Heart Disease Classification and Recommendation by Optimized Features and Adaptive Boost Learning

Pardeep Kumar, Ankit Kumar
Computer Science and Applications, Baba Mastnath University
Asthal Bohar, Rohtak, India

***Abstract*—In recent decades, cardiovascular diseases have eclipsed all others as the main reason for death in both low and middle income countries. Early identification and continuous clinical monitoring can reduce the death rate associated with heart disorders. Neither service is yet accessible, as it requires more intellect, time, and skill to effectively detect cardiac disorders in all circumstances and to advise a patient for 24 hours. In this study, researchers suggested a Machine Learning-based approach to forecast the development of cardiac disease. For precise identification of cardiac disease, an efficient ML technique is required. The proposed method works on five classes, one normal and four diseases. In the research, all classes were assigned a primary task, and recommendations were made based on that. The proposed method optimises feature weighting and selects efficient features. Following feature optimization, adaptive boost learning using tree and KNN bases is used. In the trial, sensitivity improved by 3-4%, specificity by 4-5%, and accuracy by 3-4% compared to the previous approach.**

***Keywords*—Heart disease prediction; heart disease; machine learning; optimization; multi-objective features**

I. Introduction

The cardiovascular system, which also includes the lungs, is powered by the heart, a muscular organ which circulates blood throughout the body. The cardiovascular system includes a blood vessel network in addition to arteries, veins, and capillaries. Blood is distributed by these blood vessels all over the body. Cardiac disorders, also known as heart illnesses, are defined by deviations in the normal blood circulation of heart. The leading causes of death globally are heart disorders. Based on a survey conducted by the World Health Organization (WHO), strokes and heart attacks are responsible for 17.5 million deaths worldwide. Over 75% of deaths from heart disease happen in middle-income and low-income countries. In addition, strokes and heart attacks account for 80% of CVD-related mortality [1], [2]. In light of this, the mortality rate from cardiovascular problems can be reduced with the use of early recognition of cardiac abnormalities and prediction tools. Predictive models for cardiovascular disorders can now be developed with the help of the vast amounts of patient data that are readily available thanks to the expansion of modern healthcare infrastructure (i.e. Big Data inside the Electronic Health Records System). Machine learning is a technique for finding new information by analysing large datasets from several angles. Numerous records on patients' health, disease diagnoses, and other topics are created every day in the modern healthcare sector [3], [4], [5]. Many methods for unearthing similarities or hidden patterns in data can be found using machine learning [6], [7], [8], [9]. Machine learning has proved to be beneficial when it comes to making predictions and judgments based on the massive amounts of data collected by businesses in the healthcare industry [7], [8], [9], [10], [11]. Machine learning allows computers to automatically learn from data sets and improve their performance based on past experiences with little to no human input. Each time a ML algorithm makes a good call, it gets smarter. Consequently, in this research, we present a ML algorithm for the development of a cardiovascular disease forecasting tool.

A. GAP in Previous Work

The fundamental challenge with heart disease classification is that there is a limited dataset and just five classifications, thus learning efficiently is critical. In prior work, the following problem was discovered:

- Previous research has ignored feature overlaps and increased noise during learning [1], [2].
- Formerly, the emphasis was mostly on accuracy, which ranged from 40-50% in the case of five classes [4], [5].
- Do not improve the features based on their classification capacity [8], [9].
- Learning with a single classifier that is highly polynomial and increases over fitting [11].
- The majority of research is focused on binary categorization, yet this is not a true condition [12], [13].

B. Contribution of Research

- Apply entropy and information gain constraints to optimize features.
- Optimize feature selection and feature weights by using a genetic algorithm to maximize the Pareto surface.
- Work on feature-by-feature and weighted-features analysis of several performance metrics.
- In brief, optimize feature space and learning through optimizing classifiers.
- Focusing on five classes with high accuracy, sensitivity, and specificity.

II. Related Work

The paper gives an in-depth analysis of how ML can be used to treat cardiovascular disease. We also examine numerous popular literature on predicting the course of heart disease.

Ali et al. (2021) determines which machine learning classifiers provide the most accurate performance for diagnostic applications. Several supervised ML methods were implemented and compared in the prediction of heart disease. For all deployed algorithms except KNN and MLP, feature significance scores were assessed for each feature. All of the features were sorted based on their importance score to identify which ones offer the most reliable predictions of heart illness. Using a heart illness database from Kaggle and three-classification algorithms depending on KNN, DT, and RF, the analysis revealed that the RF method achieved 100% sensitivity, specificity, and accuracy. In this study, Katarya et al. (2021) summarized a portion of the expertise automated processes. Prediction and Feature selection are key components of every automated process. By selecting features effectively, one can attain improved heart disease prediction outcomes. The researchers have demonstrated useful methods for selecting attributes, including the hybrid grid search method and random search algorithm. As per Princy et al. (2020), a cardiac database is classified utilizing multiple cutting-edge Supervised ML algorithms for disease prediction. The findings show that the DT classifying model accurately diagnosed cardiovascular problems more so than the LR, NB, SVM, RF, and KNN approaches. 73% of the time, the Decision Tree produced the best outcome. This strategy could aid physicians in predicting the onset of heart problems and providing adequate treatment. Shah et al. (2020) offers several heart disease-related variables and a model based on supervised learning techniques like DT, NB, KNN, and RF. It utilizes the current database from the Cleveland dataset of UCI's heart disease patient repository. There are 76 attributes and 303 instances in the collection. For the purpose of verifying the efficacy of different approaches, only 14 of these 76 attributes are chosen for testing. The purpose of this report is to illustrate the occurrence of heart disease among patients. As per the results, K-nearest neighbour provides the highest accuracy. Sharma et al. (2020) makes a ML model that uses the relevant parameters to predict heart disease. The scholars used a standard UCI Heart disease prediction database for this research. This database has 14 key factors that are related to heart disease. For the creation of the model, ML techniques such as RF, SVM, DT, and NB, have been utilized. The research has also attempted to identify correlations between the numerous qualities present in the dataset by employing standard ML techniques and then employing these correlations to accurately forecast the likelihood of heart disease. When compared to other ML algorithms, the RF technique provides superior prediction accuracy and processing speed. The use of this system to aid in making decisions, this model may be beneficial to medical professionals in their clinic. Krishnan et al. (2019) used two supervised algorithms for data mining on a dataset to determine the likelihood of a patients experiencing heart disease, which were analyzed using classification models such as DT Classification and NB Classifier. These two algorithms were compared on a similar dataset to evaluate which one was the most accurate. The Decision Tree model accurately predicted the cardiovascular disease patient 91% of the time, while the Nave Bayes classifier correctly guessed the heart disease patient 87% of the time. Mohan et al. (2019) strategies and related cardiovascular disease prediction via hybrid ML techniques, with the purpose of discovering essential aspects by applying ML hence boosting the accuracy in the detection of cardiovascular illness. The expectation model consists of common feature groupings and their numerous permutations. The predictive model for cardiovascular illness with hybrid RF using a linear model allows the research teams to produce an improved exhibition level with a precision level of 88.7 percent (Table I). Individuals also informed about various data mining methods and assumption methods, for example, LR, KNN, NN, SVM, and Vote, which have recently been fairly popular in distinguishing and predicting heart disease. Santhana et al. (2019) detect cardiovascular disease in male patients using categorization approaches. This document offers exhaustive information on Cardiac Heart Diseases, including Risk Factors, Facts, and Frequent Type. WEKA seems to be the Data Mining tool used, and it is a great Computational Tool for Bioinformatics Fields. All three WEKA interfaces are used here; NB, ANNs, and DT are the main methods of data mining employed in this system to forecast heart disease. DTs such as C4.5, CART, CHAID, ID3, and J48 Algorithms, and NBs Techniques are commonly used for prediction. Gavhane et al. (2018) trained and examined the dataset using the multi-layer perceptron (MLP) neural network algorithm. Any number of input layers, output layers, and hidden units may be present in this algorithm. To achieve their desired effect, these hidden layers connect all input nodes to all output nodes. This bond is allocated weights. To achieve equilibrium in the perceptron, a second identity input, bias, with weight b, would indeed be introduced into the node. The nature of the nodes' connections to one another (feedforward or feedback) is determined by the task being performed. Li et al. (2018) have created an efficient ML-based approach for the diagnosis of cardiac disease. System design utilizes ML classifiers including ANN, K-NN, NB, SVM, and DT. Four classic feature selection methods, comprising MRMR, Relief, LLBFS, and LASSO, in addition, the issue of feature selection was addressed by employing a unique feature selection method. The system uses the LOSO cross-validation approach to select the optimal hyperparameters. The system is evaluated utilizing the Cleveland cardiovascular disease database.

III. Proposed System

A. Dataset

In experiment use ''https://archive.ics.uci.edu/ml/datasets/heart+disease" data set for classification and recommendation in which total 303 instances, five classes and thirteen features (see Fig. 1 and 2).

Steps for Analysis

Entropy=$\sum_{j=1}^{N} P_i ly_s P_i$ ..... (1)

ly_srepresent classes

$P_i$probability of Instance

Information gain=1-Entropy. ....(2)

Step 1: Input heart disease dataset with features and labels.

Step 2: Features optimize by multi-objective optimization by this process given the efficient weight to features. In equation (3), E represent Entropy IG represent Information gain

TABLE I. ML PREDICTION OF A VARIETY OF HEART DISEASE AILMENTS

Ref	Year	**Aim**	Techniques	Feature/Tool	Dataset	Findings
[1]	[2021]	Model-based prediction of coronary heart disease using supervised ML	Supervised ML algorithms	Weka version 3.8.3	Kaggle	Accuracy rates of 100% were achieved by all three methods (RF, KNN, and DT).
[3]	[2020]	Using ML for early-stage prediction of heart disease	Supervised ML algorithms	Features: height, weight, Age, ap_hi ap_lo, gender, gluc, smoke, cholesterol, intake alco, cardio, active	Kaggle	The DT classification model outperformed Naive Bayes, RF, LR, KNN, and SVM in predicting cardiovascular illnesses.
[5]	[2020]	ML for Predicting Heart Disease	Supervised learning algorithms	WEKA tool	Cleveland database	According to the findings, KNN yields the best accuracy score.
[6]	[2020]	Prediction of cardiac events using ML	ML algorithms	WEKA tool	Cleveland heart disease database	RF gives more accurate predictions in less time.
[7]	[2019]	Hybrid ML techniques can accurately predict heart disease.	Hybrid RF with a linear model	Features: sex, Age, cp, chol, treetops, FBS, thalach, restecg, exang, olpeak, ca, slope, that, target	Cleveland database	HRFLM was quite accurate in predicting cardiovascular disease.
[8]	[2018]	Cardiovascular disease detection utilizing a real-time cardiac health surveillance system and ML algorithms.	ML algorithms	The WEKA data mining application version 3.8.2	Cleveland Heart Disease and Statlog Heart Disease dataset	The suggested feature selection approach is workable with SVM classifiers for building an advanced smart system for cardiac illness diagnosis.

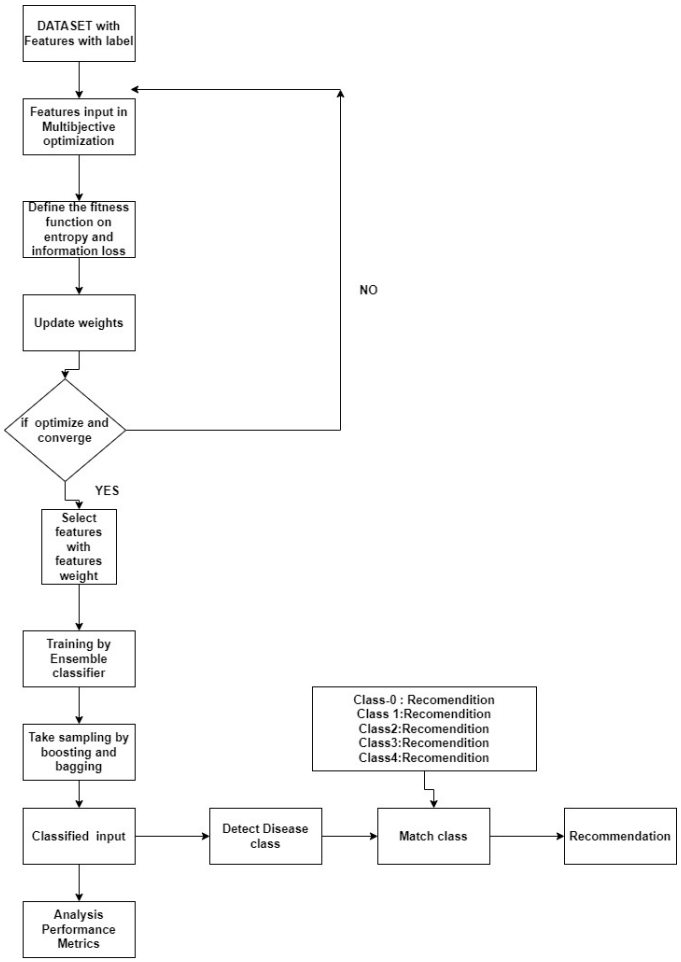

Fig. 1. Proposed classification and recommendation approach.

$$w_j^i = \begin{pmatrix} E & if\ E > IG \\ 0 & if\ IG = 0 \\ IG & otherwise \end{pmatrix} (3) \qquad (1)$$

Step 3: After crossover finish go to efficient Pareto space using following activation function or fitness function.

$$d\left(C_i, C_J\right) = \sum_{i=1}^{M} \left(W_i X_{i,J} - W_i X_i\right| \ldots\ldots\ldots\ldots..(4)$$

It finds the two-class distance and according to it finds Pareto space, here C_i, C_J are the classes, W_i are the weights as per the features.

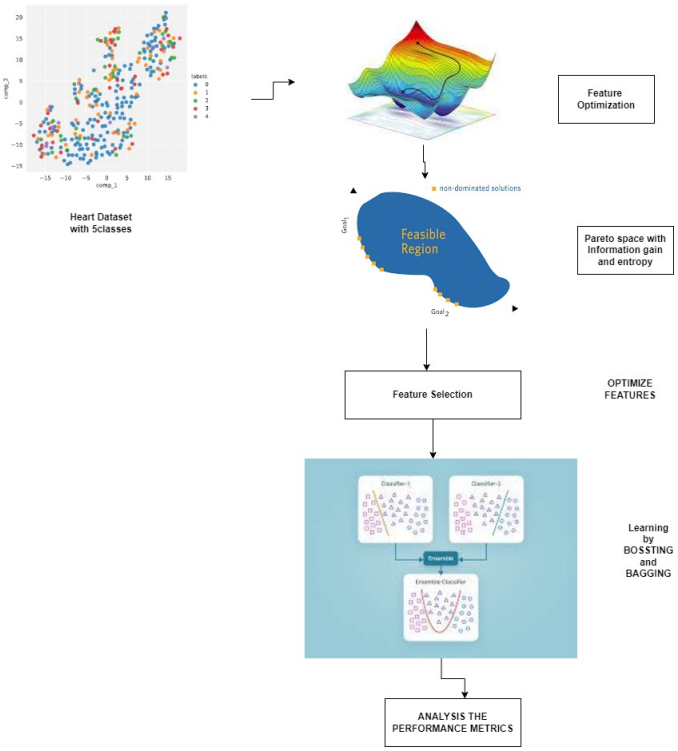

Fig. 2. Proposed classification approach.

Step 4: After finding the Pareto space optimize weight.

$$\delta^z = \frac{(\delta^z|^T.1}{\sum_{i=1}^{M}(\delta^K|^T.1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

By equation (5) find the optimal solution in space of δ^z then

Step 5:

$$z = \max_k \delta^k \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

$$w = argmax\ (z) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(7)$$

By this find the maximize optimal weights o features

Step 6: After optimizing the weights of the weighted feature learn by classifier.

$$C_N(.) = \sum_{i=1}^{N} C_N * W_N(.) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(8)$$

$C_N(.)$Boosting Classifier

C_NN number of weak classifier

$W_N(.)$weight of features

After boosting all the possibilities send it to Bagging approach

$$B = \sum_{i=1}^{K} C_K(.) * \sum_{i=1}^{K} W_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9)$$

$$\text{CM}=\alpha C_N(.)+1-\alpha(\delta_K) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(10)$$

By (9) use bagging and use (10) for combining both develop a classification model. Here α is the learning parameter $\alpha[0,1]$

Step7: After step 6 recommendation part, according to Fig. 1, test one instance and according to predict class recommend the suggestion

IV. Experimental Results and Discussion

. Adaptive Boosting KNN with 60% accuracy outperforms Adaptive Boosting-Tree (54.34%), Adaptive Boosting (45.34%), KNN (40.2%), and SVM (40.11%) while using 5 features. Adaptive Boosting KNN with 70% accuracy outperforms Adaptive Boosting-Tree (50.12%), SVM (41.2%), Adaptive Boosting (40.12%) [2], and KNN (34.34%) [3] when using 6 features. With 8 features, however, Adaptive Boosting KNN (67.34%) achieves the highest accuracy, followed by Adaptive Boosting (47.44%), Adaptive Boosting-Tree (45.23%), KNN (42.12%), and SVM (40.34%). Similarly, with 10 features, Adaptive Boosting KNN (65%) achieves the highest accuracy, followed by Adaptive Boosting (48.12%), Adaptive Boosting-Tree (47.34%), SVM (39.5%), and KNN (36.44%). Adaptive Boosting KNN delivers superior accuracy with 12 and 13 features compared to previous approaches (Table II).

The comparison of the various features depending on their degrees of accuracy is shown in Fig. 3. When compared to other approaches, the accuracy that Adaptive Boosting KNN provides is far superior (Table III).

hows the sensitivity of features derived from various approaches. Adaptive Boosting KNN with 96.23 sensitivity

Proposed Algorithm
Input dataset with features and label
1. **Mutation and crossover by eq. (3)**
2. **Extract newly generated vector**
3. **Update the fitness function of eq. (4) if optimize then go to the next step else go to the 3^{rd} step**
4. **Optimize fitness function and find the optimize Pareto space weight by eq. (5) and get weights eq. (6) and eq. (7)**
5. **Learning by Boosting eq. (8) and Bagging eq. (9)**
6. **Make Classifier model eq. (10) and analysis**
7. **Output <-Accuracy, Precision and Recall**

TABLE II. ACCURACY OF FEATURES BASED ON DIFFERENT METHODS

Features	KNN	SVM	Adaptive Boosting	Adaptive Boosting-Tree	Adaptive Boosting KNN
5	40.2	40.11	45.34	54.34	60
6	34.34	41.2	40.12	50.12	70
8	42.12	40.34	47.44	45.23	67.34
10	36.44	39.22	48.12	47.34	65
12	43.22	40	43.23	50.12	56
13	35.5	41.34	43.2	52.33	60

TABLE III. SENSITIVITY OF FEATURES BASED ON DIFFERENT METHODS

Features	KNN	SVM	Adaptive Boosting	Adaptive Boosting-Tree	Adaptive Boosting KNN
5	73.45	78.34	88.23	90.1	96.23
6	72.43	74.34	88.34	92.3	94.23
8	70.23	74.35	84.56	90	93.45
10	71.33	73.45	88.23	87.3	92.34
12	72.33	70.12	87.34	89.13	90.23
13	71.1	70.32	84.3	86.12	93

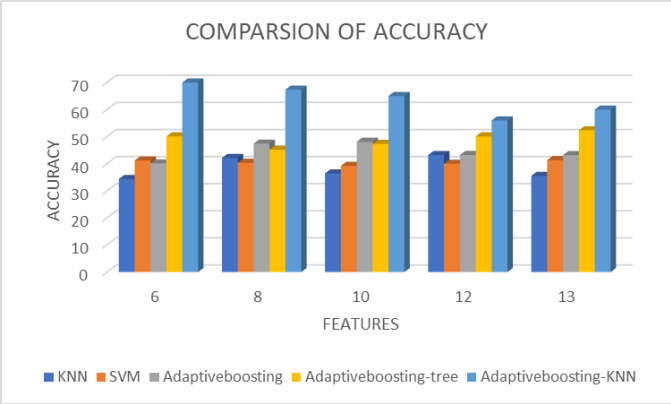

Fig. 3. Accuracy-based comparison.

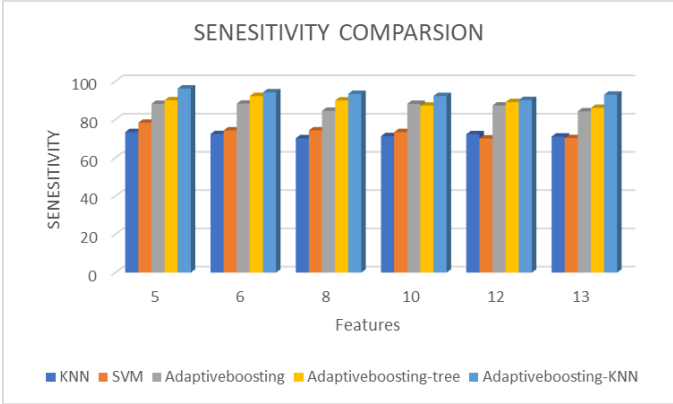

Fig. 4. Sensitivity-based comparison.

value outperforms Adaptive Boosting-Tree (90.1), Adaptive Boosting (88.23), SVM (78.34), and KNN (73.45) while using 5 features. Adaptive Boosting KNN with 94.23 sensitivity outperforms Adaptive Boosting-Tree (92.3), Adaptive Boosting (88.34), SVM (74.34), and KNN (72.43) when using 6 features. With 8 features, however, Adaptive Boosting KNN (93.45) achieves the highest sensitivity, followed by Adaptive Boosting Tree (90), Adaptive Boosting (84.56), SVM (74.35), and KNN (70.23). Similarly, with 10 features, Adaptive Boosting KNN (92.34) achieves the highest sensitivity, followed by Adaptive Boosting-Tree (87.3), Adaptive Boosting (87.34), SVM (73.45), and KNN (71.33). Adaptive Boosting KNN delivers superior sensitivity with 12 and 13 features compared to previous approaches.

Fig. 4 shows a comparison of the features according to their sensitivities. Adaptive Boosting KNN offers significantly higher sensitivity than the other competing methods (Table IV).

hows the specificity of features derived from various approaches. Adaptive Boosting KNN with 60.0 sensitivity value outperforms Adaptive Boosting-Tree (56.23), Adaptive Boosting (54.23), SVM (50.23), and KNN (45.12) while using 5 features. Adaptive Boosting KNN with 70.23 specificity outperforms Adaptive Boosting-Tree (60.23), Adaptive Boosting (53.12), KNN (46.23), and SVM (45.12) when using 6 features. With 8 features, however, Adaptive Boosting KNN (75.23) achieves the highest specificity, followed by Adaptive Boosting Tree (69.12), Adaptive Boosting (59.12), KNN (50.12), and SVM (42.34). Similarly, with 10 features, Adaptive Boosting KNN (60.13) achieves the highest specificity, followed by Adaptive Boosting-Tree (55.23), Adaptive Boosting (53.23), KNN (52.34), and SVM (40.12). Adaptive Boosting KNN delivers superior specificity with 12 and 13 features compared to previous approaches.

Fig. 5 shows a comparison of the features according to their specificities. Adaptive Boosting KNN offers significantly higher specificity than the other competing methods.

TABLE IV. SPECIFICITY OF FEATURES BASED ON DIFFERENT METHODS

Features	KNN	SVM	Adaptive Boosting	Adaptive Boosting-Tree	Adaptive Boosting KNN
5	45.12	50.23	54.23	56.23	60
6	46.23	45.12	53.12	60.23	70.23
8	50.12	42.34	59.12	69.12	75.23
10	52.34	40.12	53.23	55.23	60.13
12	55.23	35.12	68.23	70.12	72.34
13	56.12	40.12	58.12	60.12	62

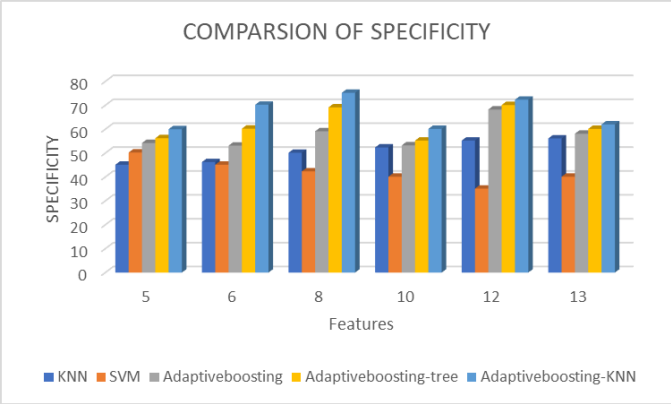

Fig. 5. Specificity-based comparison.

V. OBSERVATION OF RESULTS

- In the results, we compared existing and proposed adaptive boost approaches. There are three variants of adaptive boost in the results: one is basic adaptive boost, another is hybridized with tree, and the third is hybridized with KNN.
- Using multi–objective genetic optimization, features are given the appropriate weighting in all of the proposed methods. It makes sure that features don't overlap and boosts performance, as shown in the figures above.
- Adaptive boost tree improves all measures of performance because entropy and information gain map well on tree-based approaches.
- By maximizing performance improvement in sensitivity, the proposed model's recall value is raised.

VI. CONCLUSION

The long-term preservation of people's existence and the early detection of irregularities in heart problems will be made possible by recognizing the processing of primary health records of heart data. In order to process the raw data and deliver a new and unique insight towards heart disease, methods based on machine learning were applied in this study. Prediction of heart disease is difficult and crucial in the medical industry. However, if the disease is discovered in its initial stages and preventive measures are implemented as soon as feasible, the fatality rate can be significantly reduced. The proposed approach employs a five-class classification system to improve the diagnosis of specific heart disease and the subsequent recommendation. As a result, improving classification sensitivity is a significant task. Sensitivity is improved through feature optimization, and ensemble learning is enhanced through bagging and boosting. In comparison to traditional SVM and KNN methods, a 5% gain in sensitivity is highly significant.

In future, we enhance this work using non-linear mapping by deep learning approach and make optimize latent space for reducing overlapping between classes

REFERENCES

[1] S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system," *World Journal of Engineering and Technology*, vol. 6, no. 4, pp. 854–873, 2018.

[2] E. Maini, B. Venkateswarlu, and A. Gupta, "Applying machine learning algorithms to develop a universal cardiovascular disease prediction system," in *International Conference on Intelligent Data Communication Technologies and Internet of Things*. Springer, 2018, pp. 627–632.

[3] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, 2021.

[4] R. Katarya and P. Srinivas, "Predicting heart disease at early stages using machine learning: a survey," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 302–305, 2020.

[5] R. J. Princy, S. Preetha, P. Parthasarathy, A. R. S. H. Jose, S. Lakshminarayanan, and Jeganathan, "Prediction of cardiac disease using supervised machine learning algorithms," *2020 4th international conference on intelligent computing and control systems (ICICCS)*, pp. 570–575, 2020.

[6] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, no. 6, pp. 1–6, 2020.

[7] V. Sharma, S. Yadav, and M. Gupta, "Heart disease prediction using machine learning techniques," *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pp. 177–181, 2020.

[8] S. Mohan, , G. H. Thirumalai1, and Srivastava, pp. 81 542–81 554, 2019.

[9] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," *2018 second international conference on electronics, communication and aerospace technology (ICECA)*, pp. 1275–1278, 2018.

[10] S. Krishnan and S. Geetha, "Prediction of Heart Disease Using Machine Learning Algorithms," *2019 1st international conference on innovations in information and communication technology (ICIICT)*, pp. 1–5, 2019.

[11] S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system," *World Journal of Engineering and Technology*, vol. 6, no. 4, pp. 854–873, 2018.

[12] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, pp. 3–21, 2020.

[13] J. Li, Ping, A. U. Haq, J. S. U. Din, A. Khan, A. Khan, and Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol. 8, pp. 107 562–107 582, 2020.