An optimal multi-disease prediction framework using hybrid machine learning techniques

Aditya Gupta*, Amritpal Singh
Dept. of Computer Science and Engineering,
Dr. B R Ambedkar National Institute of Technology, Jalandhar, India
**Corresponding author: adityag.cs.19@nitj.ac.in*

Abstract

The prediction of lifestyle diseases is a vital domain in healthcare informatics research. This task is primarily achieved using the widely available machine learning algorithms. However, the high-dimensionality of data amplifies the computation complexity and significantly reduces the models' efficiency. Conspicuously, we presented a multi-disease prediction strategy for intelligent decision support using ensemble learning. The proposed work leverages genetic algorithm-based recursive feature elimination and AdaBoost to predict two prominent lifestyle diseases. Alongside the proposed approach, various benchmark machine learning techniques are also trained and validated using selected features under k-fold cross-validation. The results reveal the effectiveness of the proposed methodology in predicting multiple diseases in comparison to past works.

Keywords: AdaBoost; ensemble learning; genetic algorithms; healthcare analytics; multidisease

1. Introduction

Many chronic diseases are caused by unhealthy and unregulated lifestyles, especially in middle-aged or later-aged people. Lifestyle diseases are the long-term consequences of adopting harmful habits, as evidenced by a person's way of life (Han *et al.*, 2012). People in today's urbanized society are being pushed away from physical activities and moving forward towards sedentary habits such as processed foods, smoking and alcohol consumption, and long and improperly postured working hours. As a result, people develop chronic non-communicable diseases (NCDs) with potentially fatal consequences (Senapati *et al.*, 2015). According to the World Health Organization (WHO), NCDs kill around 41M people each year, accounting for 71% of global deaths. Of all deaths, 15M are between the ages of 30 and 69. The majority of NCD deaths are mainly due to cardiovascular disease, which accounts for 17.9M deaths, followed by cancer with 9.3M deaths. Other prominent diseases include respiratory disorders and diabetes, with 4.1M and 1.5M deaths, respectively. Figure 1 shows the top 10 reasons for death in the US according to the Centers for Disease Control and Prevention (CDC) in 2020. Due to the high mortality rate of NCDs, early detection and prediction of such diseases is a vital task.

Recent advancements in artificial intelligence (AI) have provided solutions in distinct domains, including prediction analysis (Garg, 2021) (Yahyaoui *et al.*, 2016). Machine learning, a subset of AI is a disruptive technology of the current era which helps medical professionals to make informed decisions. Despite its numerous pros, machine learning also poses varying challenges, which include accurate data acquisition (Sajjad *et al.*, 2019). Training a machine learning model on a limited-sized dataset can lead to an overfitted model (Almulla *et al.*, 2021). On the contrary, having a large set of features cause significant issues. The presence of irrelevant and excessively noisy features in a dataset can lead to classification uncertainty and poor accuracy. To create a highly efficient classification model, identifying the optimal set of features among the varied attributes is a very essential and critical task. The increased overhead and

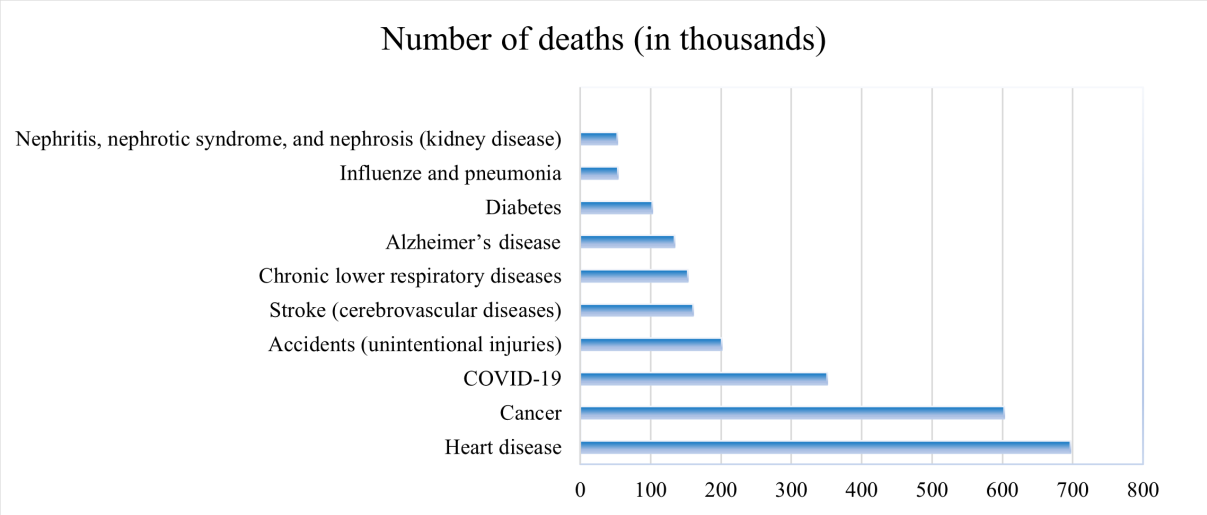

Fig. 1. Top 10 leading cause of death in the US

processing time in training and testing classification models is limited by a restricted set of features. Numerous researchers have presented distinguished frameworks in the healthcare domain, but very limited work has been carried out to eliminate the redundant features.

Considering these facts, in this study, two important lifestyle diseases, heart disease, and diabetes are considered to establish a framework for their early prediction and detection. Designing prediction models can help assist doctors in the diagnosis of critical diseases and thereby improves the overall quality of life.

The key contributions of the presented work are enlisted below:

1. A study of numerous lifestyle diseases prevailing in the US is presented.
2. The design of a machine learning-based predictive model for two important lifestyle diseases' diagnosis is proposed.
3. Data is preprocessed in terms of removing redundant data and handling missing data by utilizing the Multiple Imputation by Chained Equations (MICE) technique.
4. We proposed a modified version of recursive feature elimination based on a genetic algorithm (GA-RFE) to determine the optimal feature subset.
5. The AdaBoost classification model is trained alongside other predictive models for multi-disease prediction.
6. An extensive comparative study has been conducted to evaluate the effectiveness of the proposed model.

The work is organized into different sections. Section 2 briefly discusses the previous works. Section 3 describes various preliminaries for the proposed work. Section 4 discusses the proposed methodology adopted for the current study. The Experimental setup with results and discussions is presented in section 5. Finally, section 6 summarizes the paper.

2. Related works

Predictive models have recently demonstrated their utility in a variety of disciplines, and are not restricted to healthcare. As the health data is highly sensitive compared to other application areas, critical attention becomes unavoidable. Classical machine learning algorithms prove to be a cardinal tool to predict diseases earlier in the disease cycle. Moreover, various ensemble learning approaches are currently in trend and act as a booster to improve accuracy. Nevertheless, these classical algorithms play

an important role and act as a backbone to provide the baseline for them. There has been significant research on the literature using machine learning algorithms. Ghosh *et al.,* (2021) (Ghosh *et al.*, 2021) introduced a cardiovascular disease prediction framework using a hybrid classifier. The authors utilized the ReliefF and Least Absolute Shrinkage and Selection Operator (LASSO) method for optimal feature selection. The experimental results proved the robustness and effectiveness in the prediction of cardiac disorders. Nandy *et al.,* (2021) (Nandy *et al.*, 2021) provided a framework for heart disease prediction using the Swarm-ANN method. The proposed methodology outperformed existing works in terms of prediction accuracy, precision, and recall. However, the dataset employed for the training purpose was very small. Katarya and Meena (2021) (Katarya & Meena, 2021) presented a study to discuss the clinical manifestations of heart diseases and evaluated classification accuracy by modeling traditional classification algorithms. However, no feature selection criteria were adopted for the presented study. In 2021, (Arumugam *et al.*, 2021) presented a study on the utilization of machine learning for multiple disease prediction. The authors explored various prediction algorithms without performing any data analysis. In the study by Singh *et al.,* (2021) (Singh *et al.*, 2021), an ensemble of existing machine learning techniques was used for early diabetes prediction. The authors achieved a classification accuracy of 95%. However, the dataset used for the study only contained data from female patients. Furthermore, they did not consider any techniques for optimal feature selection from the original dataset. Similar works have presented by other researchers for the prediction of other life-threatening diseases. Table 1 provides a comparative analysis of the proposed framework with some of the most relevant works.

Table 1. Comparative analysis of related works

Authors	**Application Domain**	**Feature selection**	**Ensemble learning**	**Decision making model**
(Ghosh *et al.*, 2021)	CVD	ReliefF, LASSO	No	Random Forest Bagging Method
(Nandy *et al.*, 2021)	CVD	No	No	Swarm-ANN
(Katarya & Meena, 2021)	CVD	No	No	Random Forest
(Arumugam *et al.*, 2021)	CVD, Diabetes	No	No	NA
(Singh *et al.*, 2021)	Diabetes	No	Yes	Ensemble modeling
(Ahmed *et al.*, 2022)	Diabetes	No	No	SVM, ANN
Proposed	Multidisease	GA-RFE	Yes	AdaBoost

All of the aforementioned works mainly focused on predictive analysis. However, feature reduction before training the machine learning model is the need of the hour for improved decision making. Further, very limited researchers have utilized nature-inspired approaches in their presented works.

3. Preliminaries

This section provides a brief description of the relevant concepts used in this paper.

3.1 Genetic algorithms

John Holland introduced the basis of genetic algorithms on top of dynamic solutions in the early 1960s (Mirjalili, 2019). A genetic algorithm is a biologically inspired search strategy that is widely used in computing to find approximate solutions to complex problems (Sharma *et al.*, 2020). The algorithm is based on Darwin's theory of natural evolution and mimics the biological phenomenon of survival of the fittest. Figure 2 shows the working principle of a classical genetic algorithm.

The goal of the genetic algorithm is to obtain optimal or near-optimal solutions (Whitley, 1994). To this end, the genetic algorithm starts with the initial set of randomly generated populations. The population comprises a set of solutions or chromosomes, each of which is represented by a string of binary or real values. Every bit in the chromosome is termed as a gene and specifies the character of a particular solution. Following the development of the initial generation, the fitness of each solution is calculated. The fitness values determine the potential of a particular individual to reproduce a candidate solution. The algorithm proceeds to construct successive generations using the three operators- selection, crossover, and mutation. The selection operator is used to choose the chromosomes that will contribute to the future generation. Crossover operators choose genes from parent chromosomes and generate new offspring for the next generation. To prevent all solutions in the population from collapsing into local

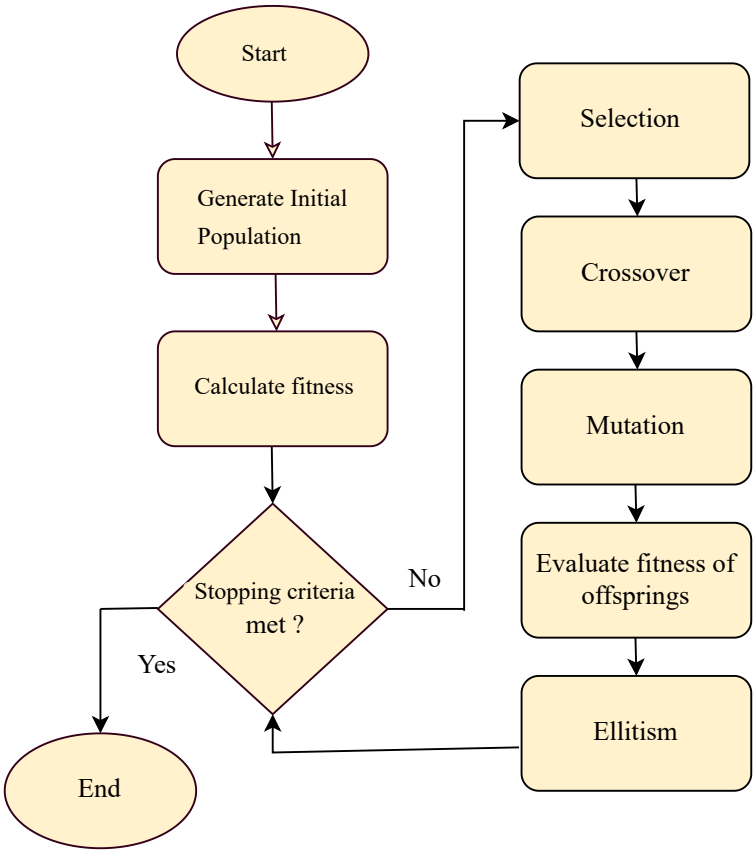

Fig. 2. Working principle of classical genetic algorithm

optima, the mutation operator is utilized. It introduces random alterations in the offspring. The same process is repeated until the desired solution is obtained.

3.2 Ensemble learning

Ensemble learning is a hybrid modeling technique that utilizes the capabilities of multiple weak learners to realize a strong learner (Dietterich *et al.*, 2002). The ensemble learning approach tends to improve or optimize the predictive performance by combining the predictions of numerous machine learning models. Another advantage of using ensemble learning results is eliminating the overfitting problem and boosts the overall learning performance (Sarmadi *et al.*, 2021). Ensemble learning encompasses three fundamental concepts, namely bagging, stacking, and boosting. Out of these three, bagging and boosting are generally considered the most prominent ensemble-based strategies. Bagging is the practice of fitting multiple decision trees to different samples of the same dataset and then averaging their decisions. On the contrary, boosting involves sequentially adding decisions made by weak learners to give a strong learner.

4. Proposed methodology

This section discusses the proposed methodology adopted for multi-disease prediction. It consists of mainly four components. Figure 3 presents the system model.

4.1 Preprocessing

Data preprocessing is one of the most inevitable tasks in the pipeline of a machine learning framework. Data preprocessing refers to a variety of possible strategies to deal with noisy, incomplete, and inconsistent data. Data preprocessing encompasses various techniques for transforming the data into a form suitable for applying machine learning algorithms. Missing values in the dataset are handled by using the widely used technique, multiple imputations by chained equations (MICE) (Royston *et al.*,

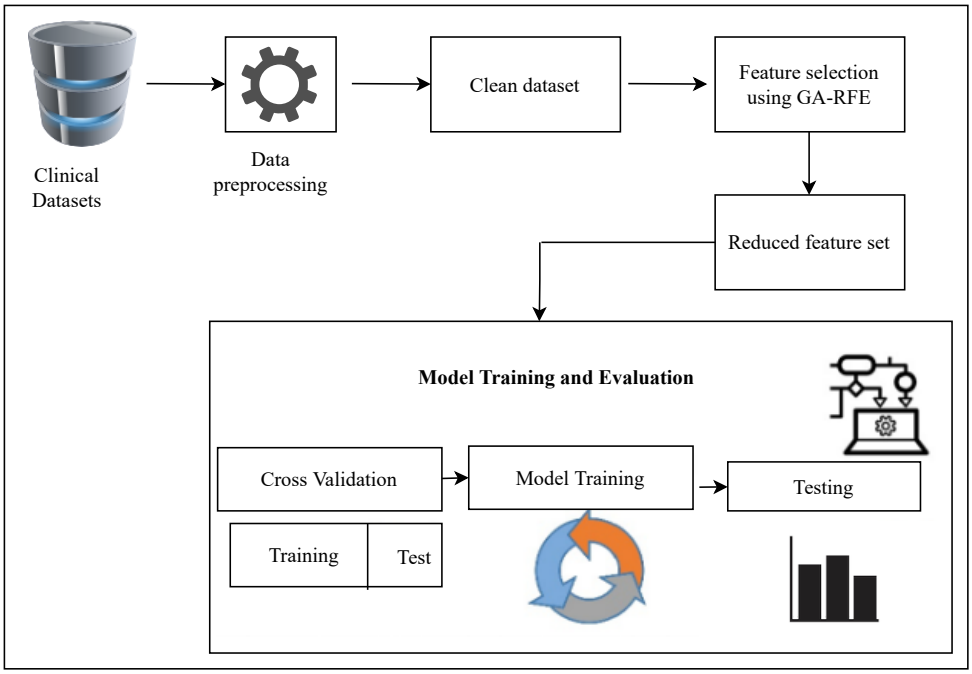

Fig. 3. System model

2011). This technique involves utilizing a regression model to obtain the missing data values from the remaining attributes in the dataset (Azur *et al.*, 2011). Figure 4 shows the implementation details of the MICE algorithm.

4.2 Feature selection

Selecting the most significant features for model training is considered an important step in any classification problem. Feature selection techniques aim to obtain important features from the original dataset by eliminating or removing irrelevant features. In this way, feature selection can be useful in terms of improving the model's performance, lowering the dataset's dimension when dealing with a high number of features, and avoiding the overfitting problem. Multiple feature selection techniques such as a wrapper, and embedded and filter-based techniques are widely utilized to obtain a set of important features. Recursive feature elimination is one of the most important techniques for feature selection which involves eliminating features at each level and re-ranking the rest of the features by retraining the existing features (Lee *et al.*, 2022). Figure 5 shows the working principle of the recursive feature elimination approach.

In this work, we integrate the recursive feature elimination with a genetic approach, GA-RFE to obtain optimal features for model training. The algorithm starts with the randomly generated initial population. A total of 25 individual solutions are selected in the initial population. An objective function based on negated mean square error is used to obtain the fitness of the individual solutions. Two individuals are selected out of the entire population using a tournament-based selection strategy. These individuals are mated using a single-point crossover strategy with a crossover probability of 0.5. The technique chooses a random crossover point and swaps the tails of its two parents to obtain a new offspring. The newly formed offsprings are mutated using the bit-flip mutation technique with the mutation probability of 0.05. Mutation aims to maintain diversity in the population. The RFE algorithm recursively eliminates the redundant features. The importance of each feature is computed while training the model on the training data. The feature importance score is mathematically computed using the equation 1.

$$Score(i) = \sum_{k=1}^{n} (r_{\min} - r_{ik}) \tag{1}$$

Where, $Score(i)$ represents the score of feature i.

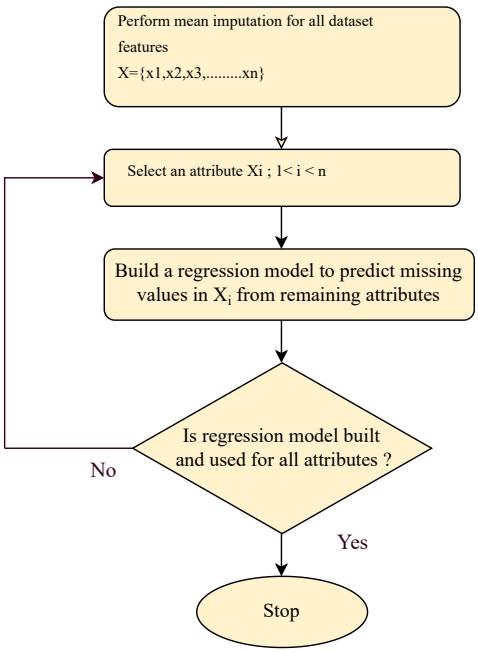

Fig. 4. MICE algorithm

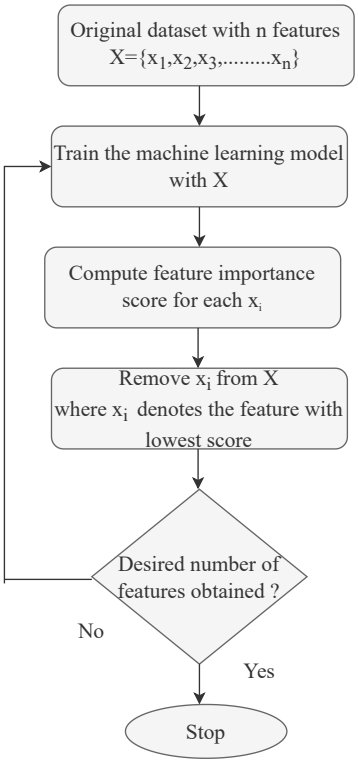

Fig. 5. Recursive feature elimination

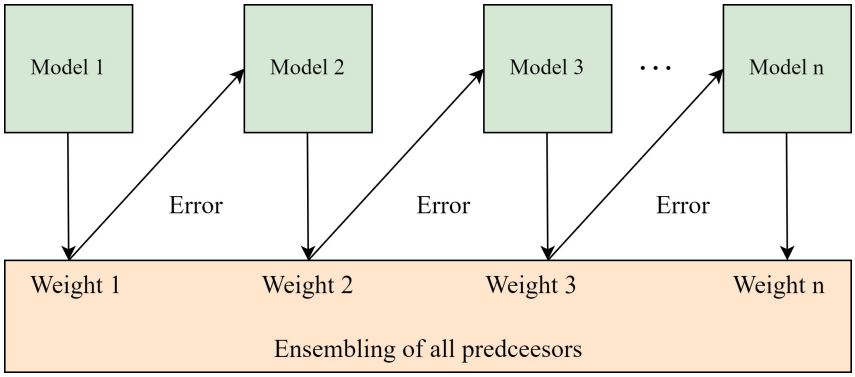

Fig. 6. AdaBoost algorithm

$r_{\min}$ denotes the minimum rank value.
n specifies the total number of features.
r_{ik} denotes the score of i^{th} feature in k^{th} feature set.

All the low score features, representing the weakest features present in the dataset are removed in each step of the algorithm and the classifier is re-trained with the rest of the features. The same procedure is followed until the desired number of generations is performed. At the end of the complete procedure, the optimal number of features are trained and passed to the classifier for training purposes.

4.3 Training of AdaBoost model

The purpose of this phase is to train the model using the GA-RFE selected features. This work utilizes the adaptive boosting (AdaBoost) machine learning technique for classification purposes. AdaBoost is an advanced machine learning technique that is based on boosting class of the ensemble learning and is widely used in diverse application domains to improve the performance of prediction. A decision tree with only one split, is the most widely used algorithm with AdaBoost (Latha *et al.*, 2019). Figure 6 illustrates the working of AdaBoost classifier. AdaBoost involves training the model on the original

dataset. Subsequently, the next weak classifier is trained to eliminate the errors present in the previous model. This procedure is iterated until the errors are reduced and the dataset is correctly predicted. By assigning weights to data objects, numerous subsets of the same dataset are generated. Each weak learner is trained using a different a subset of data. A misclassified instance is more likely to be selected for the next subset because it is assigned a higher weight. In this way, multiple models are trained sequentially. A cost function is used to combine the predictions of weak learners to realize a strong classifier. A classifier with higher accuracy is given greater importance in the final prediction. The AdaBoost algorithm can take as a parameter a weak classifier on which boosting should be performed. The complete workflow of the AdaBoost algorithm is presented in Algorithm 1.

Algorithm 1: Working of AdaBoost algorithm

Input: Dataset with n dimensions, target outcome

Output: Presence or absence of a disease

1: Assign weight to each sample i such that w_i= $\frac{1}{N}$; where i= 1,2,. . . , N

2: **for** m=1 to M **do**

3: Train a classifier $G_m(x)$ to the training dataset with weights W_i

4: Compute

$$err_m = \frac{\sum_{i=1}^{N} w_i I\left(y_i \neq G_m\left(x_i\right)\right)}{\sum_{i=1}^{N} w_i}$$

5: Compute

$$\alpha_m = \log\left(\left(1 - err_m\right)/err_m\right)$$

6: Set $w_i \leftarrow w_i \cdot \exp\left[\alpha_m \cdot I\left(y_i \neq G_m\left(x_i\right)\right)\right], i = 1, 2, \ldots, N$

7: **end for**

8: Output: $G(x) = sign\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$

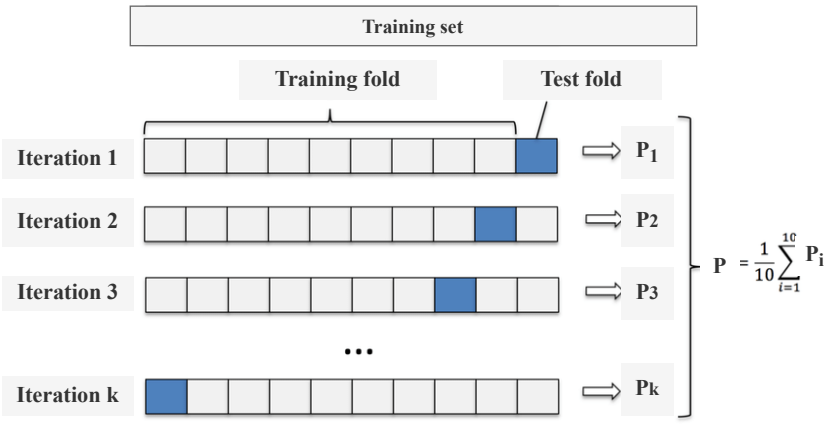

Fig. 7. k-fold cross-validation

4.4 Cross validation

To avoid overfitting the model and to improve the classification performance on the unseen data, a 10-fold cross-validation setup is employed. This strategy works well and provides a less biased model compared to the train-test split. The 10-fold cross-validation technique splits the original dataset into 10 folds, 9 of which are used for training purposes and the rest for testing and error calculations. In this way, the model is trained and tested on each sample in each iteration. The average result of each sample

is obtained and compared with the training results. The key advantage of cross-validation is to ensure that each segment of the original dataset has an equal chance of being present in both the training and test sets. Figure 7 shows the working of k-fold cross-validation.

5. Experimental studies

This section presents the simulation setup and performance analysis of the proposed work. All the experimental work is carried out using the system with the following specifications as mentioned in Table 2.

Table 2. System settings

Parameter	**Value**
Operating system	Windows 10
Primary memory	16 GB
Graphic card	NVIDIA GeForce 1060
Processing unit	Intel Core(TM) i5
Platform	Python 3.10.0

5.1 Data description

To validate the performance and model training, we consider widely available Cleveland and Pima datasets related to two important lifestyle diseases, namely heart diseases, and diabetes, obtained from the University of California (UCI) repository (Diabetes dataset, 2020)(Heart dataset, 2020). These datasets comprise numerous missing values and therefore, the missing values are handled by using the MICE imputation approach. The data distribution of both datasets is presented in Table 3.

Table 3. Datasets distribution

Dataset	**Total number of instances**	**Presence**	**Absence**
Pima	768	268 (34 %)	500 (66 %)
Cleveland	283	157 (55 %)	126 (45 %)

5.2 Performance analysis metrics

The average performance of all the classification models is evaluated on the scale of accuracy, precision, sensitivity, specificity, and F-measure. These measures primarily rely on the values such as TP, TN, FP and FN which are obtained using the confusion matrix (Gokulnath & Shantharajah , 2019) as shown in Table 4.

Table 4. Confusion matrix

Actual values	**Predicted values**	
	Yes	**No**
Yes	TP	FN
No	FP	TN

Table 5 provides the description of each performance metrics. Mathematically, these metrics can be formulated as shown below:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

Table 5. Performance assessment metrics

Measures	Description
Accuracy	Aims to determine the accuracy of predicted data instances.
Sensitivity	It measures the true positive rate of the predicted data instances.
Specificity	Measures the true negative rate of the predicted data instances.
Precision	A classification model's competence to find only the relevant data points.
F-Measure	Represents the weighted average of precision and recall.

$$Sensitivity/Recall = \frac{TP}{(TP + FN)} \tag{4}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{5}$$

$$F - Measure = \frac{(2 * (Precision * Recall))}{(Precision + Recall)} \tag{6}$$

Where, the notations TP, TN, FP and FN specify the number of true positives, true negatives, false positives, and true negatives respectively. True positive specifies that the prediction is correct and the actual value is positive. False-positive specifies that the prediction is wrong and the actual value is positive. Similarly, true negative specifies the prediction is correct and the actual value is negative, and false negative specifies that the prediction is wrong and the actual value is negative.

5.3 Results and discussion

This section deals with the results and analysis of the proposed methodology. First, we discuss the feature analysis to determine the most significant features which should be considered for the study followed by the performance analysis of machine learning algorithms.

5.3.1 Feature selection analysis

Feature selection plays an important role in improving the overall performance of the predictive model. The proposed work utilizes a hybrid GA-RFE approach to achieve the purpose. The cross-validation is employed to assess different subsets of features and select the best scoring feature set. Table 6 show the parameter settings of GA implementation. The genetic algorithm-based recursive feature elimination selects 8 and 7 most-promising features from Pima and Cleveland datasets respectively. Table 7 provides the detailed description about the selected features. The most significant features are used for training purpose.

Table 6. Genetic algorithm parameters

Parameters	Values
Initial Population	25
Encoding	Binary
Selection	Tournament selection
Crossover	Single-point
Mutation	Bit-flip
Mutation Ratio	0.05
Crossover Ratio	0.5
Number of generations	40

5.3.2 Classification performance analysis

To evaluate the performance of the proposed AdaBoost technique, several state-of-the-art algorithms that are widely used for prediction purposes are considered for comparative analysis. These algorithms

Table 7. Features selected by GA-RFE in each dataset

Dataset	Number of selected features	Description
Pima	8	Glucose, BMI, Age, Pedigree, Insulin, Skin thickness, Pregnancies, Blood pressure
Cleveland	7	Age, Cholesterol, Chest pain, Resting ECG, Max heart-rate, Depression, Peak exercise

Table 8. Parameter settings

S.No	Machine learning models	Parameters
1	Decision Tree	criterion=gini sample_split=2 impurity_decrease=0 splitter=best max_depth=8
2	Random Forests	spli_rule=gini max_depth=5 min_node_size=1 importance= impurity number of trees= 500
3	XGBoost	max_depth=6 learning_rate=0.3 reg_lambda=1 min_split_loss=0
4	AdaBoost	n_estimators=50 learning_rate=1 base_estimator= DT

include decision trees (DT), random forests (RF), and XGBoost. Alongside the proposed AdaBoost algorithm, all the aforementioned algorithms are fitted using the GA-RFE selected features. The algorithms are trained using default parameter settings as mentioned in Table 8. Furthermore, a 10-fold CV setup is employed, where 80% of the data is used for model training and the rest 20% is used for testing purposes.

Further, the performance of each algorithm is evaluated on the scales of accuracy, precision, sensitivity, specificity, and F-measure. Table 9 shows the performance of each algorithm based on Cleveland and Pima datasets. The proposed AdaBoost algorithm notable results with the highest classification accuracy of 91.9% and 96.6% on respective datasets. Furthermore, precision, sensitivity, specificity, and f-measure scores also supports the AdaBoost algorithm in classifying diabetes and heart-related diseases compared to other approaches. The performance results of different machine learning algorithms based on the Cleveland and Pima dataset with GA-RFE feature selection criteria are also graphically plotted in Figure 8 and 9. It is evident from the results that the performance of machine learning algorithm is boosted when training the model on GA-RFE selected features.

5.3.3 Comparative analysis with recent works

Numerous researchers in healthcare are using machine learning methods to predict diseases in their earliest stages. The proposed work is responsible for predicting two lifestyle diseases. To unveil the significance of the proposed work, a comparative study of previous works in the current domain is provided in Table 10. The analysis confirms the improved performance of the proposed method over all previous works.

Table 9. Performance of classifiers on Cleveland and Pima dataset with GA-RFE

Dataset	ML Algorithms	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F-measure (%)
Cleveland	DT	84.6	85.4	80.9	80.7	82.9
	Random forest	84.9	86.4	79.2	89.4	82.1
	XGBoost	87.4	87.2	80.5	88.2	82.8
	AdaBoost	91.9	90.7	87.6	91.8	90.1
Pima	DT	91.2	95.4	88.4	94.3	91.2
	Random forest	92.1	96.2	89.1	96.4	92.4
	XGBoost	94.57	96.2	94.7	96.3	95.1
	AdaBoost	96.6	98.3	97.6	98.1	96.5

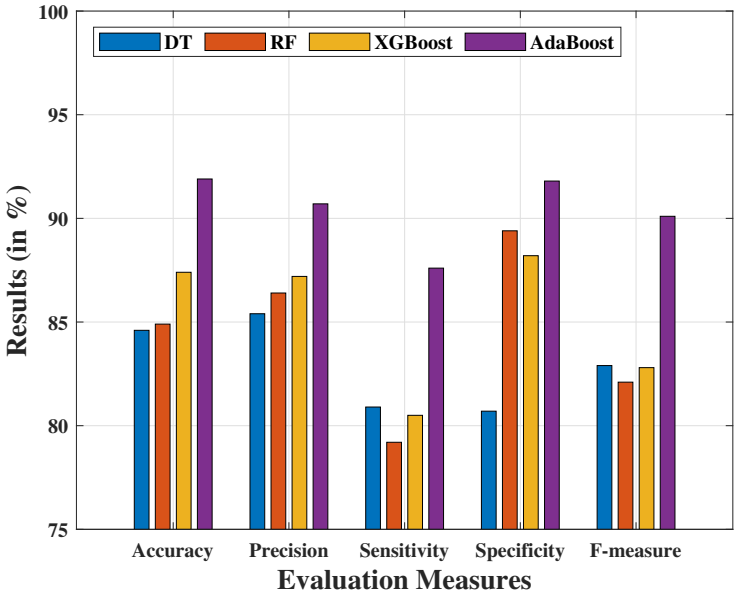

Fig. 8. Performance of the proposed on Cleveland dataset

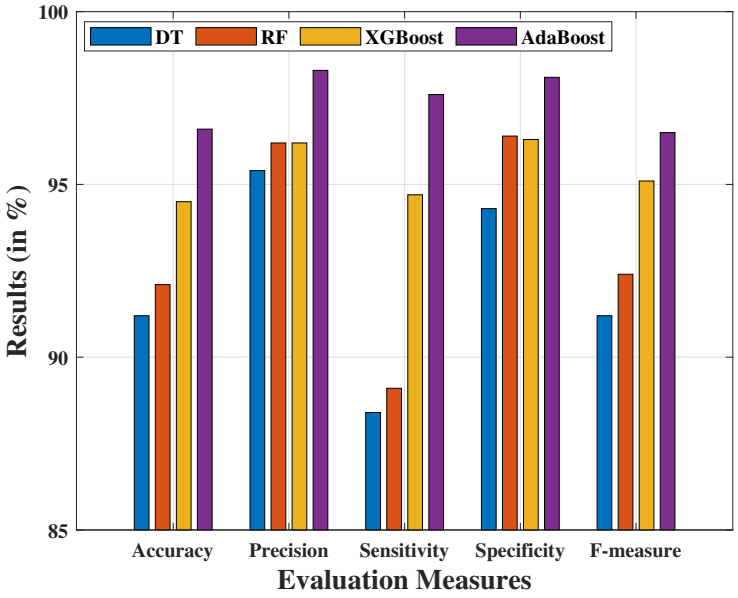

Fig. 9. Performance of the proposed on Pima dataset

Table 10. Comparative study with existing works

Study	Dataset	Feature Selection	Methodology	Validation	Average Accuracy
(Pradhan *et al.*, 2020)	Pima Indian	NA	ANN	NA	85.09
(Chatrati *et al.*, 2020)	Pima and Cleveland	Manual	SVM	NA	74
(Katarya & Meena, 2021)	UCI	NA	Logistic regression	NA	93.40
(Kondababu *et al.*)	Cleveland	NA	Hybrid Random Forest	NA	88.7
Proposed	Pima and Cleveland	GA-RFE	AdaBoost	K-fold validation	94.24

6. Conclusion

In this paper, we propose an intelligent multi-disease prediction framework using GA-RFE and AdaBoost. The proposed system is experimented by using the Cleveland and Pima datasets collected from the UCI repository. Different machine learning algorithms along with the proposed AdaBoost ensemble learning classifier are trained upon GA-RFE selected features under the k-fold cross-validation setup. The performance of the proposed method is validated using several performance evaluation metrics, such as accuracy, precision, sensitivity, specificity, and F-measure. The proposed AdaBoost algorithm attains higher classification accuracy of 91.9% and 96.6% over the Cleveland and Pima datasets respectively. Besides, AdaBoost achieves remarkable values of precision, specificity, sensitivity, and F-measure in comparison to benchmark techniques. This work can significantly be improved by using advanced meta-heuristic algorithms such as ant colony optimization, and NSGA-II for optimal feature selection. Moreover, the current framework can be extended to other real-life diseases such as chronic kidney diseases and COVID-19.

References

Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., Ghazal, T.M., & **Ahmad, M. (2022)**. Prediction of Diabetes Empowered With Fused Machine Learning. *IEEE Access*, **10**, 8529-8538.

Almulla, M. A. (2021). Location-based Expert System for Diabetes Diagnosis. *Kuwait Journal of Science*, **48**(1).

Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & **Gonzales-Yanac, T. (2021)**. Multiple disease prediction using Machine learning algorithms. *Materials Today: Proceedings*.

Azur, M. J., Stuart, E. A., Frangakis, C., & **Leaf, P. J. (2011)**. Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, **20**(1), 40-49.

Chatrati, S. P., Hossain, G., Goyal, A., Bhan, A., Bhattacharya, S., Gaurav, D., & **Tiwari, S. M. (2020)**. Smart home health monitoring system for predicting type 2 diabetes and hypertension. *Journal of King Saud University-Computer and Information Sciences*.

Diabetes dataset (2020). Accessed: 2020-05-15.
URL: `https://Archive.Ics.Uci.Edu/Ml/Datasets/Diabetes`

Dietterich, T. G. (2002). Ensemble learning. *The handbook of brain theory and neural networks*. Arbib MA, **2**, 110-125.

Garg, R. (2021). Improved energy efficiency using meta-heuristic approach for energy harvesting enabled IoT network. *Kuwait Journal of Science*.

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & **De Boer, F. (2021)**. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*, **9**, 19304-19326.

Gokulnath, C. B., & **Shantharajah, S. P. (2019)**. An optimized feature selection based on genetic approach and support vector machine for heart disease. *Cluster Computing*, **22**(6), 14777-14787.

Han, Y., Han, M., Lee, S., Sarkar, A. M., & **Lee, Y. K. (2012)**. A framework for supervising lifestyle diseases using long-term activity monitoring. *Sensors*, **12**(5), 5363-5379.

Heart dataset (2020). Accessed: 2020-05-15.
URL: `http://Archive.Ics.Uci.Edu/Ml/Datasets/Heart+Disease`

Katarya, R., & **Meena, S. K. (2021)**. Machine learning techniques for heart disease prediction: a comparative study and analysis. *Health and Technology*, **11**(1), 87-97.

Kondababu, A., Siddhartha, V., Kumar, B. B., & **Penumutchi, B. (2021)**. A comparative study on machine learning based heart disease prediction. *Materials Today: Proceedings*.

Latha, C. B. C., & **Jeeva, S. C. (2019)**. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, **16**, 100203.

Lee, M., Lee, J. H., & Kim, D. H. (2022). Gender recognition using optimal gait feature based on recursive feature elimination in normal walking. *Expert Systems with Applications*, **189**, 116040.

Mirjalili, S. (2019). Evolutionary algorithms and neural networks. *Studies in computational intelligence*, **780**.

Nandy, S., Adhikari, M., Balasubramanian, V., Menon, V. G., Li, X., & **Zakarya, M. (2021)**. An intelligent heart disease prediction system based on swarm-artificial neural network. *Neural Computing and Applications*, 1-15.

Pradhan, N., Rani, G., Dhaka, V. S., & **Poonia, R. C. (2020)**. Diabetes prediction using artificial neural network. In *Deep Learning Techniques for Biomedical and Health Informatics* (pp. 327-339). Academic Press.

Royston, P., & **White, I. R. (2011)**. Multiple imputation by chained equations (MICE): implementation in Stata. *Journal of statistical software*, **45**, 1-20.

Sajjad, M., Khan, S., Muhammad, K., Wu, W., Ullah, A., & **Baik, S. W. (2019)**. Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *Journal of computational science*, **30**, 174-182.

Sarmadi, H., Entezami, A., Saeedi Razavi, B., & **Yuen, K. V. (2021)**. Ensemble learning-based structural health monitoring by Mahalanobis distance metrics. *Structural Control and Health Monitoring*, **28**(2), e2663.

Senapati, S., Bharti, N., & **Bhattacharya, A. (2015)**. Modern lifestyle diseases: chronic diseases, awareness and prevention. *Int J Curr Res Acad Rev*, **3**(3), 215-23.

Sharma, A. K., & **Verma, K. (2020)**. NSGA-II with ENLU inspired clustering for wireless sensor networks. *Wireless Networks*, **26**(5), 3637-3655.

Singh, A., Dhillon, A., Kumar, N., Hossain, M. S., Muhammad, G., & **Kumar, M. (2021)**. eDiaPredict: An Ensemble-based framework for diabetes prediction. *ACM Transactions on Multimidia Computing Communications and Applications*, **17**(2s), 1-26.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and computing*, **4**(2), 65-85.

Yahyaoui, H., & **Al-Mutairi, A. (2016)**. A feature-based trust sequence classification algorithm. *Information Sciences*, **328**, 455-484.

Submitted: 07/03/2022
Revised: 29/05/2022
Accepted: 30/05/2022
DOI: 10.48129/kjs.splml.19321