HDDSS: An Enhanced Heart Disease Decision Support System Using RFE-ABGNB Algorithm

M. Dhilsath Fathima[1]*, S. Justin Samuel[2], S. P. Raja[3]

[1] Research Scholar, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu (India)
2 Department of Computer Science and Engineering, PSN Engineering College, Tirunelveli, Tamil Nadu, (India)
3 School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, (India)

Received 20 April 2021 | Accepted 12 July 2021 | Published 20 October 2021

Abstract

Heart disease is the leading cause of mortality globally. Heart disease refers to a range of disorders that affect the heart and blood vessels. The risks of developing heart disease become minimized if heart disease is detected early. Previous studies have suggested many heart disease decision-support systems based on machine learning (ML) algorithms. However, the lower prediction accuracy is the main issue in these heart disease decision-support systems. The proposed work developed a heart disease decision-support system (HDDSS) that can predict whether or not a person has heart disease. The main goal of this research work is to use the RFE-ABGNB to improve HDDSS prediction accuracy. The Cleveland heart disease dataset is used for training and validating the proposed HDDSS. The two significant stages of HDDSS are the feature selection stage and the classification modeling stage. The recursive feature elimination (RFE) technique is used in the first stage of HDDSS to select the relevant features of the heart disease dataset. In the second stage of HDDSS, the proposed Adaptive boosted Gaussian Naïve Bayes (ABGNB) algorithm has been used to construct a classification model for training and validating a heart disease decision-support system. An output of HDDSS is analyzed using various classification output measures. According to the results obtained, our proposed method attained a predictive performance of 92.87 percent. This HDDSS model would perform well when compared to other heart disease decision-support systems found in the literature. According to our experimental analysis, the RFE-ABGNB focused heart disease decision-support system is more appropriate for a heart disease prediction.

Keywords

ABGNB Algorithm, Heart Disease Prediction, Machine Learning, Recursive Feature Elimination, UCI Heart Disease Dataset.

DOI: 10.9781/ijimai.2021.10.003

I. Introduction

Heart disease is the leading cause of the increasing mortality rate in humans. The risk of heart disease is more in patients with uncontrolled diabetes, low high-density lipoprotein (HDL), increased low-density lipoprotein (LDL), higher Body Mass Index (BMI), smoking, and high blood pressure [1], [2]. As a result, predicting the onset of heart disease at an early stage is essential for controlling risk factors and preventing heart disease. The main objective of this paper is to build a heart disease decision-support system (HDDSS) using the RFE-ABGNB model. This HDDSS can predict heart disease risk and can be used to diagnose and prevent heart disease at an early stage [3], [4].

The HDDSS uses the RFE-ABGNB, which is a machine learning approach, to predict heart disease. Recursive feature elimination (RFE) is a feature selection method used by HDDSS to select the relevant input features from a heart disease dataset. The Adaptive Boosted Gaussian Naive Bayes Algorithm (ABGNB) is a proposed ensemble classifier used to build an HDDSS that predicts heart disease in people by evaluating heart disease risk factors. The HDDSS utilizes the UCI heart disease dataset (UCI) for training and validating the proposed ABGNB classifier. The HDDSS model will predict the probability of developing heart disease using the patient heart disease risk factors as input. The HDDSS efficiency has been evaluated using various classification performance measures.

The remainder of the paper is structured as follows: Section II reviews relevant state-of-the-art research in the automated heart disease diagnosis system; Section III describes the proposed model; and Section IV illustrates and assesses the proposed model's experimental results. Section V depicts the conclusion of the proposed work.

II. Related State-of-art Work

Machine learning (ML) techniques are increasingly being utilized to predict heart disease. This section discusses the state-of-art approaches to develop a heart disease decision-support system using ML algorithms. The Cleveland heart disease dataset is used as an input in all examined literature to build a heart disease prediction model. Haq, Amin Ul et al. [5] developed a hybrid intelligent system

* Corresponding author.
E-mail address: dilsathveltech123@gmail.com

framework for predicting heart disease. This model uses three feature selection techniques such as Relief feature selection method, the minimal-redundancy-maximal relevance method, least absolute shrinkage and selection operator methods (LASSO) for selecting the best features of the input dataset. Logistic regression, K-nearest neighbor, Artificial neural network, Support vector machine (SVM), decision tree, and naive Bayes are among the ML classifiers used to classify the selected features. Various classifier performance measures have been used to test the proposed classifier results. According to the results of the performance analysis, logical regression and support vector machine outperforms other classifiers. Logistic regression got 84% accuracy before feature selection and achieved 89% after the relief feature selection algorithm. SVM obtained 86% accuracy before feature selection and 88% after the LASSO feature selection method. The results of this study show that integrating feature selection techniques into machine learning classifiers increases classifier accuracy.

This research paper [6] compared the performance of three machine learning algorithms such as BayesNet (BN), SVM, functional trees (FT) for effective diagnosis and monitoring of the consequences of heart disease. In this work, the BayesNet algorithm and SVM achieved 83.8% accuracy, and Functional trees achieved 81.5% accuracy. Then, the Best first selection algorithm is applied to select the best feature. The accuracy of the classifiers is improved by about 3% this time after they trained using the selected features. Thus, BayesNet's accuracy increased to 84.5 percent, SVM achieved 85.1 percent accuracy, and Functional trees achieved 84.5 percent accuracy.

Mohan et al. [7] proposed a hybrid machine learning model called Hybrid random forest with a linear model (HRFLM) for predicting cardiovascular disease. HRFLM-based heart disease prediction model gave a prediction accuracy level of 88.7% which is above other ML classifiers such as naïve bayes, generalized linear model, logistic regression, deep learning, decision tree, random forest, gradient boosted trees, SVM, VOTE classifier.

This research paper [8] used various machine learning classification algorithms such as SVM, k-nearest neighbor(K-NN), artificial neural network (ANN), naïve bayes (NB), logistic regression (LR), decision tree (DT) for the identification of heart disease. This model used a feature selection algorithm called the fast conditional mutual information algorithm (FCMIM) to improve classifier accuracy with improved classifier execution time. The performance of the FCMIM method has been compared with other feature selection algorithms like Relief, Minimal-redundancy-maximal-relevance (mRmR), Least-absolute-shrinkage-selection-operator algorithm (LASSO), Local learning-based features selection algorithms (LLBFS). This outputs analysis exhibits that the FCMIM outperforms other four feature selection method on the specified ML algorithms.

Chen et al. [9] developed a heart disease prediction system (HDPS) using Learning vector quantization (LVQ) which is a prototype-based classification algorithm that works based on Artificial intelligence network concepts. HDPS achieved an accuracy score of 80%, sensitivity of 85%, and specificity of 75%.

Hidayet takci [10] proposed an improved heart attack prediction system to decide the best machine learning approach and the best feature selection technique for predicting heart disease. This author has done a comparative analysis of ML algorithms such as c4.5 classifier, Classification-Regression Tree, SVM, Iterative Dichotomiser 3, K-NN, Multi-layer perceptron (MLP), Naïve bayes, Logistic regression models, and feature selection methods like reliefF, Forward-logit, Backward-logit, Fisher filtering. This model uses a Statlog heart disease dataset which is a publicly available dataset. A computer-aided heart disease diagnosis system is built using a combination of feature selection and classification algorithms in this model. Based on the comparative analysis outcomes, a SVM with a linear kernel is suggested as the best classification model when combined with the reliefF feature selection method. This model indicates that Linear kernel SVM with ReliefF feature selection algorithm is more efficient at predicting heart disease, with an accuracy of 84.81 percent.

Thippa Reddy et al. [11] developed an automated heart disease prediction model using a firefly and BAT swarm intelligence-based OFBAT-RBFL algorithm. This model focuses on three publicly accessible heart study datasets from the UCI machine learning repository: Hungarian, Cleveland, and Switzerland. The Fuzzy logic model is used to make a classification model by generating fuzzy system rules using the selected features. Then OFBAT algorithm is applied for selecting relevant fuzzy rules, enhance the performance of the prediction model, and optimizing the output rules of the fuzzy logic system. The outcome of the experiment indicates that the RBFL algorithm outperforms the existing ML model by achieving 78 percent accuracy.

This related study demonstrates how researchers have used feature selection approaches and machine classifiers to develop an automated heart disease diagnosis model. The motivation of this study is to enhance the accuracy of the heart disease prediction model using improved feature selection and machine learning classifier. The following is the contribution of this suggested work: (1) Recursive feature elimination algorithm is used to select a relevant feature of the input dataset, (2) For building a heart disease decision-support system, the ABGNB is proposed and used as a classifier.

III. Proposed Methodology

The proposed HDDSS has been developed using the RFE-ABGNB methodology, which combines the recursive feature elimination method (RFE) for identifying significant heart disease risk factors with the adaptive boosted Gaussian Naive bayes (ABGNB) algorithm for training and validating the HDDSS. The development of HDDSS consists of two main stages: In the first stage, the recursive feature elimination algorithm is applied to the UCI input dataset to determine the optimal heart disease input features. In the second step, the proposed ABGNB classifier train and validate the heart disease prediction model using selected inputs from the RFE algorithm. The results obtained from the ABGNB classification model are evaluated with other machine learning (ML) models such as Naïve bayes (NB), K-nearest neighbor (K-NN), SVM, and Decision tree (DT). For measuring the efficiency of the proposed model, different classification performance metrics [12] were used, namely, Classifier Accuracy, Misclassification rate, Sensitivity, Specificity, Precision, F-Score, Receiver operating characteristic curve. This proposed system's process flow diagram is shown in Fig. 1.

A. Dataset Description

This work uses a Cleveland heart disease dataset from the UCI repository; this is available online [13]. This dataset consists of data about 303 individuals (303 samples), 13 heart disease predictors, and one class attribute with binary outcomes as 1 (heart disease-Positive) and 0 (heart disease-Negative). Heart Disease-Positive indicates the patient has a heart disease problem, and heart disease-Negative implies the patient has no heart disease. The input dataset contains 164 samples of the positive class and 139 samples of the negative class. There are no missing values in this dataset.

B. Recursive Feature Elimination

Feature selection is one of the data preprocessing procedures for identifying and selecting the features most associated with the output variable. The feature selection step is necessary for this proposed

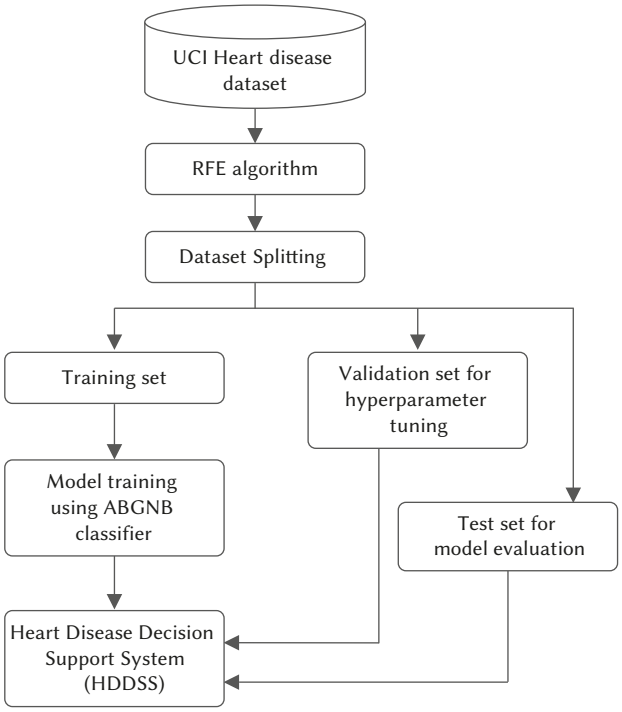

Fig. 1. The proposed HDDSS.

model for choosing relevant input features from the dataset by dropping irrelevant, redundant, noisy features [14], [15]. Keeping irrelevant features in our dataset may reduce the outcomes of the machine learning model.

For feature selection, the proposed work uses recursive feature elimination (RFE) [16]. The RFE algorithm is a recursive method to find out the statistical significance of the features. The statistical significance has calculated using criteria called hypothesis testing.

Hypothesis testing uses a p-value which is an observed significance value of input feature and it is a probabilistic measure to evaluate the hypothesis [17]. The statistical relationship exists between the input and output feature if the p-value of an input feature is less than the significance threshold (α). RFE uses 0.05 as the threshold value (α) [18]. The RFE algorithm begins with the full feature set D_s consisting of input features p_1, p_2, ... p_N, and then recursively prunes irrelevant features based on the hypothesis statement given in equations (1) & (2) at each iteration until the p-value of features is smaller than the threshold value (α). Fig. 2 depicts the steps of the RFE algorithm.

RFE uses two kinds of hypothesis called the null hypothesis and alternate hypothesis for selecting optimal features of input dataset D_s. Statements of a null hypothesis and alternate hypothesis are explained below:

Null Hypothesis: This hypothesis states that there is no relationship between the selected input feature and output feature when the p-value of the selected input feature is greater than or equal to the threshold value. According to this assumption, the input feature with a p-value greater than the threshold level is eliminated. In Equation (1), this null hypothesis statement is given as:

$$H_0: \mu \geq \alpha \tag{1}$$

Alternative Hypothesis: This hypothesis states that a strong relationship exists between the input feature and the output feature when the p-value of the input feature is less than the threshold values. According to this assumption, the input feature with a p-value lower than the threshold level is selected. In the following Equation (2), the alternative hypothesis is given as:

$$H_a: \mu < \alpha \tag{2}$$

Where H_0 is a Null Hypothesis, H_a is an Alternate Hypothesis, μ is a p-value of an input feature, and α is the threshold value. The RFE algorithm utilizes logistic regression to find the p-value of the input features to prove the alternative hypothesis claim by rejecting the null hypothesis statement. Logistic regression uses a logit function that is a form of statistical model [19] to determine the relationship between the selected input features and the output features by measuring the logarithm of odds as in Equation (3).

$$logit(p_i) = log\left(\frac{prob}{1-prob}\right) = \beta_0 + \beta_i p_i \tag{3}$$

Where *prob* is the probability of selected input features, p_i is the input feature, logistic regression parameters are β_0 and β_i. Algorithm 1 illustrates the RFE algorithm.

Algorithm 1. Recursive Feature Elimination

Input: data set D_s which consists of N training samples $D_s = ((p_1, q_1), (p_2, q_2), ..., (p_N, q_N))$ and $p_i \in P$; $q_i \in Q$ are the corresponding class labels of D_s associated with p_i. Value of $Q \in \{1, 0\}$.

1: Assign Threshold value = 0.05.

2: State the Null Hypothesis and Alternate Hypothesis.

3: Load the dataset D_s with all input features.

4: Calculate the *p*-value of each input feature using logit function.

5: Reject the alternative hypothesis if the *p*-value of the selected input feature is greater than or equal to α, and remove that feature from full feature set D_s.

6: Iterate the step 4-step 5 till getting the significant features with p-value lower than α.

Output: $D_{relevant}$ → Selected input features of D_s for training and validating the ABGNB classifier.

C. Classification Model Using Adaptive Boosted Gaussian Naïve Bayes Algorithm

Classification modeling is the next step in this proposed method. This process starts with a target dataset, which contains relevant input features obtained using the RFE algorithm. The dataset has been divided into three segments: 80% data for the training phase, 10% for the validation phase, and 10% for the testing phase. This classification modeling has two stages, which are the model training phase and the model validation phase. The ABGNB algorithm is needed to train the HDDSS. The training phase entails developing an HDDSS via learning the training algorithm (ABGNB) parameters and training dataset. Model validation is the second phase of this classification model; training results are evaluated during this phase using a validation dataset for tuning the AGBNB classifier hyper-parameters for improving efficiency and minimize the loss function of the ABGNB classifier. A test set is used to evaluate the final prediction model's working capacity using different classifier performance metrics.

1. Training Phase of ABGNB Classifier

During the training phase, an HDDSS has been developed using an ABGNB classifier. ABGNB is an ensemble of the Adaboost algorithm and Gaussian naive bayes, which outperforms conventional machine learning algorithms in prediction accuracy [20], [21]. The proposed ABGNB classifier utilizes the Adaboost algorithm to improve the prediction efficiency of the gaussian naive bayes classifier.

The Adaboost model has been trained with bootstrapped samples of the training dataset and the gaussian naive bayes algorithm. Bootstrap

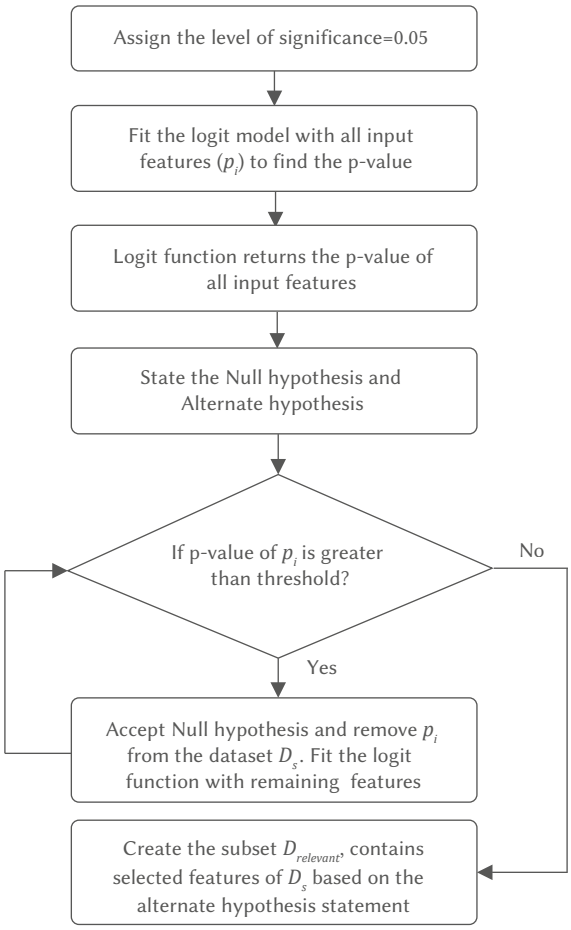

Fig. 2. Workflow of Recursive Feature Elimination Algorithm.

sampling avoids overfitting and improves the performance of training algorithms [22]. The Adaboost methodology aims to provide a correct prediction for each training instance in each iteration by training the gaussian naive Bayes classifier with differently weighted training samples.

During the first iteration of the training process, all samples in the training set are assigned the same weight, and the accuracy of the training phase will be measured after the first iteration. The weight of the misclassified samples is updated during the second iteration of the training phase to motivate the incorrect prediction in the training set, and the process continues for each iteration (s_i) of the training phase. Consequently, the Adaboost classifier constructs a linear hypothesis $h_t: P \rightarrow Q$ by the ensemble of weak hypothesis generated in iteration s_i. The resulting linear hypothesis h_t minimizes the misclassification rate by correctly classifying the given sample p_i according to the class label q_i Adaboost generates a final hypothesis by linearly combining the weak hypothesis $h_1, h_2, \ldots h_t$ for T steps and minimizing the weighted error of all training samples.

In the ABGNB framework, Gaussian Naïve Bayes (GNB) algorithm is used as a base estimator in the Adaboost classifier to calculate the class membership probability of an input sample ($Prob(p_i \mid q)$) using Gaussian probability density function which is given in the Equation (4).

$$GPDF(p_i, \mu_{classi}, \sigma_{classi}) = \frac{1}{\sigma\sqrt{2\pi}}\, exp^{-\frac{(p_i-\mu)^2}{2\sigma^2}} \tag{4}$$

So that $Prob(p_i \mid q)$ is calculated as in Equation (5)

$$Prob(p_i \mid q) = GPDF(p_i, \mu_{classi}, \sigma_{classi}) \tag{5}$$

Where $GPDF(p_i, \mu_{classi}, \sigma_{classi})$ is the gaussian probability density function of an input sample p_i; π is the mathematical constant value, σ and μ are the standard deviation and mean value of input features for each class label, exp is the mathematical constant. μ_{classi} is a mean value of input features of each class label that can be calculated using Equation (6).

$$\mu_{classi} = {}^1/_N * \sum_{i=1}^{N} p_i \tag{6}$$

Where N is the total training samples and p_i is the input value of training dataset D. σ_{classi} is a standard deviation value of each class label that can be calculated using Equation (7).

$$\sigma_{classi} = \sqrt{\frac{\sum(p_i - \mu)^2}{N}} \tag{7}$$

This ensemble of gaussian naïve bayes and Adaboost classifier (ABGNB) has more advantages as an increase in prediction accuracy and reducing overfitting problem over traditional ML algorithms. Fig.3 shows the graphical illustration of the ABGNB classifier.

2. Hyperparameter Tuning Phase of ABGNB Classifier

It is a validation phase used for tuning the hyperparameters of the ABGNB algorithm. The ABGNB hyperparameters are tuned using the grid search optimizer [23]. This grid search optimizer selects the best hyperparameter values from the hyperparameter search space. Hyperparameters are parameters used by machine learning classifiers to monitor and regulate the classifier's learning process. Tuning the hyperparameter of the classifier helps to improve the classifier prediction accuracy [24]. Grid search builds and evaluates a model for every combination of hyperparameters provided. Grid search finds the best ABGNB hyperparameter from the hyperparameter search space, and then the model is retrained with the new parameters. The validation dataset is used to measure the model's accuracy after the hyperparameters are tuned.

The generic hyperparameter statement of ML classifiers are defined below: Consider y as an ML algorithm with a M hyper-parameters (H). The hyperparameter search space of ML classifier is denoted as $H = \mathrm{H}\, H_1 \times H_2 \ldots . \times H_M$. A grid search method was used to optimize the ABGNB classifier's accuracy in this proposed HDDSS, as shown in Equation (8).

$$ABGNB\ Perf = argmax_p\ f(ABGNB, H, D_v) \tag{8}$$

Where *ABGNB Perf* returns the set of optimised hyperparameters which maximise ABGNB classifier efficiency, H denotes the hyperparameters of ABGNB classifier, D_v denotes the validation dataset, $argmax_p f$ is the grid search optimization function on *ABGNB*, H, D_v to maximise the accuracy score of training and hyperparameter tuning phase. Algorithm 2 describes the steps of the ABGNB algorithm in detail.

IV. Experimental Results and Discussions

The various experimental analysis is performed on the proposed HDDSS to demonstrate the efficiency of the Adaptive Boosted Gaussian Naïve Bayes Classifier with Recursive Feature Elimination.

A. Characteristics of Input Dataset

The proposed ABGNB and other ML classifiers are trained and validated using the UCI heart disease dataset. The characteristics of the input dataset are specified in Table I. The input features of the UCI dataset are mentioned in Table II.

Algorithm 2. Adaptive boosted Gaussian Naïve Bayes

ABGNB Input: Training data set D_s which consists of N training samples $D_s = ((p_1, q_1), (p_2, q_2), ..., (p_N, q_N))$ and $p_i \in P$; $q_i \in Q$ are the corresponding class labels of D_s associated with p_i. Value of $Q \in \{1,0\}$.

ABGNB Parameters:

- *base_learner*: Gaussian Naïve Bayes is used to train the Adaboost Classifier.
- $t = 1, 2, ..., T$ iterations
- *est_gnb*: Number of weak learners to be generated in each iteration
- *lr*: learning rate

ABGNB Output: $hyp_{op}(p)$ is a final hypothesis with improved classification performance.

1: Load $D_s = ((p_1, q_1), (p_2, q_2), ..., (p_N, q_N))$

2: Initialize weight of data sample (p_i):

$$w(p_i) = \frac{1}{N}$$

for all p_i, $i = 1, 2, ..., N$; N is a total training samples

3: for t = 1 to T do

4: for *est_gnb* =1 to T do

5: Generate a vector r with initial weight w_i

6: Apply bootstrap sampling on D_s to create a subset called s_m

7: Calculate likelihood of feature subset using Gaussian probability density function called $GPDF(s_m, r)$

8: Build weak hypothesis of gaussian naïve bayes model $hyp_{gnb}(p_i)$ using majority voting scheme

9: end *est_gnb*

10: Get Weak hypothesis $hyp_{gnb}(p_i) \rightarrow \{1, 0\}$ with error rate

$$e_w = \frac{sum\left(w_i * terror(i)\right)}{sum(w_i)}$$

where e_w is a weighted sum of an error rate, w_i is the weight for each training sample i, *terror* is the prediction error for training sample i

11: Update the weight of incorrect samples for $i = 1, 2, ..., N$ in each subset s_m

$$s_{m+1}(i) = \frac{w_i * exp\left(-a_w q_i hyp_{gnb}(p_i)\right)}{z_w}$$

where w_i is the weight of specific training sample, z_w is the normalization constant; a_w is the parameter to increase the generalization of abgnb classifier

12: end for t

13: Output the final hypothesis $hyp_{op}(p)$:

$$hyp_{op}(p) = \text{SIGN}\left(\sum_{1}^{T} a_w hyp_{gnb}(p)\right)$$

14: Calculate *training error* of ABGNB using D_v where D_v is a validation dataset;

15: Use grid search for selecting optimal hyperparameter of ABGNB from hyperparameter search space and retrain the model with optimized parameters.

16: If Validation error > Training error, Stop the retrain

17: Calculate performance of final model $hyp_{op}(p)$ using test dataset

TABLE I. Characteristics of UCI Heart Disease Dataset

Dataset	Number of input Attributes	Number of Classes in output attribute	Number of Samples
UCI heart disease dataset	13	2	303

TABLE II. Input Features of UCI Heart Disease Dataset

S.No	Feature Code	Description of features
1	AGE	The individuals' age
2	GEN	The gender of an individual
3	CP	The chest pain type of an individual
4	RBP	The resting blood pressure value
5	CHOL	The serum cholesterol
6	FBS	An individual's fasting blood sugar value
7	RESTECG	ECG resting value
8	MAXHR	Maximum heart rate achieved
9	EIA	Exercise included angina
10	OPK	Old Peak Value
11	PESS	Peak exercise ST segment
12	CF	Number of major vessels colored by fluoroscopy
13	THAL	The thalassemia

B. Performance Evaluation Measures of the Proposed Model

This HDDSS uses many classification performance metrics. Almost all evaluation measures of this proposed work are based on a Confusion matrix. This matrix assesses the classifier performance via four components named True Positive (*tp*), True Negative (*tn*), False Positive (*fp*), and False Negative (*fn*). True positive is a correctly labeled positive sample, True negative are the correctly labeled negative samples, False positive are falsely labeled negative samples, and False Negative is falsely labeled positive sample. The components of the confusion matrix (*cm*) have given in Equation (9) below.

$$cm = \begin{bmatrix} tp & fp \\ fn & tn \end{bmatrix} \tag{9}$$

Equations (10)-(17) define seven classification measures for evaluating the HDDSS model. Classifier Accuracy (Acc) is the overall effectiveness of the classifier. Misclassification rate (MCR) is the total number of incorrect predictions in the training sample. Sensitivity (Sen) refers to the number of positives that were supposed to be positive. Specificity (Spe) is the percentage of samples correctly labeled as a negative compared to the total negative samples. The number of true positives divided by the total number of true positives and false positives equals precision (Pre). F- Score (FS) is the weighted harmonic mean of the test's sensitivity and precision.

$$Acc = \frac{tp + tn}{tp + tn + fp + fn} \tag{10}$$

$$MCR = \frac{fp + fn}{tp + tn + fp + fn} \tag{11}$$

$$Sen = \frac{tp}{tp + fn} \tag{12}$$

$$Spe = \frac{tn}{tn + fp} \tag{13}$$

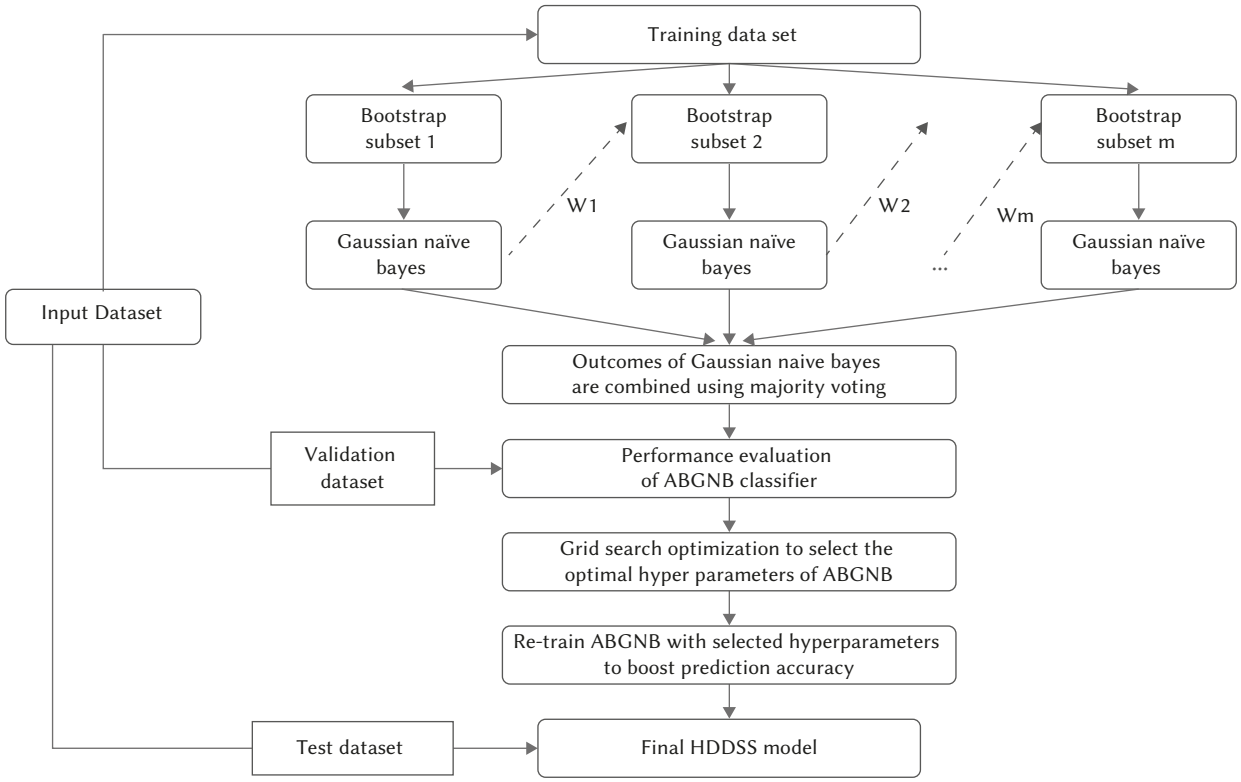

Fig. 3. Graphical Representation of Adaptive Boosted Gaussian Naïve Bayes Algorithm.

$$Pre = \frac{tp}{tp + fp} \tag{14}$$

$$FS = \frac{2}{\frac{1}{Recall} + \frac{1}{Pre}} \tag{15}$$

The recall is known as sensitivity in equation (15). The receiver operating characteristic curve (ROC) has been evaluated using the True positive rate (*tpr*) and false-positive rate (*fpr*). The best classifier shows a ROC value of 1, and the worst classifier shows a ROC value below 0.5. The equation of ROC has given below:

$$tpr = \frac{tp}{tp + fn} \tag{16}$$

$$fpr = \frac{fp}{fp + tn} \tag{17}$$

C. Experimental Result of Recursive Feature Elimination

HDDSS uses the Recursive Feature Elimination (RFE) for selecting the relevant input features for predicting heart disease. This algorithm uses a threshold value (alpha value) of 0.05 to choose the relevant input features of the heart disease dataset. The alpha value has been compared to the p-value of the input feature. If the p-value of the input attribute is less than the alpha value, then it is considered an optimal feature for predicting heart disease, so RFE selects it; otherwise, it will reject. Table III displays the p-values of the input features of the heart disease dataset. This table illustrates how the RFE algorithm selects optimal input features of the UCI heart disease dataset based on the p-value of the input features.

RFE selected seven input features from the heart disease dataset for classification modeling based on their p-values. Table IV displays the selected input features of the input dataset based on the RFE.

TABLE III. The RFE Algorithm Selects Optimal Input Features of the UCI Heart Disease Dataset Based on the P-Value of the Input Features

Feature Code	p-value	Compared with alpha -value (0.05)	Select / Reject the Feature
AGE	0.832	Greater than alpha	Rejected
GEN	0.000	Less than alpha	Selected
CP	0.000	Less than alpha	Selected
RBP	0.060	Greater than alpha	Rejected
CHOL	0.221	Greater than alpha	Rejected
FBS	0.947	Greater than alpha	Rejected
RESTECG	0.181	Greater than alpha	Rejected
MAXHR	0.026	Less than alpha	Selected
EIA	0.017	Less than alpha	Selected
OPK	0.012	Less than alpha	Selected
PESS	0.098	Greater than alpha	Rejected
CF	0.000	Less than alpha	Selected
THAL	0.002	Less than alpha	Selected

The efficiency of the proposed RFE feature selection algorithm has been compared to other feature selection methods, such as sequential forward selection (SFS) [25], sequential backward elimination (SBE) [25], univariate feature selection (UFS) [26]. Table V shows the feature selection parameters, such as the objective function and the number of features selected, for different feature selection methods like RFE, SFS, SBE, and UFS algorithms. All of the above feature selection methods are used for choosing the best feature from a heart disease dataset, and the selected attributes are utilized to build an HDDSS; Table VI displays an output assessment of the RFE with other ML feature selection methods. It is evident from Table VI that the suggested RFE performs better than the other ML feature selection algorithms for the UCI heart disease dataset.

TABLE IV. Optimal Input Features of a Heart Disease Dataset Based on the RFE Algorithm

Dataset	Input features selected by RFE for developing HDDSS
UCI heart disease dataset	GEN, CP, MAXHR, EIA, OPK, CF, THAL

TABLE V. Parameter Setting of Various Machine Learning Feature Selection Algorithms

Feature selection Algorithm	Parameter	Value
SFS	Objective function	Gini Index
		Entropy
	Number of features selected	7
SBE	Objective function	Gini Index
		Entropy
	Number of features selected	7
UFS	Objective function	Chi-square
	Number of features selected	7
Proposed RFE	Objective function	Logistic regression
	Number of features selected	7

TABLE VI. Output Assessment of the RFE With Other ML Feature Selection Methods for Heart Disease Dataset Utilizing the Proposed ABGNB Classifier

Feature Selection	Acc	MCR	Sen	Spe	Pre	FS	ROC
SFS + Gini Index	75.18	24.82	75.62	80.63	75.5	74.94	75.59
SBE + Gini Index	83.92	16.08	83.61	80.35	83.4	84.74	83.19
SFS + Entropy	80.49	19.51	80.3	83.46	80.63	79	80.32
SBE + Entropy	80.34	19.66	80.56	73.14	80.67	81.64	80.64
UFS + Chi-square	88.49	11.51	88.76	86.31	85.01	88.34	88.69
Proposed RFE	92.87	7.13	93.45	90.76	91.64	92.08	92.42

D. Experimental Result of Classification Modeling

In this phase, the selected input features of the heart disease dataset fed into the ABGNB classifier along with the target feature. Repeated 10 × 5 stratified cross-validation is applied during the validation process to build a generalized classifier on an independent dataset and avoiding over-fitting problems. This process is where the 10-fold cross-validation has been repeated five times, in which the data samples being shuffled during each repetition, providing a different split of the given data. The grid search optimizer then adjusts the ABGNB classifier hyperparameters to improve the efficiency of the trained model. The ABGNB classifier's hyperparameter range is described in Table VII.

The classification performance of the ABGNB classifier is compared with other Machine Learning models, namely, Naïve bayes (NB) [27], K-Nearest neighbor (KNN) [28], Support vector machine (SVM) [29], Decision tree (DT) [30]. Table VIII shows the output of the proposed ABGNB classifier with other conventional ML classifiers on the heart disease dataset before using the RFE algorithm.

Table IX shows the output of the proposed ABGNB classifier compared to other ML classifiers on the heart disease dataset after implementing the RFE feature selection method. It's worth noting that the proposed ABGNB classifier performs well on the heart disease dataset and has a high accuracy score of 92.87%.

TABLE VII. Hyper Parameter Search Space of ABGNB Classifier

Proposed Classifier	Hyperparameter	Hyperparameter configuration space	Selected Hyper parameters by grid search
ABGNB Classifier	No. of weak learner	[10, 50, 100, 500]	500
	Learning rate	[0.0001, 0.001, 0.01, 0.1, 1.0]	0.1
	Random state	[50,30,40]	40

TABLE VIII. Performance Comparison of ABGNB Classifier With Other ML Algorithms on Heart Disease Dataset Before Using the Suggested Recursive Feature Selection

Classifier	Acc	MCR	Sen	Spe	Pre	FS	ROC
NB	86.31	13.69	80.52	93.31	92.45	86.74	86.41
KNN	75.23	24.77	80.41	70.54	73.15	76.75	75.78
SVM	70.42	29.58	87.27	53.97	65.39	75.04	70.57
Decision Tree	80.49	19.51	83.65	76.48	78.17	81.37	80.37
Proposed ABGNB	90.12	9.88	87.56	93.14	93.45	90.67	90.47

TABLE IX. Performance Comparison of Proposed ABGNB Classifier With Other Typical ML Algorithms on Heart Disease Dataset After Applying the Suggested Recursive Feature Selection

Classifier	Acc	MCR	Sen	Spe	Pre	FS	ROC
NB	90.5	9.5	85.86	95.17	95.49	90.34	80.19
KNN	80.31	19.69	85.17	74.5	78.9	82.04	60.14
SVM	73.16	26.84	87.06	58.76	70.37	77	48.37
Decision Tree	79.24	20.76	70.46	86.27	85.31	77.52	57.06
Proposed ABGNB	92.87	7.13	93.45	90.76	91.64	92.08	84.67

Table X compares the performance of the proposed HDDSS to the output of other heart disease prediction models. The purpose of this analysis (Table X) is to show how the proposed classifier, the adaptive boosted Gaussian Naive Bayes classifier, outperforms previous studies in terms of prediction accuracy.

TABLE X. The Output of the Proposed HDDSS Compared to Other Heart Disease Prediction Models Output

Author(s)	Year	Method	Highest Accuracy (in %)
Haq et al. [5]	2018	Relief + Logistic Regression	89
Otoom et al. [6]	2015	Best first search + BayesNet	84.5
Mohan et al [7]	2019	Hybrid random forest with a linear model	88.7
Li, Jian Ping et al [8]	2020	FCMIM -SVM	92.37
Chen et al. [9]	2011	Learning vector quantization	80
Hidayet et al. [10]	2018	Linear kernel SVM + ReliefF	84.81
Thippa Reddy et al. [11]	2017	Rule Based Fuzzy Logic Model	78
David et al. [31]	2018	Random Forest	81
Das et al. [32]	2020	K-NN	86.84
Apurv Garg et al. [33]	2021	K-NN	86.88
Proposed Method	-	RFE + ABGNB	92.87

E. Discussions

The proposed research aimed to demonstrate that the RFE and ABGNB classifiers could reliably predict heart disease. The RFE is being used to select relevant input features from the heart disease dataset. RFE selects optimal features using the p-value of an input feature. Table III shows how the p-value of the input feature is utilized to choose the best input features. RFE selected seven relevant input features for building the proposed HDDSS, as shown in Table IV. The efficiency of the RFE has been compared to that of other feature selection methods such as sequential forward, sequential backward, and univariate feature selection in the proposed framework. The performance of these feature selection methods has been illustrated in Table VI. Table VI shows that the RFE is the best feature selection procedure for the UCI heart disease dataset, exceeding other feature selection strategies in identifying the best features from the input dataset. RFE outperformed other ML feature selection methods in terms of recall, precision, F-score, and ROC shows, resulting in improved classification performance. HDDSS has been developed using the proposed ABGNB classifier with the relevant input features and the target feature. The performance of the ABGNB has been compared to that of other machine learning models such as naive Bayes, KNN, SVM, and Decision tree classifiers. Tables VIII and IX illustrate that the proposed ABGNB classifier outperforms other ML classifiers before and after using the suggested RFE approach on the heart disease dataset. The prediction accuracy of the ABGNB algorithm was 90.12 percent before using the RFE algorithm and 92.87 percent after using the RFE algorithm, meaning that the RFE algorithm enhanced the ABGNB algorithm's prediction accuracy by 2 to 3 percentage points. Table IX confirms that the ABGNB classifier has the best accuracy of 92.87 percent, meaning that ABGNB can differentiate between positive and negative samples. Table IX also shows that the proposed ABGNB classifier outperforms other ML classifiers on different performance metrics for the heart disease dataset. Table X demonstrates that the proposed heart disease prediction model (HDDSS) outperforms most of the present literature for increasing prediction accuracy. According to this study, the proposed HDDSS outperformed other heart disease prediction models and is suitable for assessing the risk of heart disease in a patient.

V. Conclusion

This research aims to build an enhanced heart disease decision-support system for the prediction of heart disease. This automated diagnosis system has been experimented on the UCI heart disease dataset. This proposed HDDSS utilizes the Recursive feature elimination method for selecting the most relevant input features of the heart disease dataset. The ABGNB classifier ensembled with a grid search optimizer for enhancing the prediction accuracy of HDDSS. The experiment result illustrates that the RFE+ABGNB method gives better performance than other compared ML models on the heart disease dataset. This suggested method achieves 92.87% prediction accuracy with 93.45% sensitivity on the UCI heart disease dataset. The analysis of the proposed system showed that the proposed ABGNB with recursive feature elimination provides better heart disease prediction performance on the UCI heart disease dataset. The proposed model's efficiency can be improved even more using an automated regularization technique.

References

[1] Anand, Sonia S., Shofiqul Islam, Annika Rosengren, Maria Grazia Franzosi, Krisela Steyn, Afzal Hussein Yusufali, Matyas Keltai, Rafael Diaz, Sumathy Rangarajan, and Salim Yusuf, "Risk factors for myocardial infarction in women and men: insights from the INTERHEART study," European heart journal, vol. 29, no. 7, pp. 932-940, 2008.

[2] Frohlich, Edward D., and Patrick J. Quinlan, "Coronary heart disease risk factors: public impact of initial and later-announced risks," The Ochsner Journal, vol. 14, no. 4, 532-537, 2014.

[3] Wah, Teh Ying, Ram Gopal Raj, and Uzair Iqbal, "Automated diagnosis of coronary artery disease: a review and workflow," Cardiology research and practice, 2018.

[4] Ali, Liaqat, Atiqur Rahman, Aurangzeb Khan, Mingyi Zhou, Ashir Javeed, and Javed Ali Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on χ^2 Statistical Model and Optimally Configured Deep Neural Network," IEEE Access, vol. 7, pp. 34938-34945, 2019.

[5] Haq, Amin Ul, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mobile Information Systems, 2018.

[6] Otoom, Ahmed Fawzi, Emad E. Abdallah, Yousef Kilani, Ahmed Kefaye, and Mohammad Ashour, "Effective diagnosis and monitoring of heart disease," International Journal of Software Engineering and Its Applications, vol. 9, no. 1, pp. 143-156, 2015.

[7] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542-81554, 2019.

[8] Li, Jian Ping, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, and Abdus Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, vol. 8, pp. 107562-107582, 2020.

[9] Chen, Austin H., Shu-Yi Huang, Pei-Shan Hong, Chieh-Hao Cheng, and En-Ju Lin. "HDPS: Heart disease prediction system." In 2011 computing in cardiology, IEEE, 2011, pp. 557-560.

[10] Takci, Hidayet, "Improvement of heart attack prediction by the feature selection methods," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 26, no. 1, pp. 1-10, 2018.

[11] Reddy, G. Thippa, and Neelu Khare, "An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model," Journal of Circuits, Systems and Computers, vol. 26, no. 04, 2017.

[12] Hossin, Mohammad, and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, 2015.

[13] Janosi A, Steinbrunn W, Pfisterer M, Detrano R. Heart Disease UCI. Available online: https://www.kaggle.com/ronitf/heart-disease-uci (Cited on 23 Jan 2021)

[14] Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. "A survey of feature selection and feature extraction techniques in machine learning." In 2014 Science and Information Conference, IEEE, 2014, pp. 372-378.

[15] Miao, Jianyu, and Lingfeng Niu, "A survey on feature selection," Procedia Computer Science, vol. 91, pp. 919-926, 2016.

[16] Guyon, Isabelle, and André Elisseeff, "An introduction to variable and feature selection," Journal of machine learning research, vol. 3, no. Mar 2003, pp. 1157-1182, 2003.

[17] Massey, Adam, and Steven J. Miller. "Tests of hypotheses using statistics." Mathematics Department, Brown University, Providence, RI 2912 2006, pp. 1-32.

[18] Murphy, Kevin P. "A Probabilistic Perspective." Text book (2012).

[19] Hoffman, J. I. E. "Logistic regression." Basic Biostatistics for Medical and Biomedical Practitioners, 2019, pp. 581-589.

[20] Perez, Aritz, Pedro Larranaga, and Inaki Inza, "Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes," International Journal of Approximate Reasoning, vol. 43, no. 1, pp. 1-25, 2006.

[21] Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." In icml, vol. 96, 1996, pp. 148-156.

[22] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." In Ijcai, vol. 14, no. 2, 1995, pp. 1137-1145.

[23] Syarif, Iwan, Adam Prugel-Bennett, and Gary Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," Telkomnika, vol. 14, no. 4, pp. 1502-1509, 2016.

[24] Wu, Jia, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, and Si-Hao Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," Journal of Electronic Science and Technology, vol. 17, no. 1, pp. 26-40, 2019.

[25] Kumar, Vipin, and Sonajharia Minz, "Feature selection: a literature review," SmartCR, vol. 4, no. 3, pp. 211-229, 2014.

[26] Jović, Alan, Karla Brkić, and Nikola Bogunović. "A review of feature selection methods with applications." In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE, 2015, pp. 1200-1205.

[27] Rish, Irina "An empirical study of the naive Bayes classifier." In IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, 2001, pp. 41-46.

[28] Tan, Songbo, "An effective refinement strategy for KNN text classifier," Expert Systems with Applications, vol. 30, no. 2, pp. 290-298, 2006.

[29] Cortes, Corinna, and Vladimir Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273-297, 1995.

[30] Safavian, S. Rasoul, and David Landgrebe, "A survey of decision tree classifier methodology," IEEE transactions on systems, man, and cybernetics, vol. 21, no. 3, pp. 660-674, 1991.

[31] David, H. B. F., and Belcy, S. A., "Heart Disease Prediction Using Data Mining Techniques," ICTACT Journal on Soft Computing, vol. 9, no. 1, 2018.

[32] Das, Sumit, Manas Kumar Sanyal, and Sourav Kumar Upadhyay. "A Comparative Study for Prediction of Heart Diseases Using Machine Learning." Available at SSRN 3526776, 2020.

[33] Garg, A., Sharma, B. and Khan, R., "Heart disease prediction using machine learning techniques." In IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2021: Vol. 1022, No. 1, p. 012046.

M. Dhilsath Fathima

She received her B.E degree in CSE from Anna University, Tamil Nadu, India in 2005, Master's degree in CSE from Sathyabama Institute of Science & Technology, Tamil Nadu, India in 2011. Now, pursuing Ph.D in Computer science and Engineering Department from Sathyabama Institute of Science & Technology, Tamil Nadu, India. Since 2007, she is working as an Assistant Professor in Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Tamil Nadu, India. Her major research areas are Machine learning, Data mining, Deep learning etc.

Dr. S. Justin Samuel

Dr. S.Justin Samuel received his Ph.D in Computer Science and Engineering from Sathyabama University, India. His area of interest includes Data Mining, Wireless Sensor networks, and Image processing. He has published more than 25 research papers in International & National Journals and Conferences. He is a professor at PSN Engineering college for Department of Science and Technology, India.

S. P. Raja

S. P. Raja was born in Sathankulam, Tuticorin District, Tamilnadu, India. He completed his schooling in Sacred Heart Higher Secondary School, Sathankulam, Tuticorin, Tamilnadu, India. He completed his B. Tech in Information Technology in the year 2007 from Dr. Sivanthi Aditanar College of Engineering, Tiruchendur. He completed his M.E. in Computer Science and Engineering in the year 2010 from Manonmaniam Sundaranar University, Tirunelveli. He completed his Ph.D. in the year 2016 in the area of Image processing from Manonmaniam Sundaranar University, Tirunelveli. His area of interest is image processing and cryptography. He has more than 14 years of teaching experience in engineering colleges. He currently works as an Associate Professor in the Department of Computer Science and Engineering at the Vellore Institute of Technology, Vellore. He has published 48 papers in International Journals, 24 in International conferences and 12 in national conferences. He is an Associate Editor of the International Journal of Interactive Multimedia and Artificial Intelligence, International Journal of Image and Graphics and International Journal of Biometrics.