+False Negative}}\right) + \left(\frac{\text{True Negative}}{\text{True Negative+False Positive}}\right)\right) \tag{6}$$

- ***F1-score:*** It can also be termed as F-measure, which is the harmonic mean of recall and precision. It is utilised to determine the efficiency of measurement. This study utilises this metric for calculating the efficiency in disease classification and is denoted in equation 7.

$$\text{F1}-\text{score} = \frac{2*(\text{Precision}*\text{Recall})}{\text{Precision}+\text{Recall}} \tag{7}$$

- ***Precision:*** It is defined as the classification of correct classification count and is impacted through incorrect classification. This study considers precision for determining the count of accurate classification achieved, and is shown in equation 8.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \tag{8}$$

- ***Sensitivity:*** It denotes the overall positive segments that are correctly recognised. This study takes this metric to determine the count of correct prediction and is represented by equation 9.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \tag{9}$$

3.2 Experimental results

The proposed system was experimentally implemented and the obtained outcomes are shown in Table 4 and Table 5. Table 4 depicts the initial feature extraction of 128 features using the first level deep CNN model and the next feature extraction of 256 features are extracted using the second level deep CNN model. These features are extracted based on their relevance. Dimensionality reduction was performed to filter the features further, and finally, 50 features were extracted. This process assisted in increasing the accuracy. That is, when feature extraction was performed by the proposed deep CNN models along with dimensionality reduction by FTGM-PCA, the classifier performance increased. Table 5 shows the accuracy of the system with and without feature extraction. Accuracy was found to be 93.65% without feature extraction whereas 97% accuracy was attained with feature extraction.

Table 4. Features after dimensionality reduction

Total Number of features	Feature Extraction	Dimensionality reduction
13	128+256	50

Table 5. Accuracy based on feature extraction

With-Feature Extraction	Without-Feature Extraction
97	93.65

Hence, it was found that dimensionality reduction with efficient feature extraction improved the classifier performance, and this improved the prediction rate. The results presented in Table 5 verify the significance of feature extraction in enhancing the accuracy rate for heart disease prediction. The effective learning of deep CNN models and removal of redundant features faster through FTGM-PCA with IEB-RF gave the proposed system a high accuracy rate, which was 97%.

3.3 Comparative analysis

The performance of the proposed system was analysed through a comparative analysis where various existing systems were compared with this system. Accuracy was the significant metric, and hence it was considered at first for analysis. The obtained results are shown in Table 6. NB, SVM, J48, functional trees, bagging, K-NN, LR, ANN, DT, classification tree and RF were the traditional methods considered for analysis [37].

Table 6. Analysis of the proposed and traditional system with respect to accuracy

Authors	Technique	Accuracy
Room *et al*	Naïve Bayes	84.50%
	SVM	84.50%
	Functional trees	84.50%
Vembandasamy *et al.*	Naïve Bayes	86.42%
Chaurasia *et al*.	J48	84.35%
	Bagging	85.03%
	SVM	94.60%
Parthiban *et al*.	Naïve Bayes	74%
Seema *et al.*	Naïve Bayes	95.56%
Kumar Dwivedi	Naïve Bayes	83%
	Classification tree	77%
	K-NN	80%
	Logistic regression	85%
	SVM	82%
	ANN	84%
Existing	Naïve Bayes	88.16%
	K-NN	90.79%
	Decision tree	80.26%
	Random forest	86.84%
Proposed	FGM-PCA and IEB-RF	97%

From the obtained results in Table 6, it was found that existing systems afforded better results for RF showing 86.84%, and LR which was 85%. However, compared to the proposed system, the accuracy of other traditional systems was low. The accuracy of the introduced work was found to be 97% which was higher than other techniques. This superior performance of the proposed FTGM-PCA was due to its capability for minimizing the computational complexity as the features were analyzed with efficient learning. This assisted in computing an easily interpretable low rank approximation for forming the gram-matrix. The significance of rows and columns of the matrix that were carefully selected with consistent probability distributions for attaining reliable error-bounds for several typical matrix operations was apparent. Through the use of such operations, data with reduced noise was attained by disregarding small variations in background automatically. Due to these operations and advantages, the proposed system showed better performance than conventional strategies. To further confirm the efficacy of the proposed system, it is also analysed in terms of various other metrics such as precision, sensitivity and F1-score. K-NN, LR, SVM, LDA, RF, GB (Gradient Boosting) and CART (Classification and Regression Tree) were the various conventional methods taken for analysis [38]. The results obtained are presented in Table 7.

Table 7. Analysis of the existing and the proposed work with respect to various metrics

Algorithm	Accuracy (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
K-NN	60	61	59	58
LR	78	79	78	78
LDA	78	80	79	79
SVM	79	80	79	79
CART	68	69	68	68
GB	81	79	84	81
RF	83	81	87	84
Existing method (ensemble approach)	93	96	91	93
Proposed FTGM-PCA and IEB-RF	97	96	97	96

From Table 7, it is clear that the existing RF showed 83%, the existing ensemble approach revealed 93% accuracy. Various other methods showed different variations in accuracy: K-NN-60%, LR and LDA-78%, SVM-79%, CART-68%, and GB-81%. Though better accuracy was attained by the existing ensemble method with 93%, it was lower than the proposed method that showed a high prediction rate of 97%. Similarly, precision, sensitivity and F1-score are also analysed, and the proposed method was found to have higher values in terms of all the considered metrics, with precision-96%, 97% sensitivity and 96% F1-score.

In addition, the number of features extracted by the conventional and proposed system were assessed. Accuracy, AUC and F1-score were also used as performance metrics in this analysis and the obtained outcomes are shown in Table 8. Various traditional methods considered include Rotation forest-J48-CFS, SMO-expert based feature selection, LR-LASSO, NN, Two-tier ensemble PSO based feature selection, PSO fuzzy expert system, CFS-PSO-clustering-MLP, Boosted-C5.0 and Voting-NBLR [39].

From Table 8, it was found that the number of features extracted by different methods varied according to their efficacy. Rotation forest-J48-CFS extracted 7 features and CFS-PSO-clustering-MLP extracted 5 features. These were found to be minimum and accuracy rates were also lower for these algorithms at rates of 84.48% and 90.28%. Among all, the PSO fuzzy expert system extracted 76 features, which was more than other methods. However, the accuracy of this method was 93.27%. Relevant features have to be selected without compromising prediction rate. In that sense, the proposed GTGM-PCA and IEB-RF showed better and relevant feature extraction with no compromise in the accuracy rate of 97%. In addition, AUC and F1-score are also considered for analysis. Most of the studies have not even considered these metrics for analysis but the proposed system considered it, and also performed well showing 96.24% as F1-score and 96.14% as AUC. This reveals the superior efficiency of the proposed system to the traditional methods. Hence, the analysis carried out using three conventional studies revealed that the proposed FTGM-PCA and IEB-RF showed better performance than other traditional works. Deep CNN models performed efficient learning. Particularly, Deep CNN was exploited for extracting features with minimized non-linearity and regularization operation. This results in optimal prediction that makes the results optimal. The main merit of using Deep CNN is for attaining robust and suitable features. When feature quality is minimum, this might result in minimum performance and generalization features, in spite of the possession of ideal classification strategies. Through the use of Deep CNN, feature extraction from raw input could work automatically. These features enhanced the performance of the proposed classifier and also reduced complexity.

For confusion matrix, the experiments were assessed and the confusion matrix corresponding to it is shown in Figure 4. From the outcomes, it is clear that 26 attacks were correctly classified, while there was one misinterpretation of attack as normal. On contrary, 1 normal was misinterpreted as attack, while 33 normal cases were correctly classified as normal. In this case, correct classifications were higher than misclassification, confirming the efficacy of the proposed work.

Table 8. Analysis of the traditional and proposed system with respect to performance metrics

Technique	**Number of features**	**Accuracy (%)**	**F1 (%)**	**AUC (%)**
Rotation forest-J48-CFS	7	84.48	NR	89.5
PSO fuzzy expert systems	76	93.27	NR	NR
SMO-expert based feature selection	8	84.49	86.2	NR
CFS-PSO-clustering-MLP	5	90.28	NR	NR
LR-LASSO	6	89	NR	NR
Boosted-C5.0	12	77.8	NR	NR
NN	12	81.9	NR	NR
Voting-NBLR	9	87.41	NR	NR
Two-tier ensemble PSO based feature selection	7	85.71	86.49	85.86
Proposed FTGM-PCA and IEB-RF	**50**	**97**	**96.24**	**96.14**

*NR: Not Reported

4. Conclusions

The main purpose of this study is to predict heart disease based on data mining methods. This study used the DL method for feature extraction and the ML method for classification purpose. The proposed deep CNN models performed relevant feature extraction and feature level fusion by removing redundant data through efficient learning, FTGM-PCA reduced dimensionality of features quickly and IEB-RF classified the normal and disease-affected patient data. The performance of this system was determined by comparing it with three conventional methods. Significant metrics were used. It was found that the IEB-RF model performed accurate classification in comparison to other existing methods, achieving a rate of 97% for Cleveland heart disease dataset. The high accuracy obtained with this system makes it highly suitable for diagnosing heart disease. In the future, the accuracy of heart disease prediction can be enhanced above 97% by using different data mining techniques.

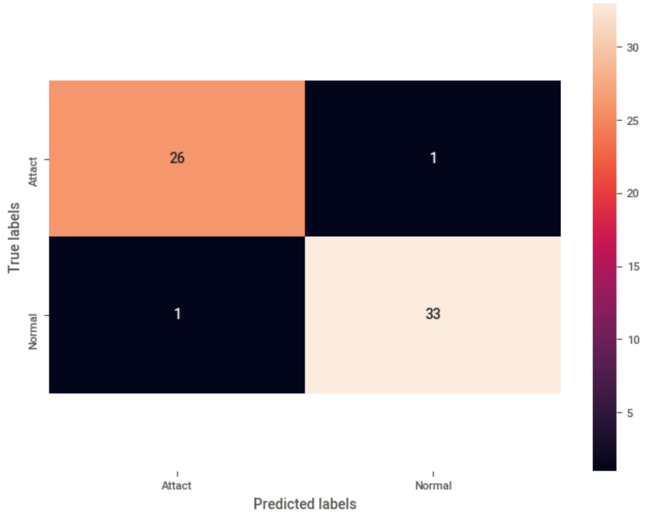

Figure 4. Confusion matrix

References

[1] Beyene, C. and Kamat, P., 2018. Survey on prediction and analysis the occurrence of heart disease using data mining techniques. *International Journal of Pure and Applied Mathematics*, 118(8), 165-174.

[2] Preetha, J., Raju, S., Kumar, A., Sayyad, S. and Vengatesan, R., 2020. Data mining technique based critical disease prediction in medical field. In: D.J. Hemanth, V.D.A. Kumar and S. Malathi, eds. *Advances in Pararllel Computing. Vol. 37. Intelligent Systems and Computer Technology*. Amsterdam: IOS Press, pp.104-108.

[3] Diwakar, M., Tripathi, A., Joshi, K., Memoria, M. Singh, P. and Kumar, N., 2021. Latest trends on heart disease prediction using machine learning and image fusion. *Materials Today*: *Proceedings*, 37, 3213-3218, DOI: 10.1016/j.matpr.2020.09.078.

[4] Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S. and Singh, P., 2021. Prediction of heart disease using a combination of machine learning and deep learning. *Computational Intelligence and Neuroscience,* 2021, DOI: 10.1155/2021/8387680.

[5] Sharma, P., Choudhary, K., Gupta, K., Chawla, R., Gupta, D. and Sharma, A., 2020. Artificial plant optimization algorithm to detect heart rate and presence of heart disease using machine learning. *Artificial Intelligence in Medicine*, 102, DOI: 10.1016/j.artmed.2019.101752.

[6] Jain, D. and Singh, V., 2018. Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, 19(3), 179-189, DOI: 10.1016/j.eij. 2018.03.002.

[7] Maji, S. and Arora, S., 2019. Decision tree algorithms for prediction of heart disease. *Proceedings of Third International Conference on Information and Communication Technology for Competitive Strategies*, Udaipur, India, December 16-17, 2017, pp. 447-454.

[8] Spencer, R., Thabtah, F., Abdelhamid, N. and Thompson, M., 2020. Exploring feature selection and classification methods for predicting heart disease. *Digital Health,* 2020, DOI: 10.1177/2055207620914777.

[9] Javeed, A., Rizvi, S.S., Zhou, S., Riaz, R., Khan, S.U. and Kwon, S.J., 2020. Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification. *Mobile Information Systems*, 2020, DOI: 10.1155/2020/8843115.

[10] Nalluri, S., Saraswathi, V., Ramasubbareddy, S., Govinda, K. and Swetha, E., 2020. Chronic heart disease prediction using data mining techniques. In: K. Raju, R. Senkerik, S. Lanka and V. Rajagopal, eds. *Data Engineering and Communication Technology*. Singaporer: Springer, pp. 903-912.

[11] Dulhare, U.N., 2018. Prediction system for heart disease using Naive Bayes and particle swarm optimization. *Biomedical Research,* 29(12), 2646-2649.

[12] Amin, M.S., Chiam, Y.K. and Varathan, K.D., 2019. Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, 36, 82-93, DOI: 10.1016/j.tele.2018.11.007.

[13] Shah, S.M.S., Batool, S., Khan, I., Ashraf, M.U., Abbas, S.H. and 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, DOI: 10.1016/j.physa.2017.04.113.

[14] Lv, N., Chen, C., Qiu, T. and Sangaiah, A.K., 2018. Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in SAR images. *IEEE Transactions on Industrial Informatics,* 14(12), 5530-5538, DOI: 10.1109/TII.2018.2873492.

[15] Le, H.M., Tran, T.D. and Tran, L.V., 2018. Automatic heart disease prediction using feature selection and data mining technique. *Journal of Computer Science and Cybernetics*, 34(1), 33-48, DOI: 10.15625/1813-9663/34/1/12665.

[16] Magesh, G. and Swarnalatha, P., 2021. Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. *Evolutionary Intelligence,* 14(2), 583-593, DOI: 10.1007/s12065-019-00336-0.

[17] Azhar, M. and Thomas, P.A., 2020. Heart disease prediction based on an optimal feature selection method using autoencoder. *International Journal of Scientific Research in Science and Technology*, 7(4), 25-38, DOI: 10.32628/IJSRST20748.

[18] Keerthika, T. and Premalatha, K., 2019. An effective feature selection for heart disease prediction with aid of hybrid kernel SVM. *International Journal of Business Intelligence and Data Mining*, 15(3), 306-326, DOI: 10.1504/IJBIDM.2019.101977.

[19] Vijayashree, J. and Sultana, H.P., 2018. A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier. *Programming and Computer Software*, 44(6), 388-397, DOI: 10.1134/S0361768818060129.

[20] Harimoorthy, K. and Thangavelu, M., 2021. Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. *Journal of Ambient Intelligence and Humanized Computing,* 12(3), 3715-3723, DOI: 10.1007/s12652-022-03971-1.

[21] Wiharto, W., Kusnanto, H. and Herianto, H., 2017. System diagnosis of coronary heart disease using a combination of dimensional reduction and data mining techniques: A

review. Indonesian. *Journal of Electrical Engineering and Computer Science,* 7(2), 514-523, DOI: 10.11591/ijeecs.v7.i2.pp514-523.

[22] Burse, K., Kirar, V.P.S., Burse, A. and Burse, R., 2019. Various preprocessing methods for neural network based heart disease prediction. In: S. Tiwari, M. Trivedi, K. Mishra, A. Misra and K. Kumar, eds. *Smart Innovations in Communication and Computational Sciences*. Singapore: Springer, pp. 55-65.

[23] Nilashi, M., Ahmadi, H., Manaf, A.A., Rashid, T.A., Samad, S. and Shahmoradi, L., 2020. Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates. *International Journal of Fuzzy Systems*, 22(4), 1376-1388, DOI: 10.1007/s40815-020-00828-7.

[24] Thiyagaraj, M. and Suseendran, G., 2018. An efficient heart disease prediction system using modified firefly algorithm based radial basis function with support vector machine. *International Journal of Engineering and Technology,* 7(2.33), 1040-1045.

[25] Sujatha, R., Ephzibah, E., Dharinya, S., Maheswari, G.U., Mareeswari, V. and Pamidimarri, V., 2018. Comparative study on dimensionality reduction for disease diagnosis using fuzzy classifier. *International Journal of Engineering and Technology*, 7(1), 79-84.

[26] Rao, G.M., Kumar, T.R. and Reddy, A.R., 2020. CNN-BD*:* An approach for disease classification and visualization. In: S. Borah, V.E. Balas and Z. Polkowski, eds. *Advances in Data Science and Management*. Singapore: Springer, pp. 149-157.

[27] Uddin, S., Khan, A., Hossain, M.E. and Moni, M.A., 2019. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19(1), 1-16, DOI: 10.1186/s12911-019-1004-8.

[28] Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7, 81542-81554, DOI: 10.1109/ACCESS. 2019.2923707.

[29] Latha, C.B.C. and Jeeva, S.C., 2019. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked,* 16, DOI: 10.1016/j.imu.2019.100203.

[30] Budholiya, K., Shrivastava, S.K. and Sharma, V., 2020. An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University-Computer and Information Sciences*, 34(7), 4514-4523.

[31] Kaya, M.O., 2021. Performance evaluation of multilayer perceptron artificial neural network model in the classification of heart failure. *The Journal of Cognitive Systems,* 6(1), 35-38, DOI: 10.52876/jcs.913671.

[32] Selvi, R.T. and Muthulakshmi, I., 2021. An optimal artificial neural network based big data application for heart disease diagnosis and classification model. *Journal of Ambient Intelligence and Humanized Computing,* 12(6), 6129-6139, DOI: 10.1007/s12652-022-04077-4.

[33] Tougui, I., Jilbab, A. and Mhamdi, J.E., 2020. Heart disease classification using data mining tools and machine learning techniques. *Health and Technology,* 10, 1137-1144, DOI: 10.1007/s12553-020-00438-1.

[34] Mathan, K., Kumar, P.M., Panchatcharam, P., Manogaran, G. and Varadharajan, R., 2018. A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. *Design Automation for Embedded Systems,* 22(3), 225-242, DOI: 10.1007/s10617-018-9205-4.

[35] Ali, L. Rahman, A., Khan, A., Zhou, M., Javeed, A. and Khan, J.A., 2019. An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network. *IEEE Access*, 7, 34938-34945, DOI: 10.1109/ACCESS.2019.2904800.

[36] Das, H., Naik, B. and Behera, H., 2020. Medical disease analysis using neuro-fuzzy with feature extraction model for classification. *Informatics in Medicine Unlocked,* 18, DOI: 10.1016/j.imu.2019.100288.

[37] Shah, D., Patel, S. and Bharti, S.K., 2020. Heart disease prediction using machine learning techniques. *SN Computer Science*, 1(6), 1-6, DOI: 10.1007/s42979-020-00365-y.

[38] Mienye, I.D., Sun, Y. and Wang, Z., 2020. An improved ensemble learning approach for the prediction of heart disease risk. *Informatics in Medicine Unlocked*, 20, DOI: 10.1016/j. imu.2020.100402.

[39] Tama, B.A., Im, S. and Lee, S., 2020. Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble. *BioMed Research International*, 2020, DOI: 10.1155/2020/9816142.