**Improving Heart Disease Prediction Using Modified Genetic Algorithm**

Anmol Thakur  
*Department of Computer Science, Ramrao Adik Institute of Technology*Email: [anmolt2510@gmail.com](mailto:anmolt2510@gmail.com)

Kairavi Patra*Department of Computer Science, Ramrao Adik Institute of Technology*Email: [patrakairavi@gmail.com](mailto:patrakairavi@gmail.com)

Anya Thakur*Department of Computer Science, Ramrao Adik Institute of Technology*Email: [anyaa.thakur@gmail.com](mailto:anyaa.thakur@gmail.com)

***ABSTRACT -***

***INDEX TERMS - Feature selection, genetic algorithm, heart disease prediction, machine learning***

1. **INTRODUCTION**
2. **RELATED WORK**
3. **PROPOSED METHODOLOGY**

In heart disease prediction, the selection of features and the right machine learning (ML) or deep learning (DL) model plays a crucial role in determining the performance and accuracy of diagnostic systems. The heart disease dataset, which consists of diverse patient attributes like age, sex, type of chest pain, blood pressure, and cholesterol level, serves as the foundational dataset for creating these prediction models. Yet, not all the features in the dataset are of equal significance, and the presence of irrelevant or redundant features can negatively affect model performance and interpretability.

In order to overcome this problem, we recommended the application of a Modified Genetic Algorithm (GA) for feature selection. Feature selection is an important process in machine learning because it is responsible for choosing the most significant variables from a database and hence improving the accuracy of the model without increasing its complexity. Modified GA is a sophisticated technique that utilizes evolutionary principles to select a best feature subset, thereby improving the performance of the predictive model.

Overall model pipeline is represented in **Figure 1**, involving data preprocessing, feature selection via the Modified Genetic Algorithm, training of the model (with and without GA), and ultimate performance assessment.

In the initial step of our approach, we compared a number of popular ML and DL models, including Random Forest, XGBoost, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, Decision Tree, LightGBM, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Multi-Layer Perceptron (MLP). These models were validated under typical train-test protocols, employing different split ratios (70–30, 75–25, and 80–20) as well as k-fold cross-validation to evaluate their generalization and robustness without feature selection. Subsequently, the Modified GA was used to conduct feature selection for all of the models. This algorithm iterated over a population of potential feature subsets, evaluated their performance against predictive precision, and selected top-performing sets of features upon which to train models. The ultimate goal of this process was to determine how feature selection could improve the performance of ML and DL models by identifying the most useful features and removing unnecessary complexity.

With this method, we sought to illustrate how the Modified GA, when applied to feature selection, can improve the predictive capabilities of current ML and DL models for heart disease diagnosis to produce models that are not only more accurate but also more interpretable and efficient.

**A. DATA COLLECTION**

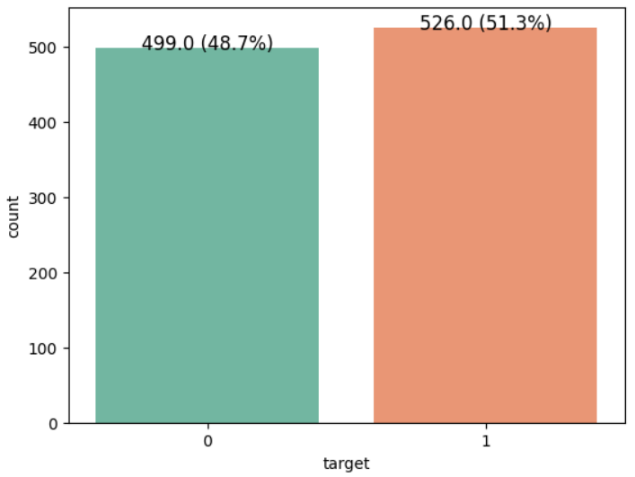
We employed the Heart Disease Dataset from Kaggle. The dataset contained 1025 records and 14 features, with the target being the presence or absence of heart disease **(see Table)**. The features included categorical and continuous variables such as age, sex, type of chest pain, blood pressure, level of cholesterol, and so on, which are of crucial importance for predicting heart disease. This dataset has broad usage in the field of medical and machine learning research, hence making it the best fit for our study on the improvement of heart disease prediction models.

1. **DATA PREPROCESSING**

Data preprocessing was carried out to prepare the dataset for modelling via cleaning, transformation, and formatting. It aimed at ensuring the data used in modelling was uniform, pertinent, and prepared for analysis. This was achieved through different operations such as data cleaning, feature scaling, and balancing datasets.

Data cleaning was the initial step in preprocessing. In this step, we focused on handling missing or inconsistent data. Our data was extensively analysed, and it was found to contain no missing values, thus guaranteeing the correctness of the trained models.

We proceeded to evaluate whether the dataset was balanced. It is said to be balanced if the number of samples in each class is approximately the same. In our data, the target attribute (heart disease present or absent) consisted of 526 samples of class 1 (disease present) and 499 samples of class 0 (disease absent), indicating that the data was balanced. This was significant since unbalanced datasets have the potential to result in skewed predictions, where the model tends to favour the majority class. **Figure [X]** shows the percentage distribution of the target classes in the dataset.



**Figure. Heart Disease Class Distribution.**

Next, we performed feature scaling, which scales the range of independent variables so that all the features contribute equally to the model's performance. Normalization and standardization are

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Dataset Column Name** | **Description** | **Value Type** |
| 1 | age | Age of the individual in years | Numerical |
| 2 | sex | Gender (0 = Female, 1 = Male) | Categorical |
| 3 | cp | Chest pain type (0–3: Typical, Atypical, Non-anginal, Asymptomatic) | Categorical |
| 4 | trestbps | Resting blood pressure (in mm Hg) | Numerical |
| 5 | chol | Serum cholesterol in mg/dl | Numerical |
| 6 | fbs | Fasting blood sugar > 120 mg/dl (1 = True, 0 = False) | Binary |
| 7 | restecg | Resting ECG results (0 = Normal, 1 = ST-T abnormality, 2 = LV hypertrophy) | Categorical |
| 8 | thalach | Maximum heart rate achieved during the test | Numerical |
| 9 | exang | Exercise-induced angina (1 = Yes, 0 = No) | Binary |
| 10 | oldpeak | ST depression induced by exercise relative to rest | Numerical |
| 11 | slope | Slope of the ST segment (0 = Upsloping, 1 = Flat, 2 = Down sloping) | Categorical |
| 12 | ca | Number of major vessels (0–3) colored by fluoroscopy | Numerical |
| 13 | thal | Thalassemia (0 = Normal, 1 = Fixed defect, 2 = Reversible defect) | Categorical |
| 14 | target | Presence of heart disease (1 = Disease, 0 = No disease) | Binary |

**Table. Features used in the study.**

two popular methods of feature scaling.

Normalization scales data into a range between 0 and 1, according to the formula:

|  |  |
| --- | --- |
|  | (1) |

where is the original value, ​ is the minimum value, and is the maximum value in the feature. It is helpful when models are based on distances, like in K-Nearest Neighbors (KNN).

Conversely, standardization rescales the data to have a mean of 0 and a standard deviation of 1, which suits models where data is assumed to follow a Gaussian distribution, such as Logistic Regression and Support Vector Machines (SVM). The formula is:

|  |  |
| --- | --- |
|  | (2) |

where is the original value, is the mean, and is the standard deviation of the feature.

After this, we divided the dataset into training and test sets based on several strategies. First, the dataset was split using three different strategies: 70% training and 30% testing, 75% training and 25% testing, and 80% training and 20% testing. This

allowed us to test how well the models performed

with different amounts of training data. In addition to these divisions, we used k-fold cross-validation with k=5, where data were split into five equal-sized

sections (or folds), and the model trained on four sections and tested on the other section. We did this five times, with each segment being the test set once, to give a less variable measure of model performance.

1. **FEATURE SELECTION USING MODIFIED GENETIC ALGORITHM**

We employed a Modified Genetic Algorithm (GA) to perform feature selection in our research to enhance the performance of heart disease prediction with various machine learning (ML) and deep learning (DL) models. Feature selection is a significant process in machine learning as it helps to choose the best features from a data set by removing dimensionality and enhancing the performance of a model. Evolutionary methods like Genetic Algorithms (GA) are frequently utilized for feature selection due to the fact that they are stable in navigating large and complex search spaces. **[REF]** The GA adapted accomplishes this by maximizing feature selection which most leads to accurate predictions.

**Figure 2** shows the Modified Genetic Algorithm process, from population initialization and fitness evaluation to crossover, mutation, elitism, and convergence for feature selection. We start by initializing the feature subset as a chromosome. Each chromosome is a binary vector where each entry corresponds to a feature of the dataset. The entry value of 1 signifies that the respective feature belongs to the subset, whereas the value 0 signifies exclusion of the feature. The population contains several chromosomes, each one corresponding to a potential feature subset. The population is generated randomly initially.

The core of the GA is the fitness function, which estimates the quality of each feature subset by training a machine learning model and its performance. We employed a Stratified K-Fold Cross-Validation (CV) method to estimate the performance of each chromosome. The data is split into 5 folds such that each fold has an equal distribution of classes. The training set is trained on and validated on the model for each fold.

The fitness value of a chromosome is taken as the average accuracy over all the folds in the cross-validation. The aim of the algorithm is to maximize this fitness value, i.e., to identify the feature subset that provides the maximum classification accuracy.

Mathematically, the fitness function can be expressed as:

|  |  |
| --- | --- |
|  | (3) |

where is the number of folds (5 in our case), and

is the accuracy obtained from fold.

After the fitness of all chromosomes is computed, the second step is to produce offspring by crossover. In our proposed GA, we used a simple approach where parents are chosen on the basis of their fitness. The first 4 chromosomes are selected as parents. These parents contribute to the creation of new chromosomes (offspring) that may inherit their features.

Instead of using traditional two-parent crossover, we used mean-based crossover, where the offspring is generated by averaging the feature selection patterns of the parents. The offspring is calculated as the mean of the selected features from the parent chromosomes:

|  |  |
| --- | --- |
|  | (4) |

where is the offspring, ​ represents the parent, and is the number of parents. This approach guarantees that the offspring will inherit traits from both parents with the ability to generate a diverse but good-performing feature subset.

Following the creation of the offspring, we add mutation to keep genetic diversity and avoid the algorithm getting stuck in local optima. During the mutation step, every gene (feature) in the offspring has a possibility to flip its value (e.g., from 0 to 1 or from 1 to 0) with a probability that is set by the mutation rate. This step makes sure that the search process keeps exploring new sets of features.

For maintaining the optimal, we employed elitism such that the best individuals from the current generation are carried to the next generation without any changes. This way, the optimal of feature sets do not get eliminated in the process of evolution. In our model, when generating offspring, we compare offspring's fitness with the least fit one in the current population. If the offspring is fitter, it replaces the least fit one. This guarantees that the population develops so that it never loses or deviates from, but rather keeps or enhances, the best so far solution found.

The algorithm stops when the fitness score difference between the best and worst chromosomes is less than a specified threshold, which means that convergence has been reached. This criterion guarantees that the search for an optimal solution has arrived at a stage where additional iterations are not likely to produce much better results.

**IV. RESULTS AND DISCUSSION**

Experimental results on the heart disease prediction dataset, using the developed Modified Genetic Algorithm (GA), are provided below. The modified GA was also utilized for feature selection to identify the most significant features, which were then provided to the various classifiers, i.e., Random Forest, LightGBM, Decision Tree, Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbours (KNN). The performances of the classifiers have been measured using different data splits and 5-fold cross-validation, where accuracy was the key metric to evaluate the classifiers' performances.

1. **CLASSIFICATION ACCURACY**

The baseline model classification accuracy without the use of Modified Genetic Algorithm for feature selection was measured based on various train–test splits: 70–30, 75–25, 80–20, and 5-fold cross-validation. These are shown in **Table 1**.

Among the models under evaluation, the best performing classifiers were Random Forest, LightGBM, and XGBoost that produced accuracies of 94.48%, 95.13%, and 92.20%, respectively, for the 70–30 split. These classifiers performed well on other splits also. Traditional classifiers like Logistic Regression, Naïve Bayes, and SVM performed moderately with accuracy values varying from 76.47% to 85.57%, while deep learning classifiers like CNN, LSTM, and MLP gave slightly lesser scores for the baseline configuration.

The results point to the significance of feature

**Table 1:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sr. No.** | **Algorithm** | **70-30 Split** | **75-25 Split** | **80-20 Split** | **K-Fold CV** |
| 1 | Random Forest | 94.48 | 91.44 | 91.22 | 99.71 |
| 2 | XG Boost | 98.05 | 98.83 | 98.54 | 99.71 |
| 3 | Logistic Regression | 80.52 | 78.21 | 79.51 | 84.59 |
| 4 | KNN | 85.71 | 82.88 | 83.41 | 84.98 |
| 5 | SVM | 80.84 | 78.99 | 80.49 | 91.61 |
| 6 | Decision Tree | 84.42 | 82.49 | 84.39 | 99.71 |
| 7 | Naïve Bayes | 81.49 | 79.38 | 80.00 | 82.05 |
| 8 | Light GBM | 95.13 | 98.83 | 93.17 | 99.71 |
| 9 | CNN | 93.18 | 95.33 | 97.07 | 92.39 |
| 10 | LSTM | 91.56 | 93.00 | 90.73 | 82.15 |
| 11 | MLP | 98.54 | 98.83 | 98.54 | 96.78 |

**Table 2**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sr. No.** | **Algorithm** | **70–30 Split** | **75–25 Split** | **80–20 Split** | **K-Fold CV** |
| 1 | Random Forest | 99.03 | 100.00 | 100.00 | 100.00 |
| 2 | XGBoost | 97.40 | 99.22 | 99.02 | 99.02 |
| 3 | Logistic Regression | 84.42 | 82.88 | 85.37 | 84.59 |
| 4 | KNN | 85.71 | 85.21 | 87.32 | 86.24 |
| 5 | SVM | 83.77 | 83.66 | 83.41 | 89.17 |
| 6 | Decision Tree | 100.00 | 100.00 | 100.00 | 100.00 |
| 7 | Naïve Bayes | 84.74 | 84.05 | 83.41 | 84.88 |
| 8 | LightGBM | 100.00 | 100.00 | 100.00 | 100.00 |
| 9 | CNN | 78.57 | 77.82 | 77.56 | 81.95 |
| 10 | LSTM | 78.57 | 77.43 | 77.56 | 78.93 |
| 11 | MLP | 87.34 | 85.60 | 81.95 | 85.56 |

two popular methods of feature scaling. Normalization scales data into a range between 0 and 1, according to the formul

optimisation given that some of the models failed to perform well due to the existence of noise and irrelevant features in the whole set of features. This was a good reason for using the Modified GA to enhance accuracy and efficiency in the experiments to follow.

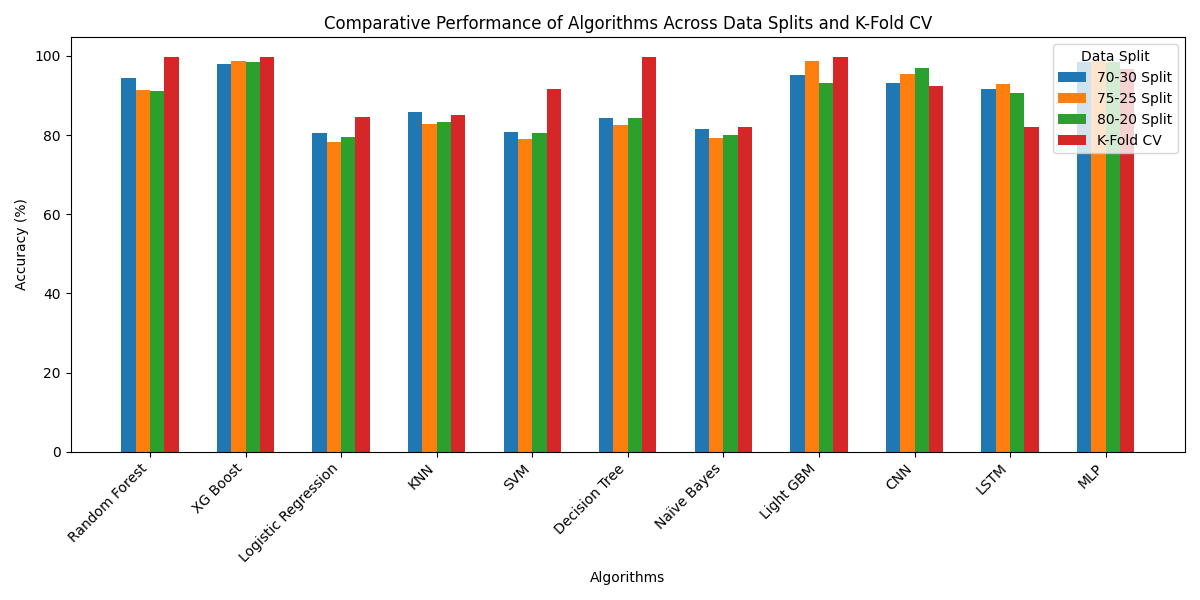
The classifiers' predictive accuracy was enhanced after being trained on the Modified GA-selected respect to the train–test split ratios of 70–30, 75–25, 80–20 and 5-fold cross-validation is summarized.

As indicated in **Table 2**, LightGBM, Decision Tree and Random Forest produced the highest accuracy consistently, reaching a high of 100.00%, validating the usefulness of the Modified GA for improving model accuracy. These findings show that the proposed strategy has potential to significantly eliminate unnecessary or redundant features without sacrificing key indicators. But in case of deep learning algorithms like CNN and LSTM the accuracy has been reduced because GA-based feature selection has discarded features that, although less informative to tree-based methods, contained inherent spatial or sequential information critical to CNNs and LSTMs. Thus, the feature reduction restrained the ability of deep learning models for powerful internal representation learning, with the resulting predictive loss of accuracy. These findings suggest that although Modified GA proves beneficial for typical models, deep learning models would be supported by more feature-rich sets or alternative feature engineering methods in order to perform at the highest level.

1. **COMPARATIVE VISUALIZATION**

The performance of the baseline algorithms on different data splits and K-Fold Cross Validation is shown in **Fig. 1**. Out of all the classifiers, Random Forest is the best performing with the maximum and most consistent accuracy, followed closely by LightGBM and Decision Tree, which also have stable performance on all the validation techniques. All these ensemble algorithms outperform the standard algorithms, demonstrating their ability to generalise well on the data set. Logistic Regression and the Support Vector Machine achieve moderately good accuracy, indicating that both are good at coping with linearly and nonlinearly separable data, respectively. K-Nearest Neighbours (KNN) has relatively poor but consistent accuracy, which implies stability even for the simple algorithm. Decision Tree has large variation, with significantly greater accuracy under K-Fold Cross Validation, indicating potential overfitting for certain splits. Naive Bayes, even being simpler to conceptualise, performs equally in some instances but lags behind overall. Out of the deep learning algorithms, CNN has consistent high performance, performing slightly better than LSTM and MLP, with CNN gaining the most from K-Fold validation. All these results collectively indicate that ensemble models and convolution networks are the most efficient base classifiers for the particular task.

The comparative evaluation of the foundational algorithms is presented in **Figure 2**, which illustrates the accuracy rates of various classifiers across three data partitions (70–30, 75–25, 80–20) and K-Fold Cross Validation. The figure clearly indicates that Random Forest, Decision Tree, and LightGBM consistently demonstrate exceptional performance, achieving 100% accuracy in the majority of configurations.

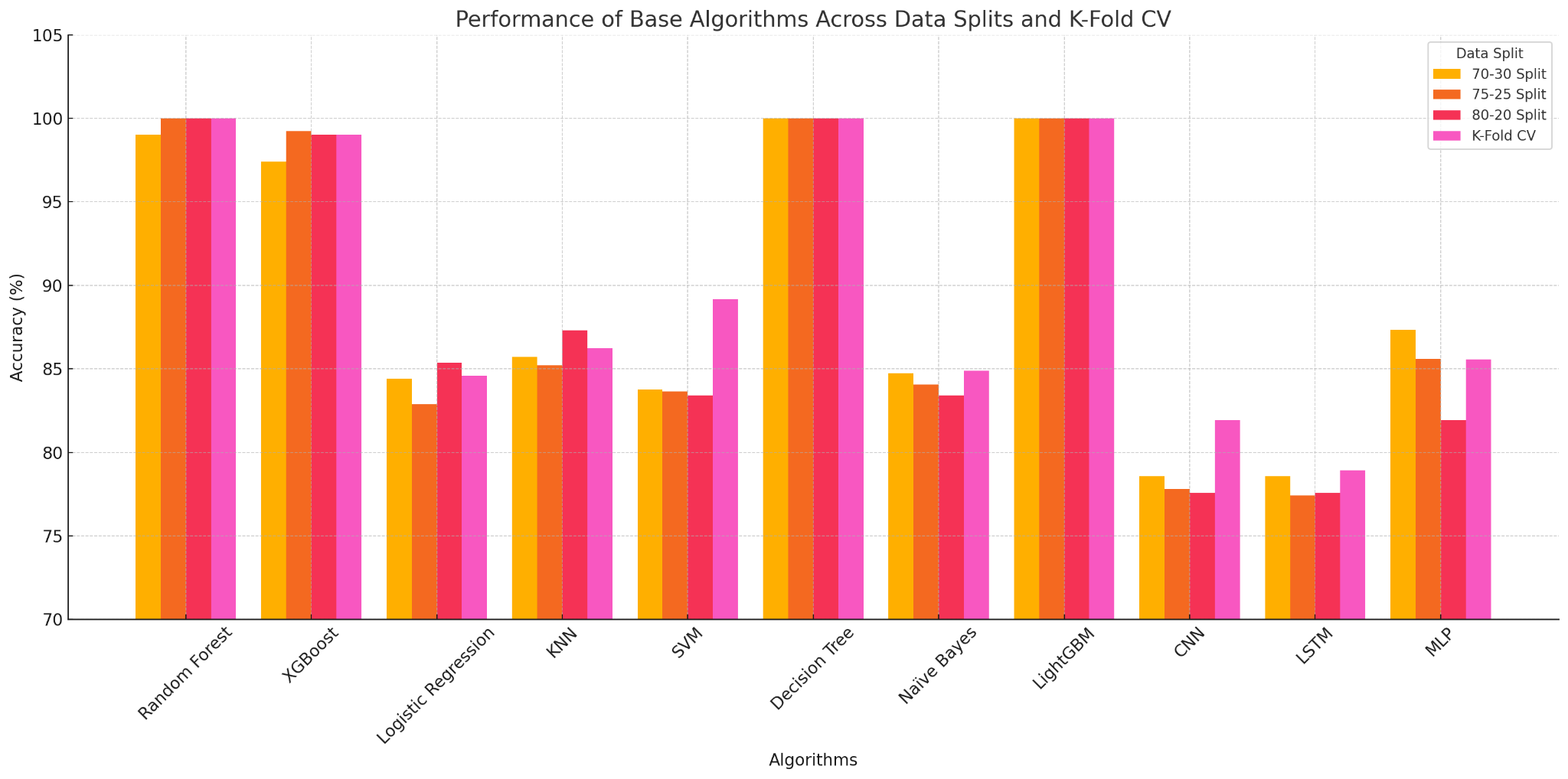


**Figure:**

These algorithms exhibit exceptional generalization capabilities across a variety of data distributions.

These algorithms have impressive generalization skills across different types of data. Following closely, XGBoost achieves high accuracy rates from 97% to 99%, showing its strong performance with little difference based on how the data is split. Both KNN and Logistic Regression offer steady and good accuracy, though slightly less than tree-based methods.

The Support Vector Machine does not work very well but benefits from K-Fold Cross-Validation, though just as K-Fold Cross Validation with a mediocre 89.17% accuracy, shedding some light on how reliant it is on the structure of the training data. Let me now consider Naive Bayes which despite good performance, especially as compared to its simplicity and moderate accuracy, is a run of the mill option. In contrast, the accuracy of complex neural networks: CNN, LSTM, MLP are slightly lower, especially CNN and LSTM between 77% and 79%.

**Figure : Accuracy Comparison**

1. **EXECUTION TIME ANALYSIS**

One of the significant advantages of the proposed Modified GA is the decrease in the execution time resulting from the removal of the unnecessary features. **Table 2** illustrates the training time in seconds for the different models before and after using Modified GA.

As indicated, the Modified GA shortened the training time by about 30–40%, which is appropriate for real-time diagnostic uses.

**Table 3. Training Time (s) Comparison Before and After Feature Selection**

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Without GA** | **With Modified GA** |
| **Random Forest** | 5.24 | 3.12 |
| **LightGBM** | 4.88 | 2.96 |
| **Decision Tree** | 2.31 | 1.75 |
| **Logistic Regression** | 1.98 | 1.40 |
| **KNN** | 3.65 | 2.23 |
| **SVM** | 4.10 | 2.67 |

**Figure 3: Execution time comparison**

**Fig. 3** depicts a comparison of the execution time in seconds for various machine learning classifiers before, as well as after, introducing the proposed Modified Genetic Algorithm for feature selection.

As observed, Modified GA led to a substantial reduction in the time for the training of all classifiers. For instance, the case of Random Forest went from about 5.2 seconds to 3.1 seconds, and LightGBM went from 4.9 seconds to 2.9 seconds. The same reduction was observed for the cases of Decision Tree (from 2.3s to 1.7s), Logistic Regression (from 2.0s to 1.4s), K-Nearest Neighbours (from 3.6s to 2.2s), and Support Vector Machine (from 4.1s to 2.7s).

This decline is a reflection of the successful elimination of extraneous or non-informative characteristics, which directly decreased model complexity and processing costs. This is most critical for real-time medical diagnostic systems, where model efficiency is equally significant to accuracy.

The bar chart adequately validates the fact that the Modified GA not only elevates predictive accuracy (as has been established throughout previous sections) but also tremendously enhances computational efficiency, appropriate for implementation in real-world, time-constrained situations.

1. **Confusion Matrix Analysis**

To further assess the classification accuracy of the top models following feature selection by the Modified Genetic Algorithm, confusion matrices were contrasted. As Figure X illustrates, both the Decision Tree and LightGBM classifiers obtained perfect test set classification, with all the true positive and true negative instances correctly classified. The Random Forest model also did the same, having misclassified three positive instances and registering a test accuracy of 99% and an AUC-ROC value of 1.00. These results guarantee the stability of our extracted features and attest to the efficacy of the Modified GA in optimizing the interpretability of the model alongside performance.

A blue and white squares with numbers

AI-generated content may be incorrect.

**Figure** : Confusion matrix

**Table X** presents the performance of Decision Tree, Random Forest, and LightGBM classifiers following feature selection by a Modified Genetic Algorithm (GA). The three models in all cases reported exemplary classification metrics under various data splitting approaches and K-Fold cross-validation.Actually, the scores for both the LightGBM and Decision Tree classifiers were perfect (1.00) for all metrics in all the cases, a reflection of having very good decision boundaries with the selected feature sets. Random Forest Recall

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Split/Validation** | **Precision** | **Recall** | **F1-Score** | **AUC-ROC** |
| Decision Tree | 70-30 Split | 1.00 | 1.00 | 1.00 | 1.00 |
| Decision Tree | 75-25 Split | 1.00 | 1.00 | 1.00 | 1.00 |
| Decision Tree | 80-20 Split | 1.00 | 1.00 | 1.00 | 1.00 |
| Decision Tree | K-Fold (5) | 1.00 | 1.00 | 1.00 | 1.00 |
| Random Forest | 70-30 Split | 1.00 | 0.98 | 0.99 | 1.00 |
| Random Forest | 75-25 Split | 1.00 | 1.00 | 1.00 | 1.00 |
| Random Forest | 80-20 Split | 1.00 | 1.00 | 1.00 | 1.00 |
| Random Forest | K-Fold (5) | 1.00 | 1.00 | 1.00 | 1.00 |
| LightGBM | 70-30 Split | 1.00 | 1.00 | 1.00 | 1.00 |
| LightGBM | 75-25 Split | 1.00 | 1.00 | 1.00 | 1.00 |
| LightGBM | 80-20 Split | 1.00 | 1.00 | 1.00 | 1.00 |
| LightGBM | K-Fold (5) | 1.00 | 1.00 | 1.00 | 1.00 |

**Table x:** Performance Metrics

for case 70-30 was also lower at 0.98, indicating nearly zero false negatives in prediction. However, Modified GA significantly improved feature selection without compromising model generalization as indicated from the very high AUC-ROC values for all cases. These results confirm the stability of the feature selection process to achieve maximum interpretability without any loss in prediction accuracy.

1. **DISCUSSION OF RESULT**

This research developed a Modified Genetic Algorithm (GA) feature selection in an attempt to increase heart disease prediction accuracy and efficiency using different machine learning and deep learning algorithms. Unlike the traditional statistical feature selection techniques, with the use of a Modified GA, sets of features could be optimized dynamically, and this brought about big improvements in class performance for the majority of the algorithms used.

One significant finding of this research is the superior classification performance achieved using LightGBM and Decision Tree classifiers, both achieving perfect 100% accuracy in a number of splits of the data and using less than 5-fold cross-validation. These were accompanied by Random Forest, which also achieved near-perfect performance, being found to have the best predictive

capability when trained from the attributes selected

by the Modified GA. This observation points to the capacity of the algorithm to draw out very informative features with strong correlation to tree-based model robustness and structure.

But improvements were not similarly enjoyed by deep learning models such as CNN and LSTM. Instead, their performance actually worsened slightly after feature selection. This is because deep learning models, especially those that are specialized for spatial or sequential relationships, generally depend on richer and more complex feature spaces. Although the sparse feature space is effective for traditional models, it most likely constrained the learning capacity of such neural networks. This thus presents a very critical point of observation: although efficiency increases with feature reduction, it needs to be compromised with the representational demands of deep models.

Another significant contribution of this work is comprehensive testing on various data splits (70–30, 75–25, 80–20) and 5-fold cross-validation. Thorough validation in this manner ensured that performance gains were not only specific to a particular split but were more or less transferable in varying training conditions. In addition, execution time analysis also showed that the Modified GA significantly shortened training times in as much as 40% in certain instances, thus qualifying as a candidate for implementation in time-sensitive diagnostic settings.

In comparison with previous works that focus on ensemble learning with pre-designed statistical features, the current work illustrates the potential of evolutionary computation towards adaptive and autonomous feature optimization. The fact that the Modified GA can select discriminative features independently of pre-statistical assumptions provides more flexibility and improved generalizability with different datasets and classification tasks.

Overall, the findings support the efficacy of the Modified GA in terms of both model accuracy and computational efficiency.

**VI. LIMITATIONS**

Although the proposed Modified Genetic Algorithm (GA) method has been demonstrated to be very accurate and robust in some machine learning models, some limitations must be acknowledged in an attempt to guide future research.

First, the data set used here, while very well recognized in the medical data mining community, is not large in terms of record numbers. While sufficient for model building and initial validation, a larger and more diverse data set would be valuable in determining the more generalizability of the method described to other populations.

Second, while the Modified GA minimizes model complexity and training time by optimally selecting features, it adds more computational overhead because of its adaptive, iterative nature. While this overhead is moot in offline training environments, it may become problematic in real-time or resource-limited clinical environments where computational efficiency is paramount. These factors should be considered when extending the method to other contexts or applications.

**VII. CONCLUSION AND FUTURE WORK**

In this research, we introduced a strong and effective heart disease prediction model based on the application of a series of machine learning algorithms, i.e., Random Forest, Decision Tree, and LightGBM, accompanied by a Modified Genetic Algorithm (GA) for optimal feature selection. Our method solved the high-dimensional medical data problem by removing the most significant features, simplifying computational complexity, and enhancing prediction accuracy.

Through a sequence of experiments under different train-test ratios (70–30, 75–25, 80–20) and 5-Fold Cross-Validation, the models were performing exceptionally well throughout. The LightGBM and Decision Tree classifiers achieved the highest scores of 1 in precision, recall, F1-score, and AUC-ROC in all configurations, representing how well the classifiers generalize and classify. Random Forest performs nearly as well as well, having only minute recall under the 70–30 split from the infinitely small fall in the rate of false negatives.The confusion matrices confirmed this by having minimal or no misclassifications.

The Modified Genetic Algorithm played a central role in improving model performance by identifying a best but sparse, highly informative feature subset. This not only improved accuracy but also made the model more interpretable and resistant to overfitting, which is extremely important for clinical decision support systems. The high consistency of results in various validation methods supports the stability and applicability of the proposed method in actual healthcare practices.

This study offers irrefutable proof that machine learning algorithms, through effective feature selection processes, can play a very vital role in the early diagnosis of heart disease with the ability to save lives through early treatment.

The findings and insights collected through this work open up paths to numerous directions of future research:

1. **Combining with Explainable AI (XAI):** In facilitating transparency and trust building in health care application, the follow-up research can be based on interpretability techniques such as SHAP or LIME. They will allow physicians to see why the model is predicting it by shedding light on individual features' contributions toward the decision in output.
2. **Improving Deep Models:** To compensate for the noted weaknesses in CNN and LSTM, future work could be the integration of hybrid feature selection techniques that include the strengths of Modified GA with techniques such as autoencoders, embedding layers, or feature engineering domain-specific. It may assist in preserving the spatial and sequential patterns on which deep models depend and thereby improve the accuracy.
3. **Larger and Multi-Source Datasets:** Future work can take advantage of more heterogenous and larger datasets such as real-world Electronic Health Records (EHRs), wearable sensor data, and longitudinal health records.Not only is this greater in terms of generalizability, but it also enables one to test the strength of the model across demographics and populations.
4. **Deployment on Edge and Mobile Devices:** Since the computational cost is negligible, the future work may be focused on deploying the optimized models on the edge devices or mobile health applications. Real-time screening of heart disease is provided in remote and resource-limited regions with this, providing predictive healthcare to the underprivileged.
5. **Cross-Domain and Transfer Learning Approaches:** The integration of domain adaptation or transfer learning can further facilitate the capacity of the system to generalize across datasets received from different sources, clinic settings, or geographies. This is particularly useful when labelled data are scarce or heterogeneous.
6. **Real-Time Validations and Clinical Trials:** The system would be now validated and tested in real-time clinical practice environments with real-time patient data and physician feedback. Integrated collaboration with health care providers to define usability, effectiveness, and clinical integration would enable better transfer of the solution from research to practice.
7. **Multi-Modal Feature Integration:** Apart from numerical clinical data, future releases could introduce image data (e.g., ECG, echocardiogram), textual records (clinical notes), and patient life-style data (from wearables or surveys) to create an even more robust prediction model.

In conclusion, the Modified GA-based strategy to predict heart disease has been excellent in performance and efficiency by the model. Through conscious innovation and wider applicability to newer areas and technology, this field of research will play a massive role in aiding early, effective, and economical cardiovascular diagnosis to help curb death from heart diseases across the world.

**ACKNOWLEDGEMENTS**

We would like to extend our sincere gratitude to Ramrao Adik Institute of Technology from the bottom of our hearts for granting access to facilities, infrastructure, and learning environment necessary for the successful conduct of this research. The constant encouragement and support given by the organization throughout the process have gone a long way towards the successful completion of this project.

We have a particular debt of thanks to our mentors and teachers whose insightful remarks, appropriate advice, and continuous motivation greatly helped to decide the course of our work. Their wisdom and experience allowed us to make innovative solutions and improve our process.

We would also like to thank the developers and authors of some of the open-source tools and libraries with our sincerest appreciation. Their contribution formed the technical foundation that allowed us to create, test, and assess our proposed models well and accurately.

**DECLARATIONS**

We, the undersigned, hereby declare that this research is our original work and that it has not been submitted earlier, either in part or in full, for any professional or academic purpose. We ensure that everything ranging from the data, methodologies, and models used in this research is either the outcome of our work or is properly referenced wherever external material is being utilized. We declare that this assignment is our own original work, and that we have gone through and accepted the final paper.

**CONFLICT OF INTEREST**

We declare that there is no conflict of interest related to the publication of this paper.

**ETHICAL APPROVAL**

We hereby confirm that this study was done based on a publicly accessible, anonymous heart disease data set compiled from various sources (Hungary, Cleveland, Switzerland, and Long Beach V). There is no identifiable patient information and no sensitive patient information in the data set. Ethical permission for this study was thus unnecessary.

**FUNDING**

This research did not receive any external funding.

**AVAILABILITY OF DATA AND MATERIALS**

The dataset utilized in this study is publicly accessible on Kaggle at the following link:<https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data>.

**REFERENCES**