Journal of Intelligent Systems and Internet of Things
Vol. 13, No. 01, PP. 177-195, 2024
DOI: https://doi.org/10.54216/JISIoT.130114

An Ensemble Machine Learning Method for Analyzing Various Medical Datasets

Chhaya Gupta [1], Nasib Singh Gill [2], Priti Maheshwary [3], Shraddha V. Pandit [4], Preeti Gulia [5], Piyush Kumar Pareek [6]

[1, 2, 5] Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, Haryana, India
[3] Rabindranath Tagore University, Bhopal, India
[4] Department of Artificial Intelligence and Data Science, PES Modern College of Engineering, Shivajinagar, Pune, India
[6] Professor and Head Department of AIML and IPR Cell, Nitte Meenakshi Institute of Technology, Bengaluru, India

Emails: chhaya.rs.dcsa@mdurohtak.ac.in; Nasib.gill@mdurohtak.ac.in; pritimaheshwary@gmail.com; shraddha.pandit@moderncoe.edu.in; preeti@mdurohtak.ac.in; piyush.kumar@nmit.ac.in
Corresponding Author: Piyush Kumar Pareek, Email: piyush.kumar@nmit.ac.in,

Abstract

In recent years, machine learning (ML) has shown a significant impact in tackling various complicated problems in different application domains, including healthcare, economics, ecological, stock market, surveillance, and commercial applications. Machine Learning techniques are good enough to deal with a wide range of data, uncover fascinating links, offer insights, and spot trends. ML can improve disease diagnosis accuracy, predictability, performance, and reliability. This paper reviews various machine learning techniques applied to different medical datasets and proposes an ensemble method for helping in the early diagnosis of different diseases. The study compares existing machine learning techniques with the proposed ensemble method. The ensemble method uses the AdaBoost algorithm to combine the traits of choice trees, random forests, and support vector machines. Three feature selection techniques, Fisher's score, information gain, and genetic algorithm, are used to select appropriate dataset features. The ensemble method also uses the K-fold cross-validation technique (where k=15) for validating results. SMOTE was employed to balance some of the datasets because they were quite unbalanced. All the methods used in this study are evaluated based on accuracy, AU Curve, Recall, Precision, and F1-score. The paper uses different medical datasets at the University of California Irvine and the Kaggle directory to compare machine-learning models with the proposed ensemble method. The encouraging results show that the ensemble method outperforms the existing machine-learning techniques. The paper thoroughly analyzes how machine learning is used in the medical industry, covering established technologies and their impact on medical diagnosis. An early diagnosis is needed to prevent people from deadly diseases. Hence, this study proposes an ensemble method that may be used to diagnose different diseases early.

Received: September 22, 2023 Revised: January 19, 2024 Accepted: June 13, 2024

Keywords: Choice Tree Classifier; Ensemble Classifier; KNN Classifier; Naïve Bayes Classifier; Random Forest Classifier; Synthetic Minority Oversampling Technique

1. Introduction

Machine Learning (ML) can aid in diagnosing medical conditions in various fields like wearable sensors, cancer detection, and medical imaging [1]. Machine learning is employed for crucial therapeutic tasks like obtaining medical information and predicting diseases and developmental stages. As a result, it helps in supporting the patient's condition. Moreover, it supports data analysis and provides intelligent alerts for effective healthcare monitoring.

Early disease detection is crucial in the field of diagnosis and prognosis. Electronic health records (EHRs) store the vast amount of medical data created daily for study. ML can automatically evaluate data from patient records recorded in PHRs (Patient Health Records) [2]. The ML models help to discover diseases by monitoring the data and predicting likely illnesses using the relevant EHR parameters [3]. Precise information about patients is necessary for the learning algorithm to function. Encoding is a simple procedure that automatically analyses the data and compares it with similar issues previously resolved. As a result, it helps the doctor diagnose new cases precisely. Non-specialists and students use it to analyze patients. ML also helps diagnose medical images nowadays to improve the field of healthcare [4], [5].

The outlook of the machine learning approach is to make predictions about a person having a particular disease at an early stage. Machine learning methods help physicians diagnose their patients effectively and promptly prescribe proper medication. Patient care, therapy support, knowledge extraction from the medical profession, and sickness progression prediction all use machine learning. Medical reasoning is essential for intelligent systems. This system uses historical patient data to predict expert systems. Machine learning techniques describe attributes that specifically identify the patients' medical issues when clinical datasets are available. Expert systems can preserve and use the patient's history to provide information about each patient uniquely and describe the patient's health condition.

Medical image diagnosis with the help of computer-based interpretations is an important field. These studies have played an essential role in diagnosing cancer tissues as malignant or benign. The research compares machine-learning algorithms on several medical datasets, including Naive Bayes, KNN, SVM, Choice trees, and Random Forests, with the proposed ensemble method [6]. This research suggests an ensemble method that combines SVM, random forest, and decision tree features with a boosting algorithm and a validation technique. The performance of all the methods has been analyzed and compared. The datasets used for the investigation are those for binary classification and are freely accessible through the Kaggle directory and University of California at Irvine (UCI) directory. The main objective and contribution of the paper are:

- The experiment uses three feature selection techniques, Fisher's score, information gain, and genetic algorithm, for selecting relevant features in all the datasets.
- The SMOTE approach balances the heavily unbalanced dataset.
- The performance of all the classifiers is analyzed and compared on reduced feature sets.

The following sections make up the organization of the paper: Section 2 surveys the related literature on recent advancements in the healthcare industry; Section 3 describes machine learning, its categories, and methods; Section 4 explains all the techniques used along with the proposed ensemble method; Section 5 reviews the results and overall performance of all the techniques used; and Section 6 concludes the paper.

2. Related Work

There has been plenty of research on machine learning in recent decades. With improvements in machine learning and continuous technological progress, ML techniques can analyze medical data more effectively. This literature review examines how machine learning has changed recent medical research. By looking at the ML approaches applied in the medical field, one can track the type of shift in medical research. Researchers might also gain information from this survey by reviewing the specific ML techniques applied to particular medical applications over the previous five years. Additionally, the medical business may significantly improve by analyzing and applying this data to suitable applications.

On the WBCD, Chhaya Gupta et al. [9] examined various machine learning approaches and extreme learning machine neural networks to determine the most accurate methodology to be applied going forward for the early diagnosis of this deadly disease.

Maza Ramzan et al. in [7] applied various techniques to medical datasets and compared the results of Naive Bayes, Random Forest, and J48 algorithms based on accuracy, precision, and recall. BV Ramana et al. [8] investigated various classification techniques to evaluate the precision, sensitivity, specificity, and accuracy of the Liver dataset.

Alexopoulos et al. [9] diagnosed stroke by inductive machine learning techniques. According to G. Waibi et al. [10], the accuracy of their chronic kidney disease (CKD) diagnosis was 99% using various ensemble learning models. To create fuzzy rules and incorporate the rules into the knowledge-based system, M. Nilanshi et al. [11] implemented it on a variety of datasets, including the PID dataset, the Wisconsin breast cancer dataset, the Statlog heart dataset, and the Mesothelioma dataset. Ali CÜvitoğlu et al. [12] applied classification techniques on both the datasets of Cryotherapy and immunotherapy and concluded Random Forest to be the best among all other classification techniques.

Ashu Garg et al. [13] presented a literature survey to investigate different machine-learning techniques used on various medical applications and discuss the current machine-learning methods.

Aada et al. [14] used R programming to diagnose Diabetes by using the Pima Indian Diabetes database. They also used bootstrap resembling technique with Naïve Bayes classifier, decision tree classifier, and KNN classifier. Bootstrap resembling is used to sample a dataset using replacement to estimate statistics on a population. It can be estimated using summary statistics like the mean or standard deviation.

Nadakinamani et al. [15] proposed a cardiovascular disease prediction system based on machine learning techniques. Models including the M5P Tree, Naive Bayes, J48, and Random Tree were compared in the study. The outcomes demonstrated random trees outperformed other techniques.

Jabbar et al. put out a Bayes network and radial basis function-based ensemble technique with a 97% accuracy [16] and trained on the Wisconsin Breast Cancer Dataset. Machine learning techniques, including Naive Bayes, SVM, Logistic Regression, Decision Tree, Random Forest, and KNN. According to the study, KNN produces higher outcomes when K=8. A dataset comprising 303 records and 76 characteristics was taken from the UCI library and used to train the model.

Hussein et al. [17] used swarm intelligence algorithms and machine learning techniques for detecting diseases. The paper discussed different swarm intelligence and machine learning techniques for diagnosing diseases. Saboor et al. [18] used nine machine learning classifiers: CART, LDA, MNB, XGB, RF, SVM, LR, AB, and ET. The classifiers were trained and tested using the heart disease dataset. The random forest classifier produced positive findings using the K-fold cross-validation approach, which is used to validate all the classifiers. Floyd et al. [19] implemented an All Convolutional Network (ACN), batch normalization, and ensemble convolutional neural network on the MIT-BIH arrhythmia database. Ensemble convolutional neural networks outperformed the SVM classifier and Random Forest classifier.

Alanazi et al. [20] used various machine learning classification techniques, such as the Lasso regressor, decision tree, SVM, logistic regression, KNN, and Naive Bayes. The article explored reinforcement, supervised, unsupervised, and other forms of machine learning applications in the healthcare domain.

The algorithm used pseudo-data-generation with transfer learning. The CCKS2020 dataset is used to train the suggested model. Mukhopadhyay et al. [21] profounded a framework for the classification of chronic liver disease by improving the prediction accuracy with the help of cutting-edge analysis techniques. The study combined the aforementioned strategies in one algorithm using global and local learner stackers. The proposed algorithm achieved high accuracy of approximately 99%.

Srivastava et al. [22] developed an automated method for diagnosing diseases at an early stage. This study used the fuzzy-logic-based random forest method on fuzzy datasets freely available on the UCI repository. Paidipati et al. [23] performed an image prediction on medical imaging with the help of SVM, KNN, LR, RF, and DT. The LR showed the best results with an accuracy of 95%. Sheetal Singh et al. [24] predicted cardiovascular health with the help of machine learning model namely, KNN, RF, NB, SVM, LR, and DT. They also employed Deep neural network and the hybrid model achieved an

accuracy of 97%. Ahmed N. Masri et al. [25] proposed an efficient machine learning model for diagnosing cervical cancer in females, and the model achieved good results. Medical data is heterogeneous by nature and has other inherent drawbacks that affect how it can be analyzed. ML techniques help in overcoming the above challenges. The detailed literature study examined the practical applications of machine learning in the healthcare industry. The Naive Bayes, KNN, SVM, Choice Tree, and Random Forest approaches make up many machine learning techniques used in the healthcare sector. It also suggests an ensemble approach for making decisions with greater accuracy. Each method has undergone in-depth mathematical analysis. The paper provides a detailed review of all the stated techniques and a comparative study of different datasets from UCI and Kaggle repositories.

3. Machine Learning

Artificial intelligence is known as "machine learning," in which a machine picks up new information. Data is gathered, pre-processed, and categorized in machine learning. New algorithms are created and trained with specific patterns from the collected information, and finally, the algorithm is tested to achieve the desired results. Technically, machine learning creates a model that can perform a specific task without a human. This section discusses the background information on machine learning types, methodologies, and approaches. Both supervised and unsupervised learning strategies are used in machine learning, with feedback being the key difference. The model learns to identify the correct title through supervised learning when it has a set of samples and their accompanying labels. Techniques for classification and regression make up supervised learning. Classification techniques take data as input and classify it into different classes. In unsupervised learning, labels for inputs are undefined. The model tries to learn the similarities between the given inputs and then predicts the labels. Clustering is the technique used in unsupervised machine learning. In clustering, similar inputs are clustered into groups. Numerous ML algorithms use both these methodologies in medical and healthcare applications. The categories of machine learning are all briefly introduced in this section.

A. Supervised Machine Learning

Supervised machine learning is creating an algorithm to understand how to map an input to an output. If the mapping is accurate, the algorithm has been successfully taught. If not, then the algorithm is modified as needed to learn properly. Supervised learning algorithms can aid in predicting upcoming and previously undiscovered data. Supervised learning algorithms are known as student-teacher networks. The teacher guides the students on how to learn from the books. The student is tested, and if successful, then passed. Otherwise, the teacher tunes the students to learn from their previous mistakes. Supervised learning is essential because:

- The algorithm gains knowledge through learning, which it may utilize to produce recommendations for unseen data.
- Additionally, knowledge aids in improving the algorithm's performance.
- Supervised Learning methods can also handle computations in the real world.

Support Vector Machine Classifier (SVMC), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), K-Nearest Neighbor Classifier (KNNC), Naïve Bayes Classifier (NBC), and Ensemble Method Classifier (EMC) are some of the supervised machine learning categorization techniques that are described in the next section.

B. Unsupervised Machine Learning

Consider a device that accepts a sequence as input, such as x1, x2, x3……, where it is input at time t. This input refers to data and can be of any form, like an image, video sequence, pixels, or sound waveform. In unsupervised learning, the machine accepts the sequence of inputs. It is strange to imagine what the machine may learn from the sequence. However, it is possible to construct a framework for an unsupervised learning machine based on the idea that the machine's goal is to categorize input data to uncover patterns that can be used for decision-making. In simple words, in unsupervised learning, the patterns are found in data that can be further used for decision-making and predicting future inputs. Unsupervised learning is divided into two categories: dimensionality reduction and clustering.

Finding patterns and similarities in the input and grouping them into various groups is called clustering. Clustering helps to organize unlabeled data so that similar data types are categorized into clusters. K-means clustering, fuzzy c-means clustering, and class-based clustering are only a few examples of the various clustering techniques [26].

C. Datasets Used

This paper uses various medical datasets to analyze and compare different machine-learning techniques and define how machine learning helps to diagnose diseases early. The datasets used are the Wisconsin Breast Cancer dataset, the Pima Indians Diabetes Database, the Hepatitis dataset, the chronic kidney disease dataset, and the Parkinson's speech dataset. Parkinson's speech dataset is the most imbalanced one. The SMOTE technique is used to balance the dataset. Each dataset used and the number of attributes and instances is briefly described in Table 1. All the medical datasets used are from the UCI and Kaggle machine-learning repositories.

Table 1: Medical Datasets from UCI

Datasets	WBCD	CKD	Cryotherapy	Hepatitis	PIDD	Parkinson's
Attributes	11	25	7	20	9	23
Instances	699	400	90	155	768	195

4. Methodology

Searching adequate data descriptors to anticipate correct classification labels in the medical domain is crucial in machine learning. Medical practitioners have always carried out medical interpretation. Using their knowledge, they create useful descriptions for particular output labels. However, based on vast data collected daily, ML approaches can produce predictions almost as accurate as subject-matter experts' predictions. To help in the early detection of symptoms of cancer treatment, ML models have demonstrated their efficacy in precisely extracting pertinent data from clinical imaging arrays. Medical professionals employ clinical applications of ML for prognosis, diagnosis, image analysis, and treatment.

ML algorithms handle various tasks, such as regression, classification, clustering, and prediction. The proposed ensemble method helps in diagnosing different diseases at a faster pace. The proposed ensemble technique includes class balance, feature selection, and classification, as shown in Fig. 1. Features are chosen using feature selection methods. Fisher's score, information gain, and genetic algorithm are the feature selection techniques used in the study to select appropriate features. Some datasets, including the Parkinson's speech dataset, are balanced using the synthetic minority oversampling strategy, or SMOTE. SMOTE synthetically generates minority samples from the existing samples in a dataset. The following subsection describes various machine learning techniques used on different datasets.

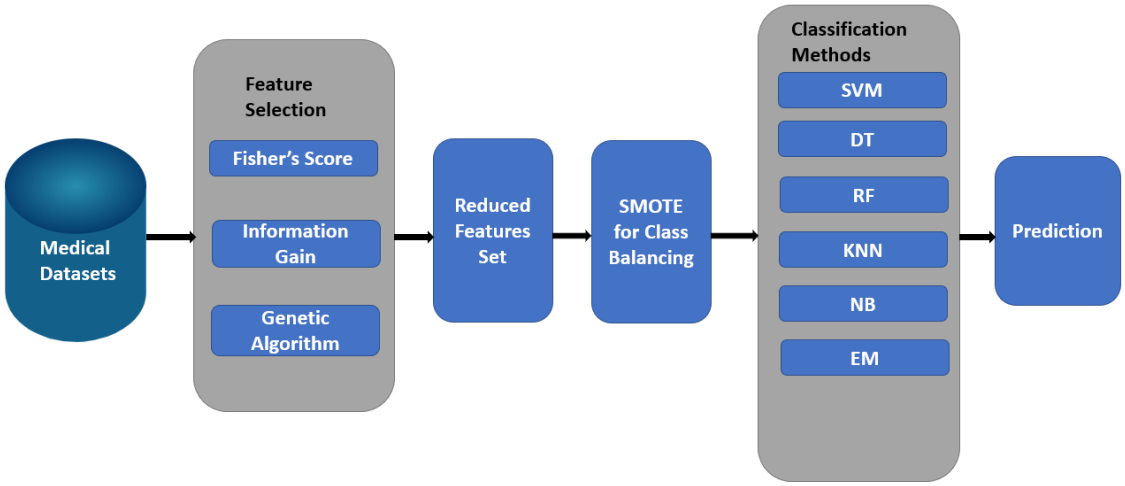


Figure 1: Methodology of the proposed ensemble method

A. Feature Selection

A pre-processing method called feature selection is used to find appropriate features in a dataset—techniques for feature selection aid in reducing the dimension of the data and improving model performance. Three feature selection methods used are described further.

1. Feature Selection using Fisher's Score

One of the methods for supervised feature selection often used is this one. This method chooses features on an independent basis based on their scores. This technique helps to reduce the size of the data. Finding a feature subset, increasing the gap among data points in different classes, and reducing the distance between points in the same class are the three key goals [27]. If there is a given training dataset Y with c classes, then the fisher score of jth features is calculated as:

$$FS(f_j) = \frac{S_b(f_j)}{\sum_{k=1}^{c} s_t^k(f_j)} \quad (1)$$

Where $S_b(f_j) = \sum_{k=1}^{c} t_k(\mu_j^k - \mu_j)^2$ = the number of samples in the kth class, the mean of the jth feature in the kth class, and the mean of the jth feature in the Y training dataset are all included in scatter between classes of the jth feature.

$S_k(f_j) = \sum_{i=1}^{n_k}(x_{ji}^k - \mu_j^k)^2$ = this is the class scatter of the jth feature in the kth class, x_{ji}^k is the value of the jth feature in the ith sample in the kth class.

2. Feature Selection using Information Gain

This method is the filter method for feature selection. This function calculates the target feature's dependence on the chosen feature, then evaluates the information learned about that feature. This technique aids in calculating the entropy decrease caused by a dataset's transformation. The technique helps in identifying the most relevant features. Information gain is calculated as follows:

Step1: Given attribute a and class b, calculate entropy (H) before observation as:

$$the\ H(b) = -\sum p\ (b) log_2 p(b) \quad (2)$$

Where p(b) = probability distribution of observations b

Step2: Calculate entropy (H) after examining attribute a:

$$the\ H(b|a) = \sum p(a) \sum p(b|a) log_2 p(b) \quad (3)$$

Step3: Evaluating information gain (I), which is the difference between entropy before examining attribute a and after examining attribute a, given as:

$$I(a|b) = H(b) - H(b|a) \quad (4)$$

Where I(b|a) = information gain of attribute a in class b.

3. Feature Selection using Genetic Algorithm

Information gain is the simplest feature selection wrapper method. This method analyzes the correct functionality of high-dimensional datasets. The genetic algorithm works on the principle of Darwin's theory of evolution, where the fittest features survive and form the base for the next generation. The general genetic algorithm is defined as follows:

Algorithm 1. Genetic algorithm

Input: Parameters are initialised aPop = p, max, d = 0;

Verify: The optimal feature subset with the highest fitness value.

1: Starting while (d<=max) do

2: Creating aPop, p, dmax; //Creating Initial population

3: For b = 1 to p do //Scoring and scaling the population

4: Select_Parents [p1, p2] = system selection (p, aPop) //Selecting Parents

5: Crossover_Child = XOR [p1, p2] //Crossover

6: Mu = mutation [Crossover_Child] //Mutation

7: End for

8: Replace p with Crossover_Child1, Crossover_Child2, …, Crossover_Childp

9: d = d+ 1;

10: Ending while

11: Storing Highest fitness value;

Where p = population size

GA creates an arbitrary initial population, aPop, and fixes some arbitrary size. The algorithm constructs the sequence of the populations by using current generation features. The algorithm computes the fitness value of each feature of the current population, known as the raw fitness score. Parent features are selected based on their expectations. The features with low fitness scores are called elite and are passed to the next population. Child features are chosen either by mutation or by crossover techniques. The genetic algorithm used in this paper uses the mutation technique. Children eventually replace the present population to create the following generation. When the number of generations approaches the maximum, the algorithm terminates.

B. Machine Learning Classifiers

This section discusses various machine learning classifiers used for analyzing different datasets.

1. Support Vector Machine Classifier (SVMC)

The Support Vector Machine Classifier (SVMC) is a controlled machine learning method to solve binary classification problems. A traditional SVMC model, which employs a hyperplane to split two classes with the least amount of overlap, as shown in Fig. 2, frequently maps the data points into a higher-dimensional space. The goal of quadratic programming is to use a hyperplane that can be linearly split to optimize the gap between the two classes. SVMC is unfamiliar with mapping before this. It chooses the optimum hyperplane separator using a kernel-based dot product in the feature space of the map, as shown in Eq. (5). The value of the weight vector and bias vector make up the dot product for the input vectors x and b. The model employs "Sigmoid," "Radial Basis Function," "Linear," and "Poly" as its kernels. The data values can be visualized in n-dimensional space using this method. The data points are classified into two classes using the linear plane (w*x-b=0), and their placements are determined by their weights and bias.

$$wTx+b = 0 \quad (5)$$

Although intended for binary classifications, support vector machines (SVMs) can also provide non-linear answers. Kernel functions give SVM more for categorizing the data more features. SVMs are effective with datasets with fewer samples and employ a data-driven algorithmic approach; as a result, they are useful for disease prognosis and diagnosis. SVMs are helpful for many medical classification problems like Alzheimer's disease, urinary tract infections, pulmonary hypertension, Parkinson's disease, and thyroid diseases.

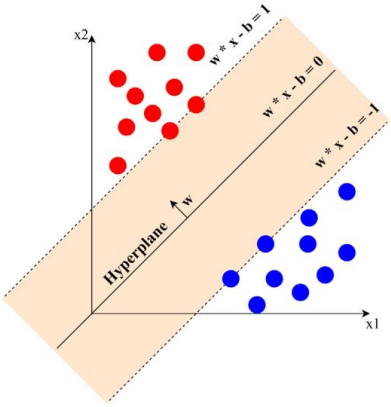


Figure 2: Solution to a problem produced by SVM

2. K-Nearest Neighbour Classifier (KNNC)

KNNC is a classification method that labels an unclassified data point using information from a classified data point. In this, the value of K is determined, which is, by default, taken as 5. An unclassified data point is taken as the input vector and assigned to a class to which most of the K belongs, as shown in Fig. 3. Each data point category is counted from the K neighbors of a given data point. The freshly collected data points are categorized into groups with the most data points among their neighbors. In Figure 3, the central star represents a data point: when K is 3, the star will be designated to the orange class, and when K is 10, it will be given to another blue class. The distance that exists between data points of the same class described in equations (6) and (7) is measured using the Euclidean Distance (De) or Manhattan Distance (Dm), respectively. The Euclidean or Manhattan distance is applied when the data points are continuous variables. Hamming Distance (DH) (Eq. 8) is used when data points are categorical variables.

$$\text{de} = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{\frac{1}{2}} \tag{6}$$

$$\text{dm} = \sum_{i=1}^{n}|x_i - y_i| \tag{7}$$

$$\text{dh} = \sum_{i=1}^{n}|(x_i - y_i)^2| \tag{8}$$

KNNC considers all features equally, making it the simplest machine-learning technique. When data has a lot of redundant properties, KNNC is utilized.

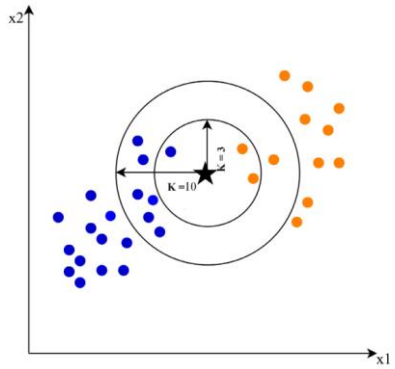


Figure 3: Solution to a problem by KNN

3. Naïve Bayes Classifier (NBC)

Naive Bayes is another machine learning method for classifying data and predicting a class based on an example from a set of attributes and knowledge of prior occurrences. This method uses the Bayes Theorem to predict and analyze datasets. The Bayes theorem uses a straightforward formula, as demonstrated in Eq. 9, to determine the likelihood of an event based on the information available. NBC helps make features independently contribute to the probability of an event even when they are not correlated. NBC helps eliminate unrelated features and thus increases accuracy.

$$P(i|j) = \left(\frac{P(j|i)*P(i)}{P(j)}\right) \quad (9)$$

where,

P(i|j) = the probability of hypotheses "i" on given data "j."

P(j|i) = the probability of data "j" given hypotheses "i."

P(i) = the probability of hypothesis "i" being accurate.

P(j) = the probability of data given.

4. Choice Tree Classifier (CTC)

Data from the dataset are used to build a tree of choice using supervised machine learning. CTC is utilized to address binary classification issues. Each node represents a variable that evaluates how accurately each node is classifying the labeled data. Each node learns about classified data by calculating entropy and information gained at each node, as shown in Eq. 10 and 11, respectively. This method aids in choosing the optimal node to serve as the parent node. The child node consists of values of selected input data. The dataset is partitioned to give each region access to the most data points possible. The procedure repeats until there are no more splits possible.

$$\text{Entropy(S)} = -\sum_{i=1}^{n} p_i log_2 p_i \quad (10)$$

where, pi = probability of an element being in class i

S = dataset

$$\text{Gain(S,A)} = \text{Entropy(S)} - \text{Entropy(S|A)} \quad (11)$$

where, S = dataset

A = given feature

Entropy aids in determining if data is certain or disordered. Considering entropy and information gain can assist determine whether or not the data has to be separated. If entropy is high, then the variance in data is also high. Entropy also helps to calculate the information gain, as shown in Eq. 11. The more information is held by a feature that aids in a better split, the larger the information gain. A dataset's information gain is equal to the entropy it has lost. Entropy and information gain must be decided on, and a perfect dataset split is needed.

5. Random Forest Classifier (RFC)

The random forest classifier is made up of multiple-choice trees. Choice trees cast votes for categorization tasks in this ensemble machine-learning method. This method evaluates the average of predictions of trees. The prediction of a random forest with n trees and individual weight Wj has been shown in Eq. 12. When a tree becomes deeper and deeper, an overfitting problem arises. RFC helps eliminate the overfitting problem by taking inputs in vector forms. The vectorized input passes each decision tree present in RFC. Each decision tree classifies the vector input. Then, the RFC decides which class to take as the output, either the highest-voted class or the average of all votes class. Voting rules depend on the input vectors.

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} W_j(x_j, x') \quad (12)$$

Fig. 4 shows an example of RFC having three decision trees. The decision trees work on input vectors from the same dataset and classify the classes as Class A or B. The Random Forest technique achieves high accuracy as compared to traditional classification techniques.

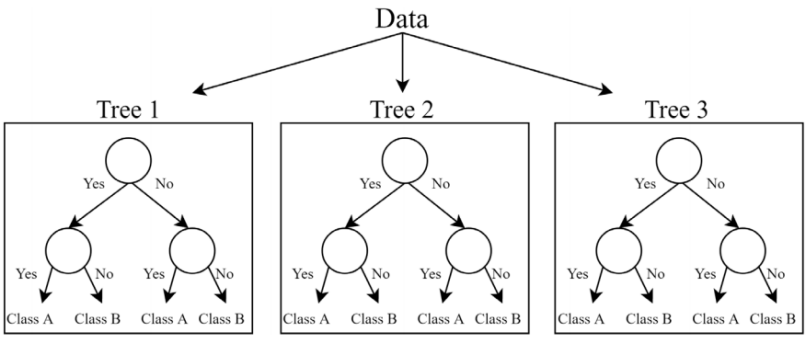


Figure 4: Solution to a problem by Random Forest

6. Ensemble Method Classifier (EMC)

Different constraints of classifiers are used to train ensemble methods. The results of numerous classifiers are combined with ensemble classifiers to provide a single, incredibly accurate result. Ensemble methods are machine learning techniques that combine obtained results to improve general classification performance. Researchers have coined many terms for ensemble methods like multi-strategy learning, multiple integrations, classifier synthesis, and grouping classifier. Ensemble classification is categorized into two standard techniques: Boosting and Bagging.

Training data is affected by the Boosting method. To create a classifier Ci(x), identical weights are first applied to each occurrence of xi, and with each iteration of I, the knowledge technique reduces the errors present in the training set. Weights are added to training instance xi, and the Ci(x) error is assessed. Finally, the weight of xi increases, which affects the classifier's outcome and impacts the weights of misclassified xi and classified xi.

Bagging is a technique in which N items change distinct T training data. The training sets are known as bootstrap duplicates, some of which are undesirable. The proposed ensemble method combines the features of a Choice tree, support vector machine, and random forest classifiers. All these supervised controlled machine learning classifiers enhance the capabilities of the ensemble method. The method further uses the AdaBoost technique to boost the results. In contrast with existing strategies, the suggested ensemble method classifier (EMC) is more efficient and produces better precision, recall, accuracy, and F1-score outcomes. The hypothesized ensemble method results are validated using the K-fold cross-validation procedure with a value of k equal to 15.

C. Performance Parameters

The classifiers are evaluated based on the following metrics:

- Accuracy: The correct forecasts out of all the predictions are referred to as accuracy and are calculated as;
- Accuracy = (TP+TN)/(TN+FP+TP+FN)
- Recall: it is the measurement of the completeness of a classifier.
- Recall = TP/(TP+FN)
- Precision: Precision is the capability of the system to produce relevant results.
- Precision = TP/(TP+FP)
- F1-score: F1-score is the mean of recall and precision
- F1-score = (2*TP)/(2*TP+FP+FN)
- AUC: Area under curve and plots TPR vs. FPR where TPR is True Positive Rate and FPR is False Positive Rate.

Where TP = True Positive, TN = true Negative, FP = False Positive, and FN = False Negative

5. Experimentation

This section displays the results of different machine learning classifiers that used three feature selection strategies on different medical datasets. The report also includes the findings of comparisons between ensemble classifiers and other techniques. Three feature selection strategies are integrated with the ensemble method. The unbalanced datasets are balanced using the SMOTE approach, and

the results are validated using K-fold cross-validation with K equal to 15. The k-fold cross-validation technique is used to evaluate any model and also helps to understand the learning rate of the model. Accuracy, recall, F1-score, precision, and AUC are used for performance evaluation. For categorization, features with high information gain values are used. The effectiveness of each classifier is assessed after the three feature selection techniques have chosen features.

Fisher's score selects 25 features out of 32, information gain also selects 25 out of 32, and the genetic algorithm selects 20 out of 32. Table 2 displays the features that were chosen using various feature selection techniques. Table 3 displays the performance of each classifier on the WBCD (Wisconsin Breast Cancer Dataset). The ensemble method outperforms the other classifiers by achieving 97.66% accuracy, AUC of 96.97%, recall of 94.83%, precision of 98.21%, and F1-score of 96.49%. The confusion matrix, learning rate, performance evaluation, and ROC curve of the ensemble method are shown in Table 4.

Table 2: Features selection by different methods on Wisconsin Breast Cancer Dataset

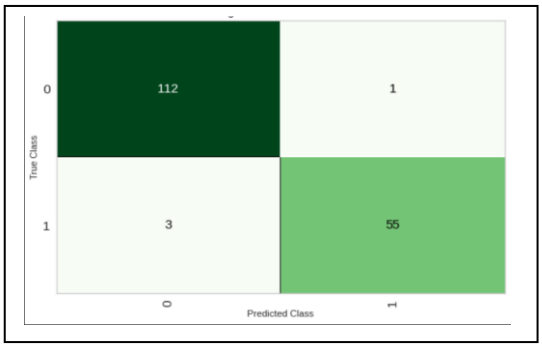
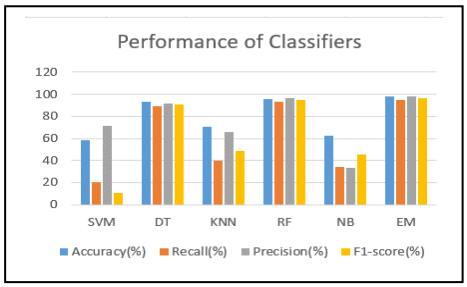
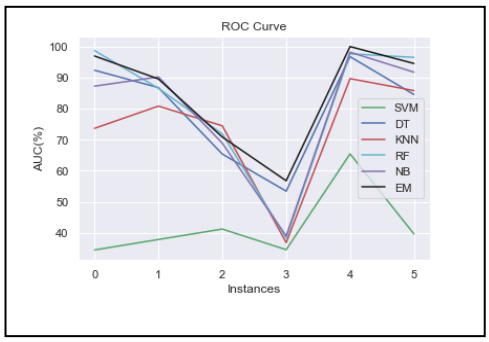
Feature Selection Method	No. of features selected	Features
Fisher's Score	25	id, diagnosis, mean_radius, mean_texture, mean_perimeter mean_area, mean_smoothness, mean_compactness, mean_concavity, mean_concave_points, mean_symmetry, mean_fractal_dimension, se_radius, se_texture, se_perimeter, se_area, se_smoothness, se_compactness, se_concavity, se_concave_points, se_symmetry, se_fractal_dimension, w_radius, w_texture, w_perimeter
Information Gain	25	id, diagnosis, mean_radius, mean_texture, mean_perimeter mean_area, mean_smoothness, mean_compactness, mean_concavity, mean_concave_points, mean_symmetry, mean_fractal_dimension, se_radius, se_texture, se_perimeter, se_area, se_smoothness, se_compactness, se_concavity, se_concave_points, se_symmetry, se_fractal_dimension, w_radius, w_texture, w_perimeter
Genetic Algorithm	20	id, diagnosis, mean_radius, mean_texture, mean_perimeter mean_area, mean_smoothness, mean_compactness, mean_concavity, mean_concave points, mean_symmetry, mean_fractal_dimension, se_radius, se_texture, se_perimeter, se_area, se_smoothness, se_compactness, se_concavity, se_concave_points

Table 3: Performance of all classifiers on the Wisconsin Breast Cancer Dataset

Model	Accuracy(%)	AUC(%)	Recall(%)	Precision(%)	F1-score(%)

SVM	58.32	34.50	20.00	71.25	10.64
DT	93.23	92.38	89.52	91.84	90.48
KNN	70.59	73.67	39.71	66.02	48.80
RF	95.99	98.71	93.05	96.16	94.38
NB	62.81	87.25	34.05	33.45	45.67
EM	97.66	96.97	94.83	98.21	96.49

Table 4: Confusion matrix, Learning rate, Performance evaluation, and ROC Curve of Ensemble method on Wisconsin Breast Cancer Dataset

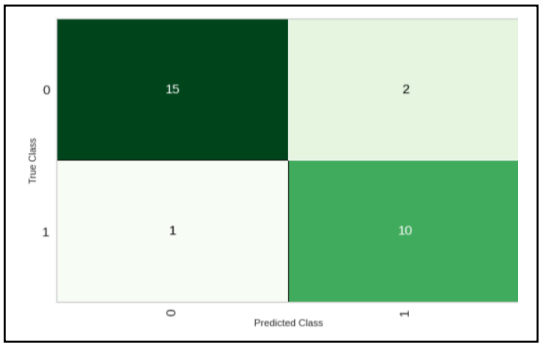
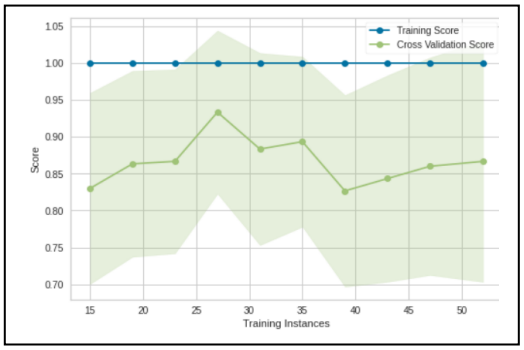
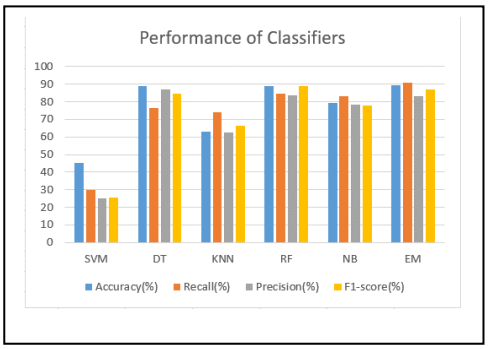
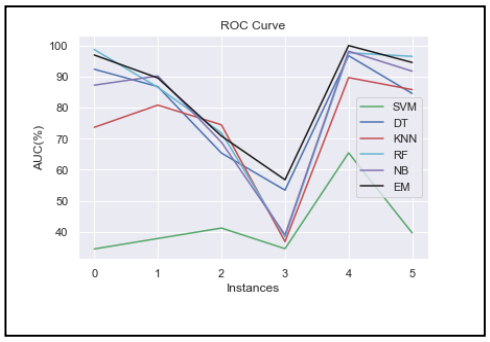
Confusion Matrix of EM	Learning Rate of EM
	
Performance of Classifiers	**ROC Curve**
	

Fisher's score, information gain, and genetic algorithm select all the features in the Cryotherapy dataset due to a smaller number of features present in it. The performance of all the classifiers on this dataset is presented in Table 5. The ensemble method achieves an accuracy of 89.29%, AUC of 89.57%, recall of 90.91%, precision of 83.33%, and F1-score of 86.96%. The confusion matrix, learning rate, performance evaluation, and ROC curve of the ensemble method are presented in Table 6.

Table 5: Classifier performance on the Cryotherapy Dataset

Model	Accuracy(%)	AUC(%)	Recall(%)	Precision(%)	F1-score(%)
SVM	45.00	37.87	30.00	25.00	25.33
DT	88.98	86.78	76.56	86.79	84.56
KNN	62.86	80.83	74.17	62.50	66.29
RF	88.67	86.56	84.54	83.56	89.00
NB	79.05	90.22	83.33	78.17	77.77
EM	89.29	89.57	90.91	83.33	86.96

Table 6: Confusion matrix, Learning rate, Performance evaluation and ROC Curve of Ensemble method on Cryotherapy Dataset

Confusion Matrix of EM	Learning Rate of EM
	
Performance of Classifiers	**ROC Curve**
	

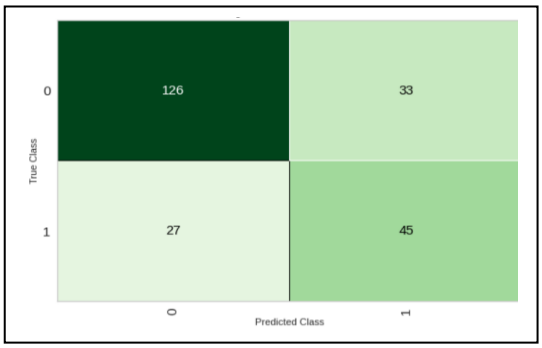
Fisher's score, information gain, and genetic algorithm select all the features in the PIDD due to the number of features. Each classifier's performance on this dataset is presented in Table 7. The ensemble method achieves an accuracy of 74.03%, AUC of 70.87%, recall of 62.5%, precision of 57.69%, and F1-score of 60.00%. The confusion matrix, learning rate, performance evaluation, and ROC curve of the ensemble method classification technique are shown in Table 8.

Table 7: Classifiers performance on the Pima Indian Diabetes Dataset

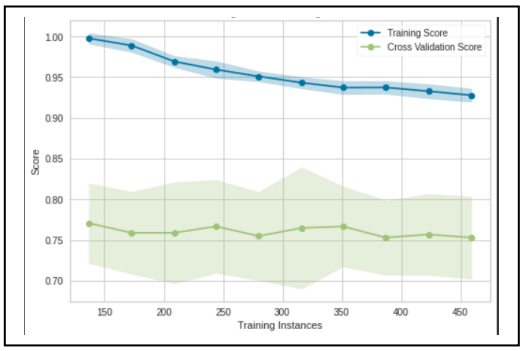
Model	Accuracy(%)	AUC(%)	Recall(%)	Precision(%)	F1-score(%)
SVM	65.52	41.23	27.43	39.40	27.44
DT	68.33	65.39	56.20	53.94	54.47
KNN	69.27	74.48	49.21	55.88	51.85
RF	73.66	71.89	52.98	56.67	58.25
NB	71.11	68.89	42.69	68.85	52.32
EM	74.03	70.87	62.50	57.69	60.00

Table 8: Confusion matrix, Learning rate, Performance evaluation and ROC Curve of Ensemble method on Pima Indian Diabetes Dataset

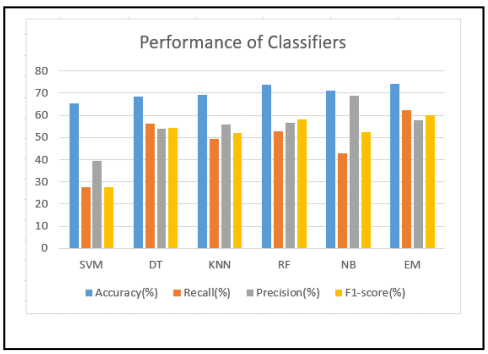
Confusion Matrix of EM	Learning Rate of EM

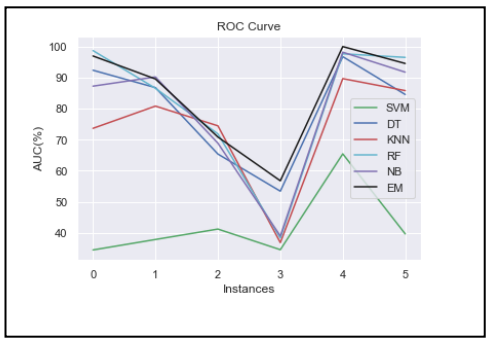


Performance of Classifier



ROC Curve





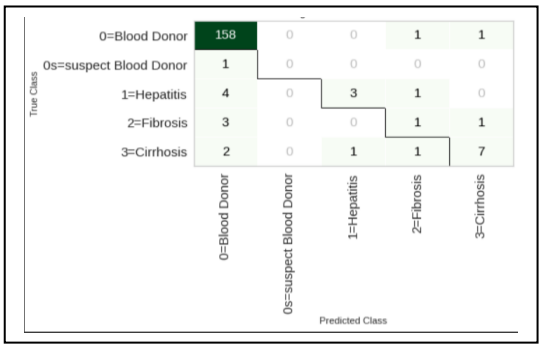
Fisher's score, information gain, and genetic algorithm select all the features in the Hepatitis dataset due to a smaller number of features present in it. The performance of all the classifiers on this dataset is presented in Table 9. The ensemble method achieves an accuracy of 91.35%, AUC of 56.78%, recall of 43.98%, precision of 89.88%, and F1-score of 90.25%. The confusion matrix, learning rate, performance evaluation, and ROC curve of the ensemble method are presented in Table 10.

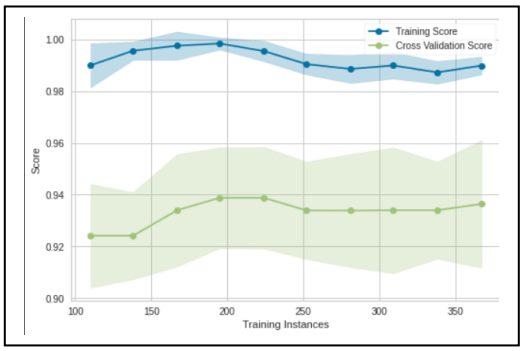
Table 9: Classifiers performance on the Hepatitis Dataset

Model	Accuracy(%)	AUC(%)	Recall(%)	Precision(%)	F1-score(%)
SVM	80.70	34.58	39.99	85.33	79.56
DT	89.89	53.43	45.20	67.89	54.47
KNN	90.11	36.86	48.20	88.87	58.45
RF	90.23	38.34	53.23	84.56	90.85
NB	89.53	39.02	44.19	89.04	88.96
EM	91.35	56.78	43.98	89.88	90.25

Table 10: Confusion matrix, Learning rate, Performance evaluation and ROC Curve of Ensemble method on Hepatitis Dataset

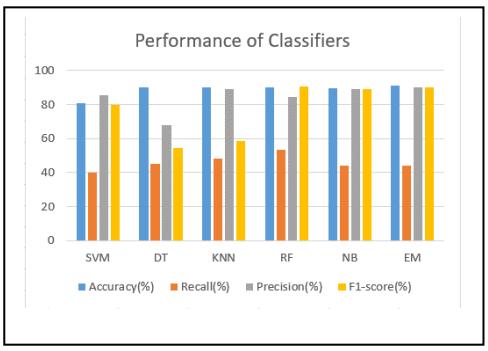
Confusion Matrix of EM	Learning Rate of EM

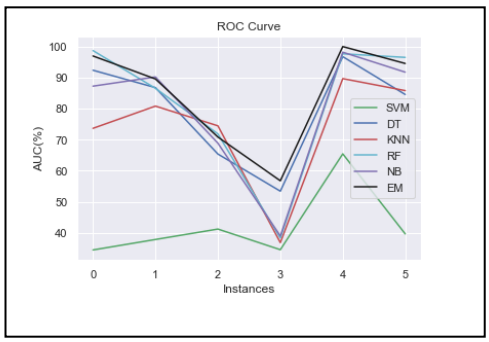




Performance of Classifiers

ROC Curve





Feature selection methods select features in the chronic kidney disease dataset. Fisher's score selects 15 out of 26, information gain also selects 15 out of 26 features, and the genetic algorithm selects 10 out of 26. The selected features are presented in Table 11. Table 12 displays the performance of each classifier on the Chronic Kidney Disease Dataset. The ensemble method outperforms the other classifiers by achieving 99.17% accuracy, AUC of 100%, recall of 100%, precision of 98.73%, and F1-score of 99.36% in comparison with Abdelrahim et al. model that achieved an accuracy of 97.91% [28], and Narasimha Swamy et al. [29] the method that achieved an accuracy of 97.98% with the CKD dataset. The confusion matrix, learning rate, performance evaluation, and ROC curve of the ensemble method are shown in Table 13.

Table 11: Features selection by different methods on Chronic kidney Disease Dataset

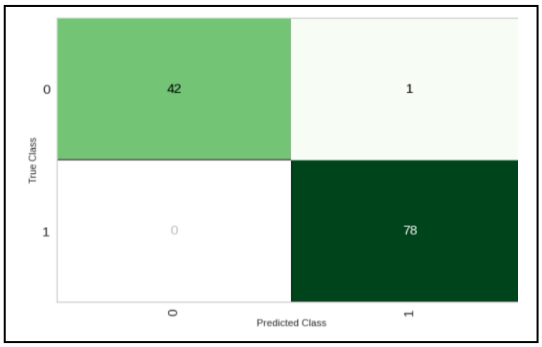
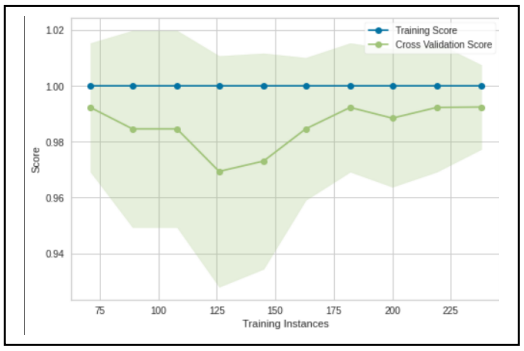
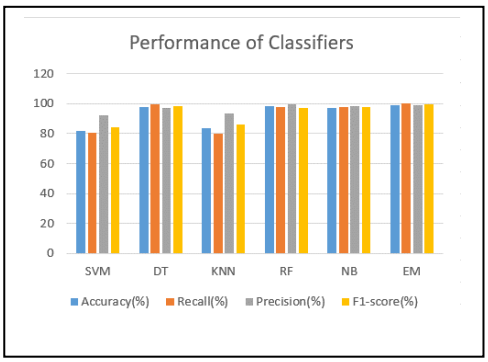
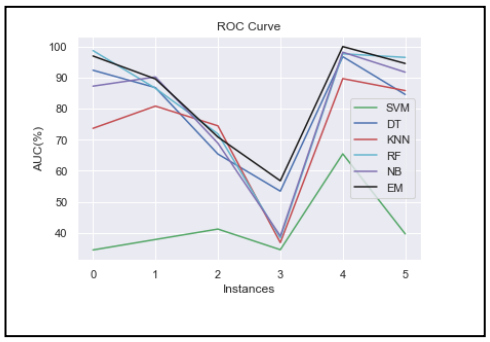
Feature Selection Method	No. of features selected	Features
Fisher's Score	15	Id, Age, bld_prs, sugr, brg, bu, sd, sod, pot, hemoglo, pcv_count, w_count, r_count, appet, classification
Information Gain	15	Id, Age, bld_prs, sugr, brg, bu, sd, sod, pot, hemoglo, pcv_count, w_count, r_count, appet, classification
Genetic Algorithm	10	Age, bp, sugr, brg, bu, sd, sod, hemoglo, pcv_count, classification

Table 12: Classifiers performance on the Chronic Kidney Disease Dataset

Model	Accuracy(%)	AUC(%)	Recall(%)	Precision(%)	F1-score(%)

SVM	81.68	65.45	80.75	92.33	84.03
DT	97.50	96.75	99.41	96.93	98.10
KNN	83.48	89.68	80.23	93.18	86.10
RF	98.66	97.67	97.89	99.44	96.89
NB	97.47	98.19	97.78	98.36	98.02
EM	99.17	100	100	98.73	99.36

Table 13: Confusion matrix, Learning rate, Performance evaluation and ROC Curve of Ensemble method on Chronic Kidney Disease Dataset

Confusion Matrix of EM	Learning Rate of EM
	
Performance of Classifiers	**ROC Curve**
	

Feature selection methods select different features in Parkinson's speech dataset. Fisher's score selects 11 out of 24, information gain also selects 11 out of 24 features, and the genetic algorithm selects 5 out of 24. Table 14 displays the features chosen using various feature selection techniques. Table 15 displays the results of all the classifiers on the dataset of Parkinson's speech. The ensemble method achieves 99.80% accuracy, 96.85% AUC, 95.83% recall, 99.21% precision, and 98.49% F1-score values. The confusion matrix, learning rate, performance evaluation, and ROC curve of the ensemble method are shown in Table 16.

Table 14: Features selection by different methods on Parkinson's Speech Dataset [30]

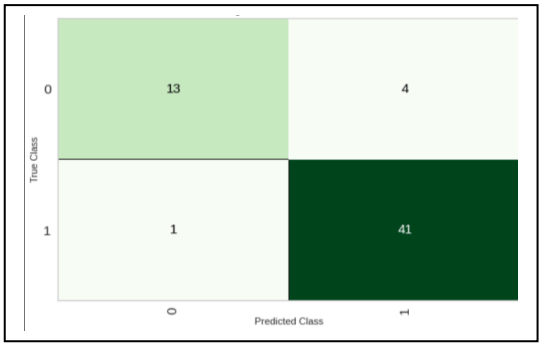
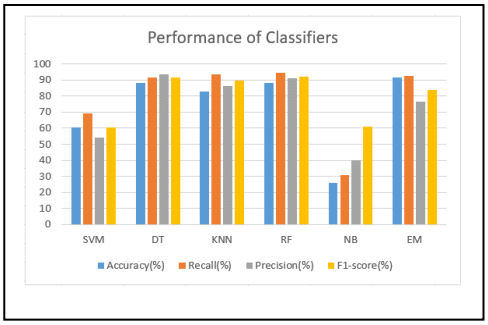
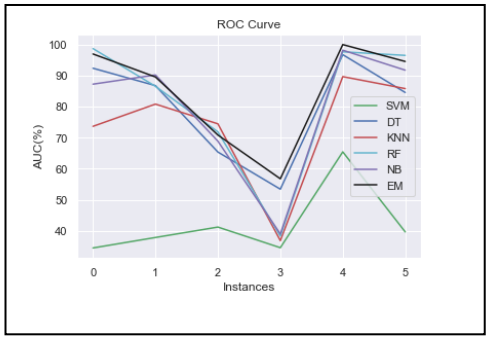
Feature Selection Technique Used	Features selected	Features

Fisher's Score	11	MDVP:Fo(Hz), MDVP:Flo(Hz), MDVP:Fhi(Hz), MDVP:Jitter(Abs), Shimmer:APQ5, DFA, NHR, PPE, RPDE, spread1, spread2
Information Gain	11	MDVP:Fo(Hz), MDVP:Flo(Hz), MDVP:Fhi(Hz), MDVP:Jitter(Abs), Shimmer:APQ5, DFA, NHR, PPE, RPDE, spread1, spread2
Genetic Algorithm	5	MDVP:Fo(Hz), MDVP:Shimmer(dB), MDVP:PPQ, spread1, spread2

Table 15: Classifiers performance on Parkinson's Speech Dataset

Model	Accuracy(%)	AUC(%)	Recall(%)	Precision(%)	F1-score(%)
SVM	60.33	39.67	69.09	53.96	60.56
DT	88.19	84.57	91.64	93.68	91.81
KNN	83.08	85.82	93.45	86.43	89.60
RF	88.24	96.52	94.45	90.97	92.15
NB	25.77	91.76	30.82	40.00	60.97
EM	91.50	94.56	92.85	76.47	83.87

Table 16: Confusion matrix, Learning rate, Performance evaluation and ROC Curve of Ensemble method on Parkinson's Speech Dataset

Confusion Matrix of EM	Learning Rate of EM
	
Performance of Classifiers	**ROC Curve**
	

The proposed ensemble method outperformed the current approaches. The tables given show the performance analysis of each classifier.

6. Conclusion

In addition to an ensemble method, numerous machine-learning approaches have been applied in the research. Accuracy, precision, recall, F1-score, and AUC are used to evaluate each technique. The ensemble method uses AdaBoost and K-fold-cross-validation techniques with value of k equal to 15. These techniques enhance the efficiency of model and hence it outperforms the other machine learning techniques. In this paper, six different medical datasets have been used. Three feature selection techniques, namely Fisher's score, information gain and genetic algorithm are also used. These techniques help select the relevant features from the given input and helps in increasing the efficiency of the classifiers. Feature selection techniques are always beneficial for reducing complexity and increasing any classifier's efficiency. The proposed ensemble method outperforms the existing machine learning classifiers with improved accuracy. The proposed method cannot replace healthcare professionals but can be an early diagnosis tool for various diseases. There is always room for improvement in every field, and medical diagnosis is a vast field for researchers. The proposed ensemble approach will be examined on additional healthcare datasets in the future.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] M. Shehab et al., "Machine learning in medical applications: A review of state-of-the-art methods," Comput. Biol. Med., vol. 145, no. November 2021, 2022, doi: 10.1016/j.compbiomed.2022.105458.

[2] S. David, J. Andrew, K. Martin Sagayam, and A. A. Elngar, "Augmenting security for electronic patient health record (ePHR) monitoring system using cryptographic key management schemes," Fusion Pract. Appl., vol. 5, no. 2, pp. 51–61, 2021, doi: 10.54216/FPA.050201.

[3] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities," IEEE Access, vol. 5, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[4] J. Abed Eleiwy and N. Jaafar, "Novel Filter of DWT for Image Processing Applications," Fusion Pract. Appl., vol. 4, no. 2, pp. 32–41, 2021, doi: 10.54216/FPA.040205.

[5] A. G, A. Kumar, and V. R, "Query-Based Image Retrieval using Support Vector Machine (SVM)," J. Cogn. Human-Computer Interact., vol. 1, no. 1, pp. 28–36, 2021, doi: 10.54216/jchci.010104.

[6] G. Akhila, H. K, and J. R. Jaramillo, "Indian Premier League Using Different Aspects of Machine Learning Algorithms," J. Cogn. Human-Computer Interact., vol. 1, no. 1, pp. 01–07, 2021, doi: 10.54216/jchci.010101.

[7] M. Ramzan, "Comparing and evaluating the performance of WEKA classifiers on critical diseases," India Int. Conf. Inf. Process. IICIP 2016 - Proc., pp. 1–4, 2017, doi: 10.1109/IICIP.2016.7975309.

[8] B. V. Ramana and R. S. Kumar Boddu, "Performance comparison of classification algorithms on medical datasets," 2019 IEEE 9th Annu. Comput. Commun. Work. Conf. CCWC 2019, pp. 140–145, 2019, doi: 10.1109/CCWC.2019.8666497.

[9] E. Alexopoulos1, G. D. Dounias, and K. Vemmos, "Medical diagnosis of stroke using inductive machine learning," Mach. Learn. Appl. Mach. Learn. Med. Appl., no. September 1999, pp. 20–23, 1999, [Online]. Available: http://www.researchgate.net/publication/2819899_Medical_Diagnosis_Of_Stroke_Using_Inductive_Machine_Learning/file/9fcfd51407a635db88.pdf

[10] G. W.H.S.D, "Performance Evaluation on Machine Learning Classification Techniques for Disease (CKD)," Ieee, pp. 291–296, 2017, doi: 10.1109/BIBE.2017.00056.

[11] M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, "An analytical method for diseases prediction using machine learning techniques," Comput. Chem. Eng., vol. 106, pp. 212–223, 2017, doi: 10.1016/j.compchemeng.2017.06.011.

[12] A. Cüvitoğlu and Z. Işik, "Evaluation machine-learning approaches for classification of Cryotherapy and Immunotherapy datasets," Int. J. Mach. Learn. Comput., vol. 8, no. 4, pp. 331–335, 2018, doi: 10.18178/ijmlc.2018.8.4.707.

[13] A. Garg and V. Mago, “Role of machine learning in medical research: A survey,” Comput. Sci. Rev., vol. 40, p. 100370, May 2021, doi: 10.1016/J.COSREV.2021.100370.
[14] A. Aada and S. Tiwari, “Predicting Diabetes in Medical Datasets Using Machine Learning Techniques,” Int. J. Sci. Res. Eng. Trends, vol. 5, no. 2, pp. 257–267, 2019.
[15] R. G. Nadakinamani et al., “Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques,” Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/2973324.
[16] P. Jabbari and N. Rezaei, “Artificial intelligence and immunotherapy,” Expert Rev. Clin. Immunol., vol. 15, no. 7, pp. 689–691, 2019, doi: 10.1080/1744666X.2019.1623670.
[17] E. H. Houssein, E. Saber, Y. M. Wazery, and A. A. Ali, “Swarm Intelligence Algorithms-Based Machine Learning Framework for Medical Diagnosis: A Comprehensive Review,” Stud. Comput. Intell., vol. 1038, pp. 85–106, 2022, doi: 10.1007/978-3-030-99079-4_4/COVER.
[18] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, “A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms,” Mob. Inf. Syst., vol. 2022, 2022, doi: 10.1155/2022/1410169.
[19] M. W. Floyd, J. T. Turner, and D. W. Aha, “Using deep learning to automate feature modeling in learning by observation,” FLAIRS 2017 - Proc. 30th Int. Florida Artif. Intell. Res. Soc. Conf., no. June, pp. 50–55, 2017.
[20] A. Alanazi, “Using machine learning for healthcare challenges and opportunities,” Informatics Med. Unlocked, vol. 30, no. March, p. 100924, 2022, doi: 10.1016/j.imu.2022.100924.
[21] A. Pan, S. Mukhopadhyay, and S. Samanta, “Liver Disease Detection,” Int. J. Healthc. Inf. Syst. Informatics, vol. 17, no. 2, pp. 1–19, 2022, doi: 10.4018/ijhisi.299956.
[22] S. Mall, A. Srivastava, B. D. Mazumdar, M. Mishra, S. L. Bangare, and A. Deepak, “Implementation of machine learning techniques for disease diagnosis,” Mater. Today Proc., vol. 51, pp. 2198–2201, 2022, doi: 10.1016/j.matpr.2021.11.274.
[23] P. Dinesh, A. S. Vickram, and P. Kalyanasundaram, “Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: SVM, KNN, logistic regression, random forest and decision tree to measure accuracy,” FIFTH Int. Conf. Appl. Sci. ICAS2023, vol. 3097, no. 1, p. 020140, May 2024, doi: 10.1063/5.0203746/3290220.
[24] M. S. Singh, K. Thongam, P. Choudhary, and P. K. Bhagat, “An Integrated Machine Learning Approach for Congestive Heart Failure Prediction,” Diagnostics, vol. 14, no. 7, pp. 1–21, 2024, doi: 10.3390/diagnostics14070736.
[25] A. N. Al Masri and H. Mokayed, “An Efficient Machine Learning based Cervical Cancer Detection and Classification,” J. Cybersecurity Inf. Manag., vol. 2, no. 2, pp. 58–67, 2020, doi: 10.54216/jcim.020203.
[26] A. Abdelhafeez and H. K. Mohamed, “Skin cancer detection using neutrosophic c-means and fuzzy c-means clustering algorithms,” J. Intell. Syst. Internet Things, vol. 8, no. 1, pp. 33–42, 2023, doi: 10.54216/JISIoT.080103.
[27] L. Sun, T. Wang, W. Ding, J. Xu, and Y. Lin, “Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification,” Inf. Sci. (Ny)., vol. 578, pp. 887–912, 2021, doi: 10.1016/j.ins.2021.08.032.
[28] Abdelrahim Koura and H. S. Elnashar, “Data Mining Algorithms for Kidney Disease Stages Prediction,” J. Cybersecurity Inf. Manag., vol. 1, no. 1, pp. 21–29, 2020, doi: 10.54216/jcim.010104.
[29] B. N. Swamy, R. Nakka, A. Sharma, S. P. Praveen, V. N. Thatha, and K. Gautam, “An Ensemble Learning Approach for detection of Chronic Kidney Disease (CKD),” J. Intell. Syst. Internet Things, vol. 10, no. 2, pp. 38–48, 2023, doi: 10.54216/JISIoT.100204.
[30] R. Lamba, T. Gulati, H. F. Alharbi, and A. Jain, “A hybrid system for Parkinson’s disease diagnosis using machine learning techniques,” Int. J. Speech Technol., vol. 25, no. 3, pp. 583–593, 2022, doi: 10.1007/s10772-021-09837-9.