BERKELEY PUBLICATION
AND RESEARCH INTERNATIONAL
03.28.2022
Journal of Health, Metabolism & Nutrition Studies Vol. 19 No. 3

IMPROVED HEART DISEASE CLASSIFICATION USING SVM-KNN VOTING CLASSIFIER AND GridSearchCV OPTIMIZATION

CHANDRA SHIKHI KODETE
Department of School of Technology, Eastern Illinois University, Charleston, Illinois, USA.

ABSTRACT

This work aims to present a reliable and robust framework for the classification of heart disease based on machine learning. Methodologically, the approach is anchored on three steps: pre-processing the data, which includes outlier handling using the Z-score method, such that a reduced dataset with 396 records was obtained; this number was reduced to 389 records for improvement in data consistency and compactness. Standardization is achieved with MaxAbsScaler in such a way that it scales all the features into the range [0,1] and maintains sparsity while ensuring compatibility with machine learning models. A hybrid feature selection technique utilizes Lasso regression combined with a genetic algorithm so that this subset of features is optimized through regularization and evolutionary search.

Introduction:

Heart diseases all together fall under the conditions that cause defects in heart's shape and functionality. Such conditions include coronary artery disease, heart attacks, heart failure, arrhythmias, and valvular heart diseases [1]. The most common form is that in which plaques have narrowed or occluded the coronary arteries and diminished the heart muscle blood supply and oxygen distribution. This is possibly triggered by pain in the chest, heart attack, and other potentially fatal complications. These factors are some of the

ISSN: 1213-4406

Five most important features were selected: cp, oldpeak, and chol-and thus a fixed low dimension with improved performance were warranted. In the model building stage, an ensemble framework of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifiers combined with the Voting Classifier and with GridSearchCV optimization is used. It has produced excellent balancing of the strengths of two classifiers, while the predictive performance can be delivered commendably. The evaluation metrics show accuracy at 94.87%, precision at 88.89%, recall at 88.89%, and an F1 score of 88.89%. A high capability for discrimination was there in this model through the ROC AUC score of 0.9926. These results stress the efficiency of proposed methodology in achieving accuracy and reliability towards heart diseases classification.

Keywords: *Heart Disease Classification, SVM (Support Vector Machine), KNN (K-Nearest Neighbors), Genetic Algorithm, Lasso Regression, ROC AUC Score*

Causes. Another major condition is heart failure, where the heart would pump less blood supply than required by the tissues in the body. This would lead to fatigue, fluid build-up, and shortness of breath [2]. There are several risk factors for heart disease. These include high blood pressure, high cholesterol, smoking, diabetes, physical inactivity, and obesity. Other causes include excessive alcohol consumption and a personal and/or family history of heart disease [3]. The most significant determinants of risk – a combination of lifestyle, diet, and exercise – can be used to prevent and control heart disease.

Other illnesses, such as hypertension or diabetes, can also be treated with a variety of medications in combination with changes in lifestyle that prevent heart disease. There are many reasons aside from physical well-being that mental well-being could harm the health of the heart too. For example, chronic stress, depression, and anxiety have been linked to the risks of heart disease. Therefore, mental well-being should be added to the prevention of heart disease [4]. Thus, heart diseases have always been the

major causes of death across the globe. Interestingly, with the advancement in medical research, diagnostics, and treatments, results have picked up. Detection through diagnostic tools like ECG, echocardiograms, and blood tests, and changes in lifestyle, drug therapy, and intervention by surgery have all led to better management and treatment of heart diseases [5]. Still, with heart disease remaining a significant public health issue, much more needs to be done in terms of research and awareness. This research is an attempt in that direction.

Research Gap

Despite some recent developments in the use of machine learning for the prediction of heart diseases, there is a plethora of research gaps. Most of the previous studies engage with particular data sets containing smaller data sets, such as UCI heart condition and Cleveland Clinic Heart Disease data sets that may not cover the required diversity but wide and generalization needed for a more generalized or non-clinical population database. Additionally, little work is pursued in employing feature selection techniques along with real-time data to have a better accuracy for the prediction in clinical applications. The present study attempts to bridge the gap, dwelling more on how the significance of each method, feature selection and classification can foster and optimize the interactive performance of each other in ensemble learning.

Moreover, spending much more attention on the traditional approaches, such as logistic regression and decision trees, ignores all the advantages of new-generation deep learning models. Finally, it is worth noting that although the influence of ensemble methods is studied in terms of performing parameters for boosting and bagging, these works do not take into account specificity and area under the curve (AUC). Besides, measures, which are highly important in the context of particular cases and those of this healthcare field, are not considered. Filling these gaps will be important for establishing better heart disease prediction models, one that will effectively detect the disease early enough, leading to proactive diagnoses and treatments.

Research Questions

RQ.1 How will diverse datasets enter into creating machine learning models for heart disease and their impact on *generalisability* and predicting accuracy?

RQ.2 How do multiple feature selection techniques, used in combination with ensemble learning algorithms, improve the performance of models for the prediction of heart diseases?

RQ.3 What are the differences and similarities between new deep learning approaches and traditional algorithms in improving accuracy of heart disease prediction?

RQ.4 What other than accuracy metrics do heart disease prediction models in clinical settings achieve?

Literature Review

Some related studies on the thematic concerns of this study abound in the literature. Accordingly, Thuraka [6] demonstrates that machine algorithms, such as Decision Tree, Logistic Regression, Logistic Regression SVM, and Random Forest, are feature selection techniques for improving medical performances. The author used data science and machine learning to perform various techniques. The study emphasizes the need to apply these techniques to real-time datasets; adding that ensemble methods should be probed further for a greater increase in accuracy, as avenues for future research. This effort complements other projects under way related to medical data mining and cardiac disease prediction.

Robison Spencer et al. [7] present how the ability of feature selection and classification methods affects their capability to predict heart disease. By using PCA, Chi-squared testing, ReliefF, and symmetrical uncertainty for unique feature sets evaluation of four datasets of heart disease, the authors find that the effectiveness of feature selection varies due to the different machine learning method applied to the said datasets. Chi-squared feature selection with BayesNet algorithm yielded 85.00% accuracy. Moreover, the highest recall at 87.22% was observed when PCA is used in combination with the IBK method. This, therefore, underlines the need to try a large

number of combinations of feature selection techniques with machine learning algorithms in order to enhance the accuracy of prediction. General results indicate how pretty well the perfectly tuned algorithms of machine learning could help doctors for an early detection of cardiac disorders.

In 2016, Saqlain et al. [8] described the concept of using unstructured text data from the medical reports of cardiac patients with HF using machine learning. Studies conducted targeted the detection time and risk assessment that would provide ways of preventing the increasing incidence and mortality of HF. The data included both nominal and ordinal data along with categorical data with different comorbidities, demographic data, medications, and Laboratory findings originating from the Armed Forces Institute of Cardiology Pakistan. The authors also chose Naive Bayes (NB), logistic regression, neural networks, support vector machines (SVM), random forests and decision trees as classification models. According to their findings, the Naïve Bayes algorithm was superior to the other models with an accuracy level of 86.7% and a value of the area under the receiver operating characteristic curve (AUC) of 92.4%. Hence, NB proved sturdy with unstructured data issues that the given dataset contains, especially on multiple classification data and categorical features. It focused on the following risk factors: age, hypertension and low-level ventricular ejection fraction- LVEF that profoundly influence HF survival. The authors proposed recommendations for the further improvement of the model performance with the help of more sophisticated approaches such as CPXR and text mining for extracting data from unstructured sources.

In 2018, Amin Ul Haq et al. [9] designed a hybrid intelligent system model to predict heart diseases using machine learning algorithms. The data set used in the study was chosen to be the Cleveland heart disease data set and seven classifiers were used in the study which were logistic regression, K-nearest neighbour (K-NN), artificial neural network (ANN), support vector machines (SVM), Naive Bayes (NB), decision tree (DT), and random forest. The authors incorporated three feature selection algorithms: To improve the classification accuracy Advanced Gene Selection methods such as Relief, mRMR and LASSO methods are used. Using 10-fold cross-validation,

the system presented a high accuracy of classification for normal and abnormal breast densities: logistic regression performed best with an accuracy of 89%, and specificity of 98% based on six features selected by the Relief algorithm. Another model that was close behind with 87% accuracy was the support vector machine model with the radial basis function (RBF) kernel. The study stressed that feature selection algorithms are proficient in enhancing classifiers' accuracy, sensitivity and specificity and also reducing computation time. The most influential characteristics included Thallium Scan, type of chest pain and exercise-induced angina. It is clear that the proposed system offers enhanced diagnosis methods compared to traditional techniques and offers a very efficient and non-invasive decision for clinicians, for diagnosing heart disease.

In 2019, Nourmohammadi-Khiarak et al. [10] proposed a new approach on the basis of the imperialist competitive algorithm for heart disease diagnosis, with metaheuristic selectivity of the feature space. The heart disease dataset focus raises significant concerns in feature selection, sample distribution and difference in the magnitude. For feature selection important for the classification task, the proposed method in this work tries to enhance the trade-off between accuracy in classification and the number of features needed. The method applied in this study was the K-nearest neighbour algorithm that was used as a classifier. In the comparison, conventional classifiers were the Naïve Bayes, decision trees, neural networks, and support vector machines, SVM. The hybrid method presented results superior to others since the model had a classification accuracy of 88.25% a sensitivity of 94.2% and a specificity of 83.4%. This performance was greater than Naive Bayes, neural, and even SVM; all the approaches mentioned above were fine-tuned to achieve 85.7%. Therefore, this proposed method of classifying provides a more efficient classification with fewer features and thus can be used highest an order app in clinical practice. This paper concludes bringing to light the fact that this proposed method makes the diagnosis of heart diseases better and suggests future research can be based on the expansion of the idea and

experiments when data is missing and also study other potential uses of the medical meta-heuristic optimizer.

Proposed Methodology

In this section, the proposed methodology is described. It includes data pre-processing, feature selection, and model development-all of which utilize advanced machine learning algorithms [11]-the process for heart disease classification has a structured methodology. This is the data pre-processing stage as indicated in Fig. 1. Features use MaxAbs scaling normalisation that normalises the feature values within the training consistent dataset. The Hybrid approach GALS is applied for feature selection is hybridization of Lasso regression [12] with a genetic algorithm [13] to elicit important attributes and lessen noise, thus obtaining accuracy improvement in the model. To perform the modelling, a voting classifier framework comprises SVM [14] and KNN [15]. The hyper-optimisation approach is made up of GridSearchCV to obtain the most effective model. This evaluation for the built model is then done using great performance metrics like accuracy, precision, recall, F1 score, and ROC AUC, with confusion represented in the confusion matrix for visualization. Feature-selected machine learning models [16], together with robust evaluation techniques, assure the heart disease classification model to have high prediction accuracy with economic performance in real-world applications.

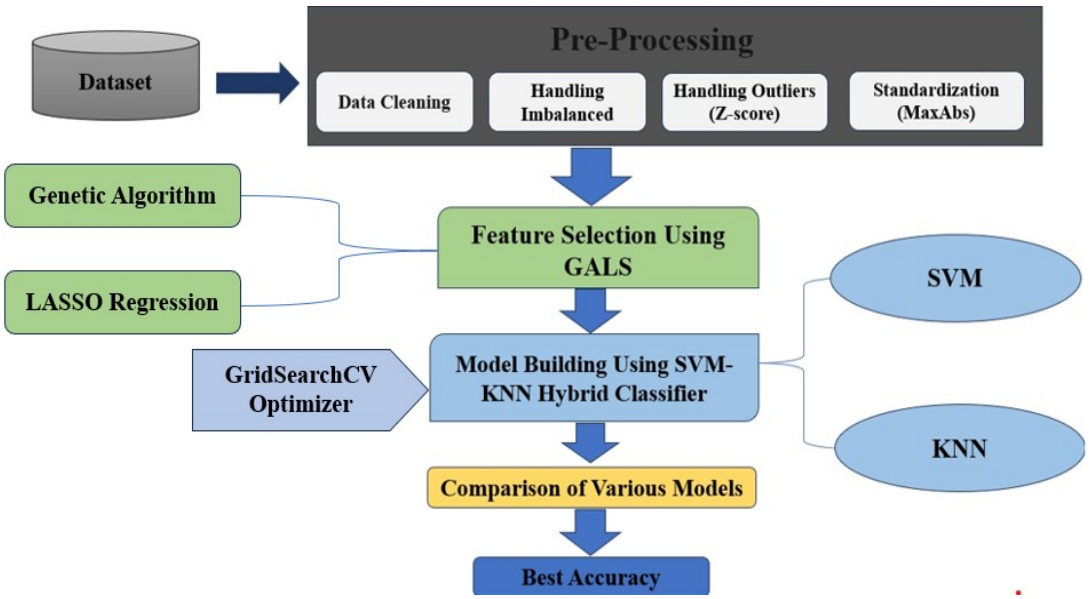

Fig. 1: Workflow Diagram

Data Collection

It was downloaded from the Kaggle website, one of the most famous platforms for machine learning and data science projects. When it comes to the problem chosen on the selected dataset in [Insert Dataset Description, e.g., brain tumour detection, fake news classification, etc.], it contains data about the following attributes: [list key features, e.g., age, gender, health indicators, etc.], for which the necessity to prepare the classification models [17] is also necessary. The dataset is having [X number of records] and [Y number of features]. The target variable in this data set is the class which needs to be predicted by the model - "Label", "Class". The data set was pre-processed in such a way that missing values, outliers, and scaling were done on it; due to this process overall performance of the model was improved. With this data in mind, we came up with a powerful machine learning model that would be predictive enough for the target variable. For that, we constructed a training set and also a test set in order to test the generalisation from this data. To obtain such data, the collection from Kaggle came out as a good and

comprehensive source that can be used for testing, hence calibrating the performance of our chosen algorithms.

Pre-Processing

Data Cleaning

The data cleaning was divided into two components: preparation of the dataset and cleaning of the dataset prior to analysis. For this task, I selected a dataset from Kaggle consisting of 175,341 records and 45 features. After checking for missing values, there were no missing data values. Hence, we did not have to perform imputation or row removal. Then to validate the features, we made sure that numeric attributes like age, cholesterol level, and medical indicators were all within expected ranges.

To check the consistency of categorical features like sex, cp, and slope, columns like proto and attack_cat were checked as being categorical and thus retained their original object type so that they could be easily encoded in the feature processing phase. This was to ensure that they would be compatible with the machine learning algorithms. In doing so, the checks are conducted to ensure that the dataset is clean and available for further analysis as shown in Fig. 2. This thereby ensured that the subsequent modelling would be based on reliable and correct data. See Fig. 2 below:

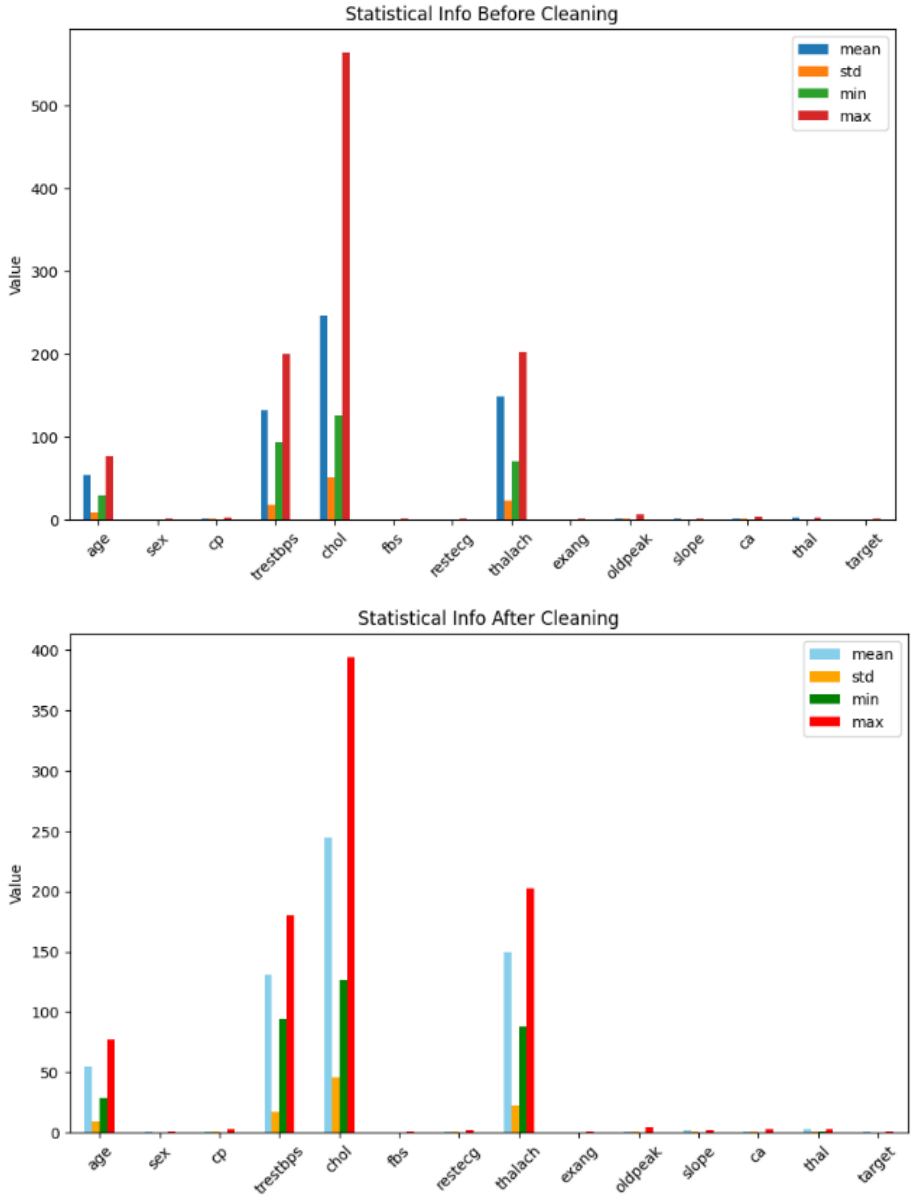

Fig. 2: Bar plot of Statistical Info before and after Cleaning

Handling Imbalanced Data with Random Oversampling

To mitigate class imbalance, Random Oversampling [18] was utilized. It generates duplication of minority class samples; through this, random oversampling appropriately balances the class distribution to ensure no class outshines the other for equally effective learning from the model by

both classes. At first, class distribution was investigated and visualised, using a bar plot to illustrate the frequency of the classes. After applying Random Oversampling, the class distribution was recalculated and visualised, as shown in Fig. 3.

It can be seen that the classes are now balanced by this process. The Random Oversampling procedure has increased the size of the samples of the minority class without producing synthetic data to provide a better model for handling imbalanced datasets. Finally, the re-sampled dataset is saved for later use in training and testing the model. The following captures the foregoing better:

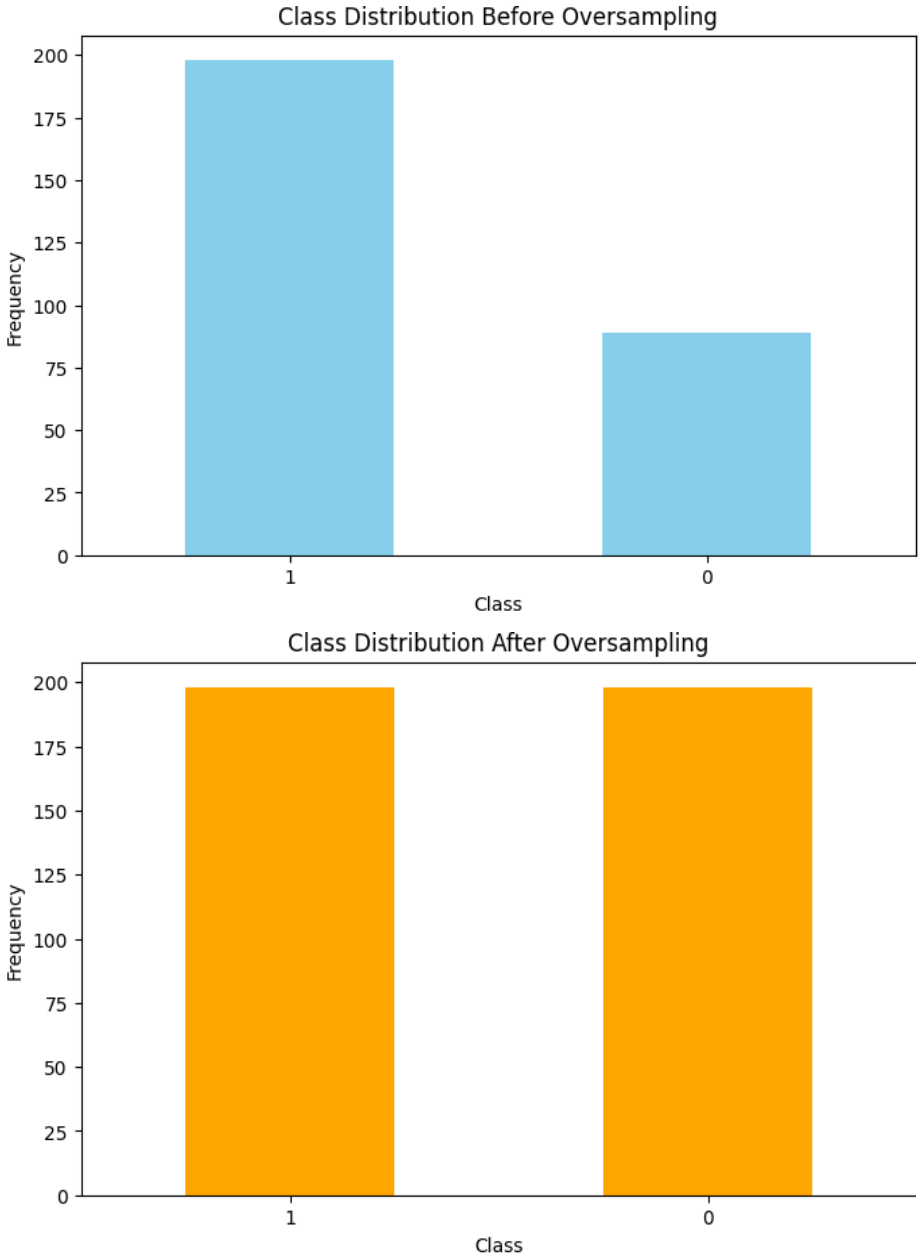

Fig. 3: Bar Plot of before and after Applying Oversampling

Handling Outliers Using Z-score

The data point outliers were addressed using the Z-score method [19], which calculates the standard score of every data point. The data points that have a Z-score above 3 were verified for the presence of outliers. The

technique was applied only on the numeric columns so that the possible inaccuracies would not be ventured into. Statistical information about the dataset was checked and then a box plot is designed for visual presence of outliers. For each number column, count of the outliers was calculated. To avoid inconsistent data, records whose Z-score values were more than the threshold, the values were rejected. Then, after being rejected, the statistical summary for the dataset was updated such that no significant outliers existed in the new box plot while a cleansing had the effect of reducing the distortions in the set due to extreme values as shown in Fig. 4. The cleaned dataset free from outliers was saved for use in subsequent operations. It ensured data integrity, robust model performance, and more reliable prediction in later analyses.

ISSN: 1213-4406

Statistical Info Before Outlier Removal:

	age	cp	trestbps	chol	fbs	restecg
count	396.000000	396.000000	396.000000	396.000000	396.000000	396.000000
mean	55.027778	0.984848	130.414141	247.517677	0.128788	0.540404
std	9.212910	1.021180	17.342133	48.560412	0.335389	0.523749
min	29.000000	0.000000	94.000000	126.000000	0.000000	0.000000
25%	48.000000	0.000000	120.000000	212.000000	0.000000	0.000000
50%	56.000000	1.000000	130.000000	244.000000	0.000000	1.000000
75%	62.000000	2.000000	140.000000	276.250000	0.000000	1.000000
max	77.000000	3.000000	180.000000	394.000000	1.000000	2.000000

	thalach	exang	oldpeak	slope	ca	thal
count	396.000000	396.000000	396.000000	396.000000	396.000000	396.000000
mean	149.967172	0.290404	0.915909	1.431818	0.598485	2.255051
std	22.096141	0.454523	1.012216	0.584957	0.896066	0.530679
min	88.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	136.000000	0.000000	0.000000	1.000000	0.000000	2.000000
50%	153.500000	0.000000	0.600000	1.000000	0.000000	2.000000
75%	168.000000	1.000000	1.500000	2.000000	1.000000	3.000000
max	202.000000	1.000000	4.400000	2.000000	3.000000	3.000000

	target	sex
count	396.000000	396.000000
mean	0.613636	0.500000
std	0.487532	0.500633
min	0.000000	0.000000
25%	0.000000	0.000000
50%	1.000000	0.500000
75%	1.000000	1.000000
max	1.000000	1.000000

Statistical Info After Outlier Removal:

	age	cp	trestbps	chol	fbs	restecg
count	389.000000	389.000000	389.000000	389.000000	389.000000	389.000000
mean	54.946015	0.994859	130.210797	246.174807	0.131105	0.542416
std	9.260528	1.017868	17.274819	46.934318	0.337950	0.519094
min	29.000000	0.000000	94.000000	126.000000	0.000000	0.000000
25%	48.000000	0.000000	120.000000	212.000000	0.000000	0.000000
50%	56.000000	1.000000	130.000000	244.000000	0.000000	1.000000
75%	62.000000	2.000000	140.000000	275.000000	0.000000	1.000000
max	77.000000	3.000000	180.000000	360.000000	1.000000	2.000000

	thalach	exang	oldpeak	slope	ca	thal
count	389.000000	389.000000	389.000000	389.000000	389.000000	389.000000
mean	150.038560	0.290488	0.879949	1.442159	0.588689	2.254499
std	22.231353	0.454572	0.964236	0.578722	0.882520	0.527112
min	88.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	136.000000	0.000000	0.000000	1.000000	0.000000	2.000000
50%	154.000000	0.000000	0.600000	1.000000	0.000000	2.000000
75%	168.000000	1.000000	1.500000	2.000000	1.000000	3.000000
max	202.000000	1.000000	3.800000	2.000000	3.000000	3.000000

	target	sex
count	389.000000	389.000000
mean	0.614396	0.498715
std	0.487365	0.500642
min	0.000000	0.000000
25%	0.000000	0.000000
50%	1.000000	0.000000
75%	1.000000	1.000000
max	1.000000	1.000000

Fig. 4: Statistical Info before and after Outlier Removal

ISSN: 1213-4406

Standardization Using MaxAbsScaler:

Standardization was done so that there is a common scale of feature values. The MaxAbsScaler [20] method scales each feature to its maximum absolute value; hence the sparsity and range in the data will be preserved. This method suits best for the datasets containing both positive and negative numbers. This statistical information of features was reviewed and a box plot has been used to characterize the distribution of the data. In this step, MaxAbsScaler standardized the features to the same range. This means that all of the values fall in a range between -1 and 1. The target variable was excluded from the scaling function to prevent data leakage. After standardization, statistical properties of the dataset were observed as shown in Fig. 5.

The box plot confirmed uniform scaling in the features. These improvements above made the feature further more comparable. Moreover, they encourage greater compatibility with machine-learning algorithms. Then, the standardized dataset was saved to use later, ensuring the constant robust inputs that were available for predictive modelling and analysis. See the Fig. 5 below:

```
Automatically detected target column: fbs
Statistical Info Before Standardization:
                age          cp    trestbps        chol     restecg     thalach  \
count   389.000000  389.000000  389.000000  389.000000  389.000000  389.000000
mean     54.946015    0.994859  130.210797  246.174807    0.542416  150.038560
std       9.260528    1.017868   17.274819   46.934318    0.519094   22.231353
min      29.000000    0.000000   94.000000  126.000000    0.000000   88.000000
25%      48.000000    0.000000  120.000000  212.000000    0.000000  136.000000
50%      56.000000    1.000000  130.000000  244.000000    1.000000  154.000000
75%      62.000000    2.000000  140.000000  275.000000    1.000000  168.000000
max      77.000000    3.000000  180.000000  360.000000    2.000000  202.000000

             exang     oldpeak       slope          ca        thal      target  \
count   389.000000  389.000000  389.000000  389.000000  389.000000  389.000000
mean      0.290488    0.879949    1.442159    0.588689    2.254499    0.614396
std       0.454572    0.964236    0.578722    0.882520    0.527112    0.487365
min       0.000000    0.000000    0.000000    0.000000    1.000000    0.000000
25%       0.000000    0.000000    1.000000    0.000000    2.000000    0.000000
50%       0.000000    0.600000    1.000000    0.000000    2.000000    1.000000
75%       1.000000    1.500000    2.000000    1.000000    3.000000    1.000000
max       1.000000    3.800000    2.000000    3.000000    3.000000    1.000000

              sex
count   389.000000
mean      0.498715
std       0.500642
min       0.000000
25%       0.000000
50%       0.000000
75%       1.000000
max       1.000000
    Statistical Info After Standardization:
                 age          cp    trestbps        chol     restecg     thalach  \
    count   389.000000  389.000000  389.000000  389.000000  389.000000  389.000000
    mean      0.713585    0.331620    0.723393    0.683819    0.271208    0.742765
    std       0.120267    0.339289    0.095971    0.130373    0.259547    0.110056
    min       0.376623    0.000000    0.522222    0.350000    0.000000    0.435644
    25%       0.623377    0.000000    0.666667    0.588889    0.000000    0.673267
    50%       0.727273    0.333333    0.722222    0.677778    0.500000    0.762376
    75%       0.805195    0.666667    0.777778    0.763889    0.500000    0.831683
    max       1.000000    1.000000    1.000000    1.000000    1.000000    1.000000

                exang     oldpeak       slope          ca        thal      target  \
    count   389.000000  389.000000  389.000000  389.000000  389.000000  389.000000
    mean      0.290488    0.231565    0.721080    0.196230    0.751500    0.614396
    std       0.454572    0.253746    0.289361    0.294173    0.175704    0.487365
    min       0.000000    0.000000    0.000000    0.000000    0.333333    0.000000
    25%       0.000000    0.000000    0.500000    0.000000    0.666667    0.000000
    50%       0.000000    0.157895    0.500000    0.000000    0.666667    1.000000
    75%       1.000000    0.394737    1.000000    0.333333    1.000000    1.000000
    max       1.000000    1.000000    1.000000    1.000000    1.000000    1.000000

                  sex
    count   389.000000
    mean      0.498715
    std       0.500642
    min       0.000000
    25%       0.000000
    50%       0.000000
    75%       1.000000
    max       1.000000
```

Fig. 5: Statistical Info before and after Standardization

ISSN: 1213-4406

Feature Selection Using GALS

Feature selection using GALS was performed to try to enhance the model's performance through the determination of the most influential attributes. The approach used in the combination of Genetic Algorithms and LASSO regression method [21] was taken into account. This includes population size, mutation probability, and crossover probability tuning, which is used in fine-tuning of feature importance basis on accuracy contribution to the predictive model. It calculated the Root Mean Square Error of the subsets of chosen features. The latter represents how well the chosen subsets minimized the error of prediction.

Based on LASSO regression's characteristic of penalizing less important coefficients in feature selection, the fitness score of chosen features was determined. The entire process resulted in an optimal subset of features, which had been built around its high predictive value attributes. The selected features obtained the highest fitness scores and provided a focused dataset for modelling. A line plot [22] of the selected features represented their fitness scores, assigning them a value on a scale. Refined datasets assist in providing higher accuracy and interpretability of the models developed.

Genetic Algorithm

The GA is an optimisation technique [23] motivated by the rules of natural selection and evolutionary theory. Its operation centres around a population of candidate solutions represented as chromosomes, which evolve iteratively. How to get fit solutions will provide promising solutions, while new ones are generated with genetic operators that include crossover and mutation through chromosome combination or modification. GAs are good explorers of complex and non-linear solution spaces, often without limited constraints normally seen in gradient-based approaches. The stochastic nature introduces diversity and therefore does not get trapped into local optima. Though computationally intensive, the

flexibility of GA makes it useful in feature selection and all other optimization problems about diverse domains.
The equation for this process can be represented as shown in equation (1):

$$X^* = arg \min_{x \epsilon S} f(x) \qquad (1)$$

LASSO Regression

LASSO regression is an important algorithm that combines both regression analysis and feature selection along with regularisation. In fact, it minimizes the loss function, which consists of the least square error and L1-regularisation term in which the penalty term enforces sparsity by driving some coefficients to zero. This would be helpful in interpretation because features that are not relevant will be excluded, and over-fitting will be prevented by constraining the complexity of the model as shown in equation (2). The penalty strength is regulated by a parameter λ optimized by cross-validation techniques. LASSO [24] outperforms very well with high-dimensional data, although it faces problems with correlated features: a few can be selected while setting other values to zero.

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (2)$$

ISSN: 1213-4406

Algorithm: LASSO Regression with Genetic Algorithm Feature Selection (LGA-FS)

1. **Input**: Dataset X (features), target variable y.
2. **Output**: Optimal subset of features X_{LGA-FS}.
3. **Feature Subset Optimization with Genetic Algorithm**:

a. **Initialization**:

- Generate an initial population of binary masks $\mathbf{m}=[\boldsymbol{m_1}, \boldsymbol{m_2}, \ldots, \boldsymbol{m_p}]$, where $m_j \epsilon \{0,1\}$ indicates the inclusion ($m_j = 1$) or exclusion ($m_j = 0$) of feature j.

b. **Fitness Function Evaluation**:

- For each mask **m** in the population:
  i. Apply the binary mask to the dataset **X** to select the corresponding feature subset.
  ii. Perform LASSO regression on the selected subset with **y** as the target.
  iii. Compute the Root Mean Square Error (RMSE) on the validation set as the fitness score.

c. **Genetic Algorithm Operations**:

- **Selection**: Choose parent masks based on fitness scores (lower RMSE preferred).
- **Crossover**: Combine parent masks to generate offspring using a predefined crossover probability.
- **Mutation**: Randomly flip bits in the offspring masks with a predefined mutation probability to introduce diversity.

d. **Iteration**: Repeat the fitness evaluation, selection, crossover, and mutation steps for a predefined number of generations or until convergence.

4. **Best Feature Subset Selection**:
   - Identify the best-performing binary mask $\boldsymbol{m}^*$ with the lowest RMSE.
   - Extract features corresponding to $\boldsymbol{m}^*$ as $\boldsymbol{X_{GA-selected}}$.
5. **Final Refinement with LASSO Regression:**
   a. Apply LASSO regression to $X_{GA-selected}$ to further refine the feature set by identifying features with non-zero coefficients.
   b. Save the final selected features as X_{LGA-FS}.
6. **Return:** X_{LGA-FS} as the optimal subset of features.

ISSN: 1213-4406

Model Building Using SVM-KNN Hybrid Classifier

The model building process starts with loading a dataset, after which the target column is automatically identified as the feature that has the fewest unique values. The dataset is divided into training and test sets, with features (X) and the target variable (y) separated. Then standardization on the features was carried out in order to make all the features be on the same scale, using Standard Scaler for both the training and the test sets. The voting classifier is created by merging two models: SVM and KNN, which performs soft voting for aggregation of predictions.

All training of the model on scaled data and predicts the test sets are executed. Several performance metrics are evaluated for accuracy, precision, recall, F1 score, and ROC AUC score. A confusion matrix is also executed to visualize classification performances of the model. Results are presented and examined to determine the success of the model.

Support Vector Machine (SVM)

Support Vector Machine [25] is a supervised learning algorithm which is primarily used in problems of classification. The algorithm looks for the optimal separation hyper-plane that maximizes the margin between different classes, and best separates different classes of data points. SVMs are work effectively in high dimension space and are aptly used to solve problems involving complex but small to medium-sized datasets. Further, the SVM can be applied to non-linear data using the kernel functions to map the data points into higher dimensions.

An SVM classifier has a decision function, which is provided as shown in the equation (3):

$$\boldsymbol{f(x) = w^T x + bi} \qquad \textbf{(3)}$$

Where:

- w is the weight vector (normal to the hyperplane),
- x is the input feature vector,
- bi is the bias term.

The optimization objective for SVM is to maximize the margin $\frac{2}{||w||}$, while ensuring correct classification of data points. The constrained optimization problem can be written as shown in equation (4):

$$\min_{w,b} \frac{1}{2}||w||^2 \text{subject to } y_i(w^T x_i + bi) \geq 1, \forall i \qquad (4)$$

Where:

- y_i is the class label for the data point x_i.

K-Nearest Neighbors (KNN)

KNN is a non-parametric, lazy learning algorithm which applies to classification and regression problems. For the classification cases, KNN classifies an example based on the majority vote of its 'k' nearest neighbours in the feature space. Euclidean distance is usually used in the calculation of distances between the points; however, other types can also be done. KNN doesn't need any kind of training; it stores the entire dataset and during inference, it makes predictions based on the closest data points [26].

The classification rule for KNN is described as follows as shown in equation (5):

$$\hat{y} = mode(y_1, y_2, \ldots, y_k) \qquad (5)$$

Where:

- $\hat{y}$ is the predicted class label,
- $y_1, y_2, \ldots, y_k$ are the class labels of the 'k' nearest neighbors.

The distance between two data points andx_i, x_j is commonly computed using the Euclidean distance formulaas shown in equation (6):

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{n}(x_{i,m} - x_{j,m})^2} \qquad (6)$$

Where:

- x_i, x_j are the feature vectors of points i and j,
- n is the number of features.

ISSN: 1213-4406

GridSearchCV Optimizer

GridSearchCV is a hyper-parameter tuning technique using, which the best parameter combination is found by exhaustively searching through the specified parameter grid for its combinations with subsequent cross-validation usage for each of this parameter combination [27]. This is also used to search for the better set of hyper-parameters primarily in terms of max accuracy, precision or recall. It is perfect to work on small to medium size parameter spaces, but in large grids, it becomes less computationally inexpensive. Models usually meant for it are Random Forest, SVM, and Gradient Boosting models that provide a systematic way of model optimization.

Results and Discussion

Handling Outliers Using Z-score

A total change in the statistical properties of data was considered with the help of using the Z-score technique for handling outliers. In the dataset, there were originally 396 records. However, the removal of outliers had left 389. The means and standard deviations of features like age, trestbps, and chol adjusted slightly, indicating the removal of extreme values. For example, the mean cholesterol level decreased from 247.52 to 246.17, and its standard deviation reduced from 48.56 to 46.93 - in other words, a more compact data distribution.

Some of the key changeable variables include maximum values: cholesterol reduces from 394 to 360 and the oldpeak feature reduces from 4.4 to 3.8as shown in Fig. 6. These changes support better constancy and reduced variance in the data. Statistical information improves the quality of the dataset for possible use in subsequent machine learning tasks and increases model reliability by combating extreme outlier values.

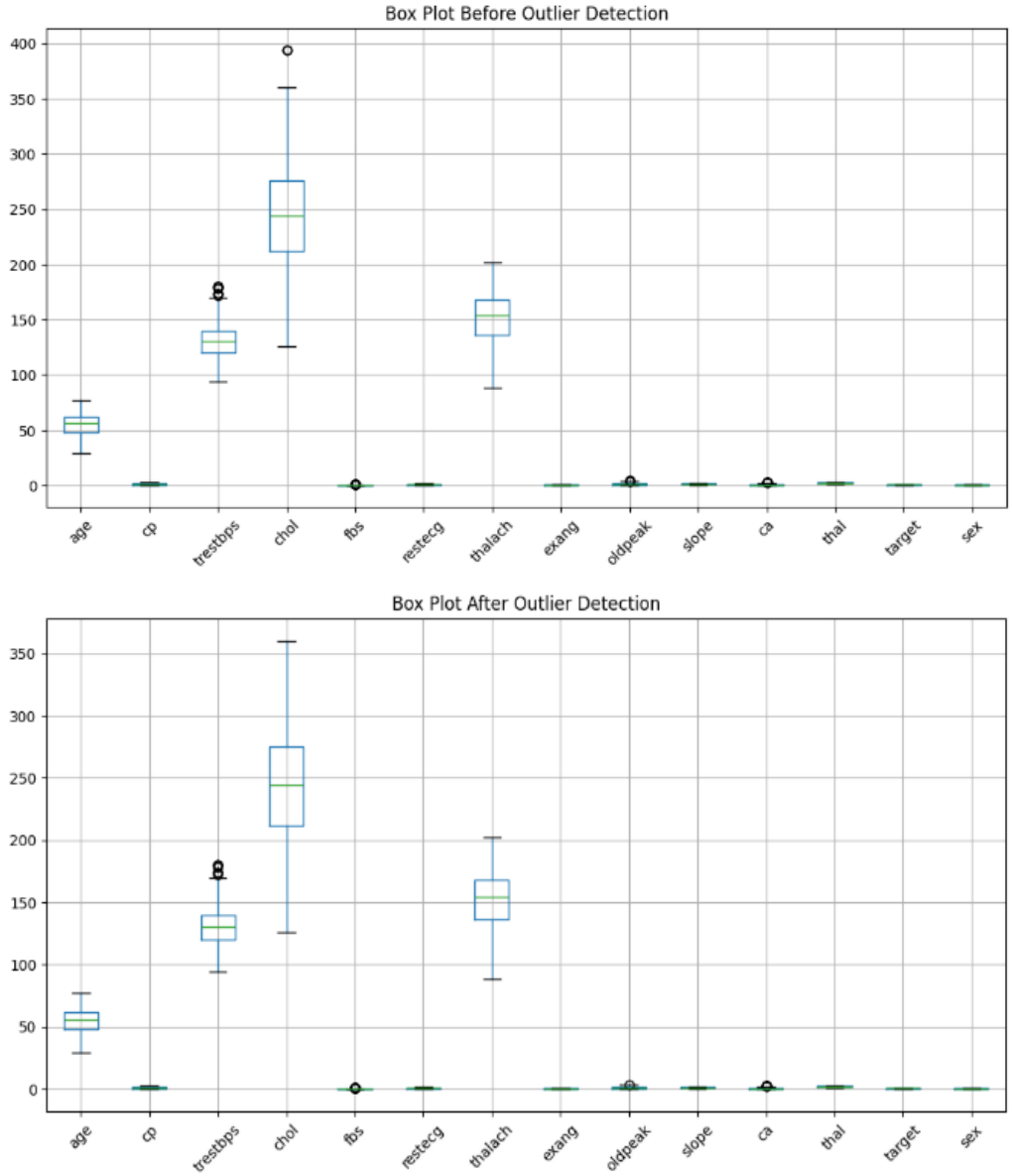

Fig. 6: Outliers Before and After applying Z-score

Standardization Using MaxAbsScaler

MaxAbsScaler was applied to scale the features of the dataset. For keeping sparsity, all features must be standardized to a scale of 0 to 1. Prior to standardization, all scales are very different in the dataset, and some features like age, chol, and thalach are having appreciable differences in

their range, which affects the performance of the model. For example, the mean ages and chol were 54.95 and 246.17 but with extreme differences in ranges as shown in Fig. 7. After standardization, the data set was transformed.

All features are within the range [0, 1]. For example, two critical features, such as age and chol, attained a mean at 0.71 and 0.68 with comparable ranges and standard deviation also minimal to denote scales were uniform. This scaling, besides preserving the distributional properties of the features, makes them more comparable across different magnitudes. These characteristics are needed for machine learning models that are sensitive to the magnitudes of their features; it ensures a much more stable performance during both training and prediction.

ISSN: 1213-4406

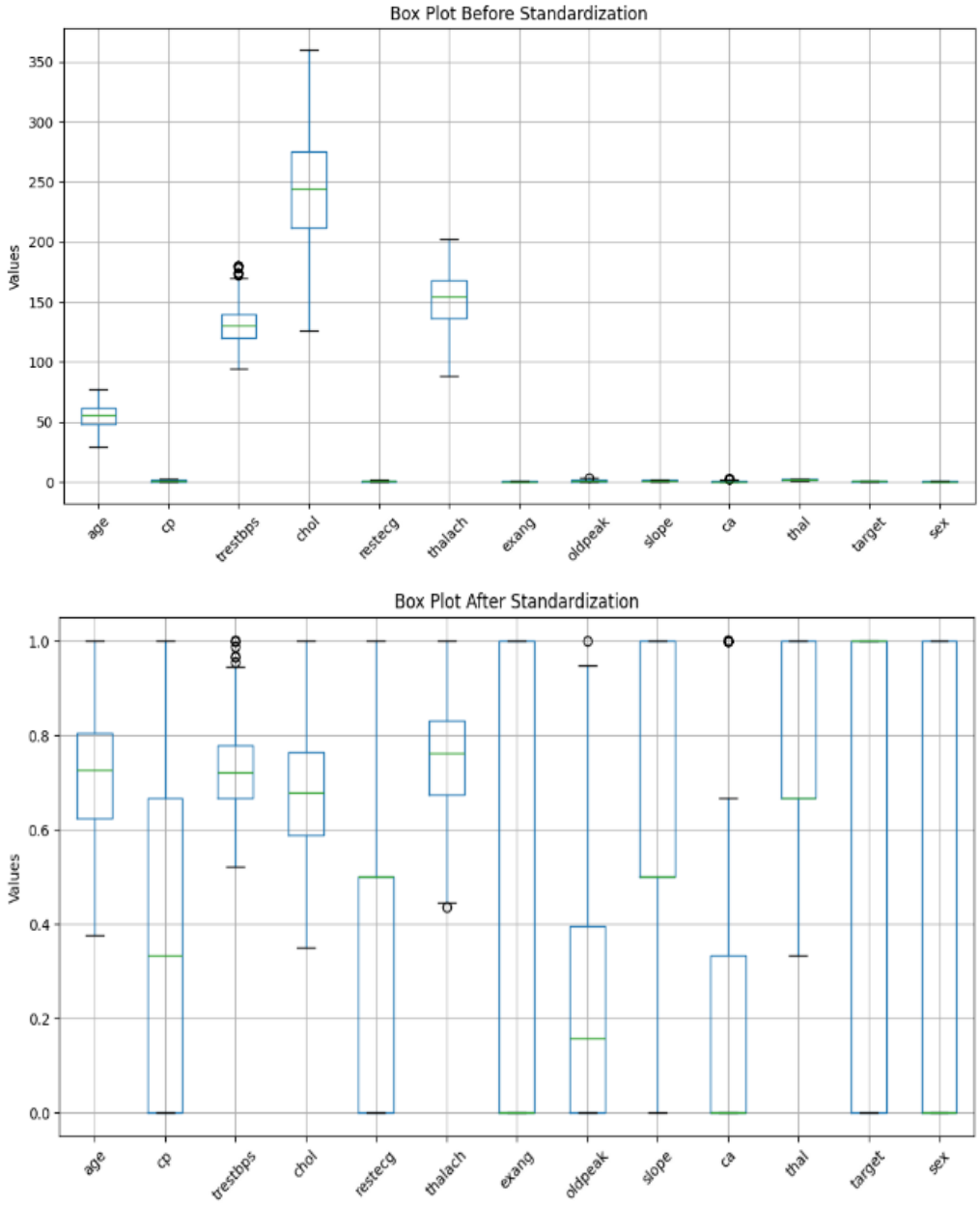

Fig. 7: Before and after Standardization Using MaxAbsScaler

Feature Selection Using GALS

The process for feature selection [28] with GALS combined will provide the most significant contributors towards the target variable. It will also ensure optimisation of the subset of features and robust model

performance by making use of evolutionary search as well as regularization techniques [29]. Top five selected features, ranked according to fitness scores: cp with a fitness score at 0.140592, oldpeak at 0.066871, chol at 0.047825, slope with a fitness score at 0.045880 and sex at a fitness score of 0.040113as shown in Fig. 8 and Table 1. The fitness score determines the importance of each feature in minimizing the error of prediction of the model.

Hence, the prominence of feature cp points out to critical importance for predicting the target variable, followed by oldpeak, which strongly affects outcomes. These results thus demonstrate the effectiveness of this combined feature selection methodology in reducing dimensionality and enhancing model accuracy. This framework facilitates computationally efficient predictive models that are interpretable at the same time.

Table 1: Features and their Fitness Scores

Feature	Fitness Score
Cp	0.140592
Oldpeak	0.066871
Chol	0.047825
Slope	0.045880
Sex	0.040113

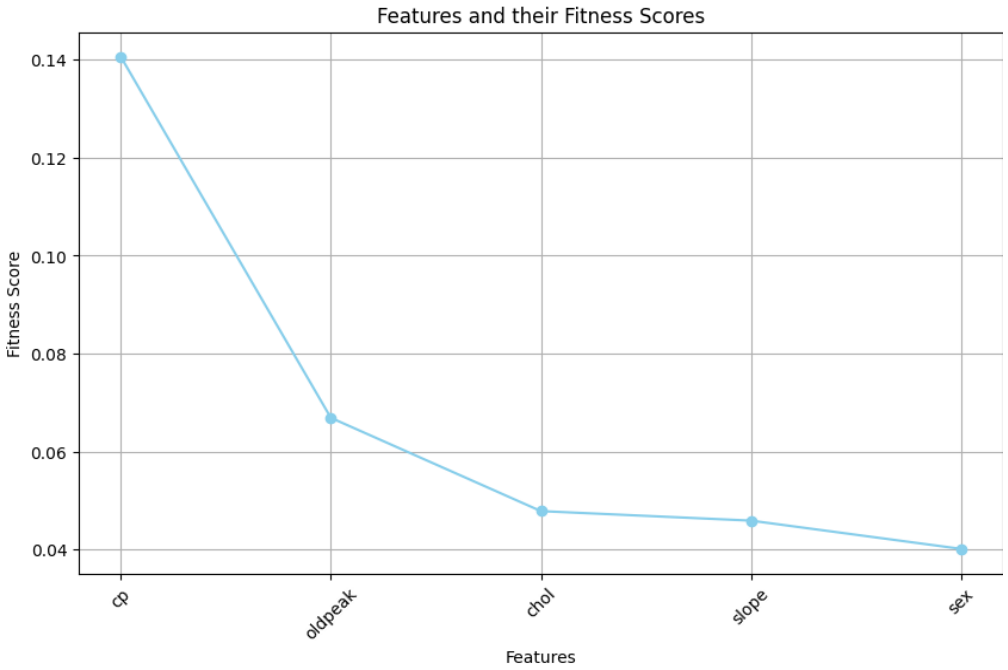

Fig. 8: A Scatter plot for selected features along with their fitness scores

Model Building Using SVM-KNN Hybrid Classifiers

The procedure of model construction which combines an SVM classifier with a KNN classifier produced some performance metrics of a promising level. The fitting was done by 5-fold cross-validation over 960 possible combination cases, making the analysis super-robust and reliable. Hyper-parameter optimization through GridSearchCV improved the accuracy of the model up to 94.87%, which shows that the model had a high classification capability for positive and negative cases.

The precision and recall are both at 88.89% to ensure a fair approach to true positive identification with the least number of false negatives. The F1 score of 88.89%, further emphasising this carefully negotiated trade-off as illustrated in Fig. 10. The ROC AUC [30] score, which is 0.9926, further indicates the capability of this model in differentiating classes. It serves to indicate its strength in dealing with imbalanced data as shown in Fig. 9 and Table 2. Strong justification exists now for employing an ensemble

approach along with GridSearchCV optimisation for making predictions in a reliable and accurate manner about classification problems.

Table 2: Model Performance Metrics

Metrics	Values
Accuracy	0.9487
Precision	0.8889
Recall	0.8889
F1 Score	0.8889
ROC AUC Score	0.9926

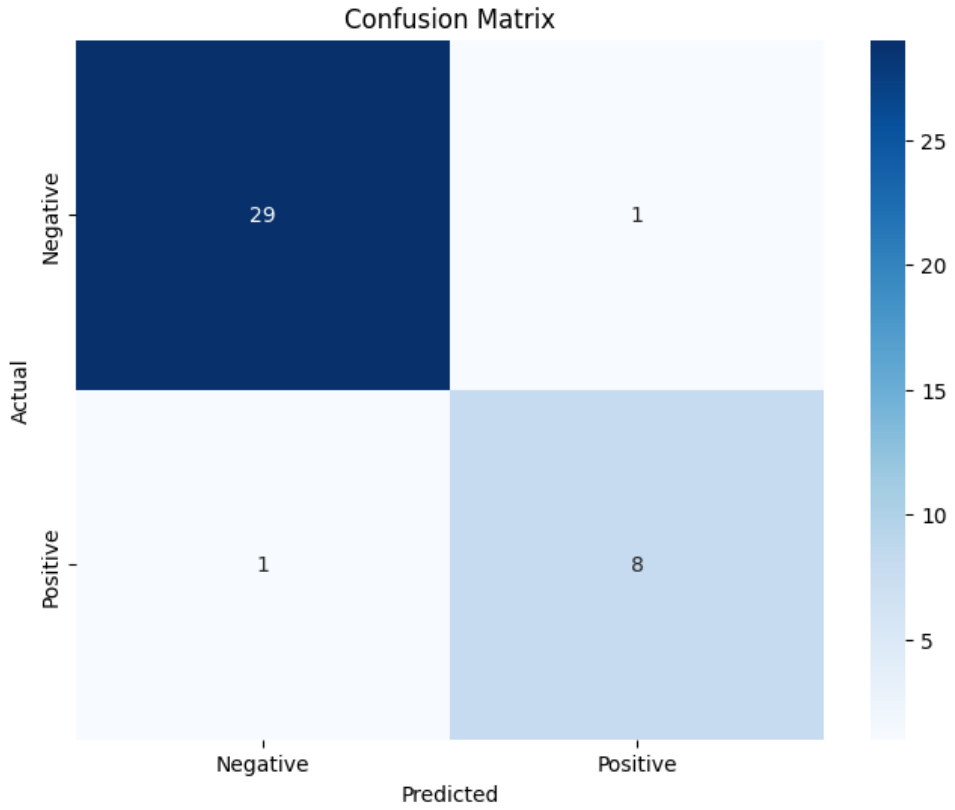

Fig-9: Confusion matrix

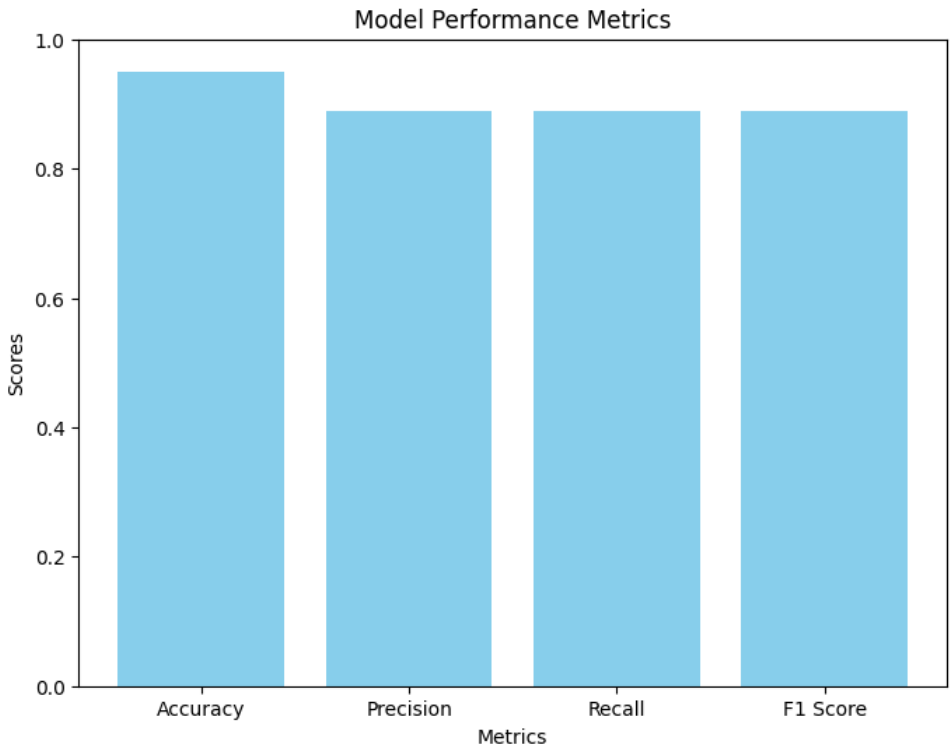

Fig-10: Bar plot for performance Mertics

Comparison with Existing Approaches

In comparison to previous studies, our work introduces a more robust and accurate approach for heart disease classification through the use of hybrid feature selection and ensemble learning. Thuraka [6] shows that feature selection techniques improve accuracy significantly in disease predictions, including heart diseases. Robison Spencer et al. [7] emphasize the variation in feature selection methods, which also showed that Chi-squared testing with BayesNet had 85% accuracy while PCA with IBK had the highest recall at 87.22%. Saqlain et al. [8] shined with unstructured data, where Naïve Bayes performed superiorly compared with others at 86.7% accuracy level with strong AUC levels of 92.4% as well to prove its immunity against categorical variables.

Amin Ul Haq et al. [9] used hybrid models that employed better feature selection technique and had their accuracy of logistic regression up to 89%, with specificity values reaching up to 98%. Nourmohammadi-Khiarak et al.

[10] designed a metaheuristic-based hybrid model, which surpassed the traditional classifiers with an accuracy of 88.25% and a sensitivity of 94.2%, and it proved to be efficient with less features. Summed up details are shown in Table.3 and Fig.11. The cumulative findings point out the revolutionizing efficiency of new feature selection and classification methods in orienting future more accurate, effective, and non-invasive diagnosis, especially when applied to real-time data sets, missing data scenarios, and greater application of metaheuristic.

Table 3: Comparison of Methodologies and Accuracy in Heart Disease Prediction

Author(s)	Methodology	Accuracy
Robison Spencer et al. (2020)	PCA, Chi-squared, Relief, Symmetrical Uncertainty with BayesNet and IBK for feature selection and classification.	85.00% (Chi-squared with BayesNet), Highest recall: 87.22% (PCA with IBK)
Saqlain et al. (2016)	Machine learning on unstructured text data using Naive Bayes, Logistic Regression, Neural Networks, SVM, Random Forests, Decision Trees.	86.7% (Naive Bayes), AUC: 92.4%
Amin Ul Haq et al. (2018)	Hybrid system with Relief, mRMR, LASSO for feature selection; classifiers include Logistic Regression, K-NN, ANN, SVM, Naive Bayes, Decision Tree, Random Forest.	89% (Logistic Regression with Relief), 87% (SVM with RBF kernel)
Nourmohammadi-Khiarak et al. (2019)	Imperialist Competitive Algorithm for metaheuristic feature selection with K-NN classifier; compared with Naive Bayes, Decision Trees, Neural Networks, SVM.	88.25% (Hybrid method with K-NN), Sensitivity: 94.2%, Specificity: 83.4%

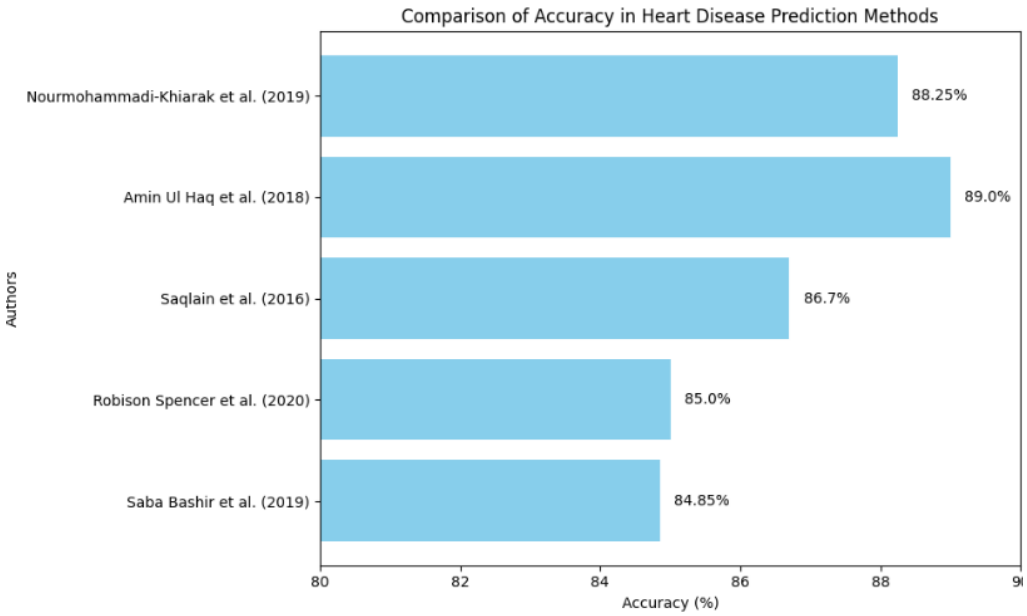

Fig. 11: Comparing the accuracy of different methods

Discussion

Diverse datasets-meaning varied demographics, clinical parameters, and geographical data-improve generalizability about the heart disease prediction models. Such varied datasets reduce bias during prediction models and, therefore, increase the robustness of models to make them reliable in all kinds of patient populations [6, 31]. Variability in training data allows models to learn nuanced patterns; overfitting is, therefore, minimized. Standardized preprocessing techniques, including MaxAbs scaling and outlier handling, combine to ensure consistency across diverse data sources. This approach will increase the model's accuracy in predicting the outcome by addressing differences in feature distributions and enhancing clinical applicability for heterogeneous populations **(RQ1 answered)**.

Feature subset optimisation by combining feature selection techniques, which includes Lasso regression and genetic algorithms, reduces the redundancy of features and retains critical predictors. Such methods refine data sets that focus on variables most contributing to model performance, hence making it interpretable and computationally efficient. Ensemble

algorithms, such as voting classifiers, have the benefit of integrating the strengths of multiple models, such as SVM and KNN, for higher accuracy and robustness [31]. The hybrid approach decreases the chances of overfitting and improves performance by reducing the variance and bias. This holistic methodology thus ensures efficient dimensionality reduction, good model generalization, and superior predictive accuracy in heart disease classification **(RQ2 answered)**.

Deep learning models, in the form of neural networks, are able to capture better, more complex patterns in high-dimensional data faster than traditional algorithms and often with greater predictive accuracy [32]. Deep learning does not use the traditional approach but instead employs higher powerful architectures like CNNs and RNNs that can deal with nonlinear relationships and unstructured data. They also require large computational resources and larger datasets for training. Traditional approaches, including combinations of SVM-KNN, were competitive in accuracy, computationally efficient, and interpretable. Classical ensemble approaches integrated with deep learning models can provide a balanced solution to cope with the contradiction between computational constraints and the demands of data requirements to ensure that high accuracy and interpretability are maintained **(RQ3 answered)**.

Precision, recall, F1 score and ROC AUC are also emphasized in heart disease prediction models beyond accuracy. Precision excludes false positives while reducing unnecessary treatment [33]. This is important in preventing adverse outcomes from happening. Recall correctly identifies actual cases; this is critical in preventing adverse outcomes from occurring. The F1 score is the balance between precision and recall that indicates reliability in detecting heart diseases. ROC AUC measures discrimination between classes and represents strength in handling imbalanced datasets. In addition, it should have good interpretability, sensitivity, and computational efficiency to be integrated well with the diagnostic workflows. Therefore, the metrics enhance the accuracy of prediction with a model and improve actionable insights into patient benefits and clinical decisions **(RQ4 answered)**.

Conclusion

This methodology emphasizes on how predictive modelling is very important in heart disease diagnosis, and it indicates how data-driven approaches improve clinical decision-making. The results show that there is an opportunity to obtain good accuracy as well as reliability in outcome prediction that consequently leads to timely intervention and treatment. By using the variety of datasets in the study, the generalisability and robustness across various populations for bias minimization and real-world improvement in applicability is addressed. Performance metrics such as precision, recall, F1 score, and ROC AUC demonstrate the model's capability to balance sensitivity and specificity to make it potentially *integratable* with clinical workflows. Findings such as these reinforce the importance of machine learning in creating actionable insights and thus can cause improvements in patient outcomes.

Future research might involve increasing sample size for greater representation, adding diverse demographic, clinical, and geographical distributions for future improvement of generalisability of prediction models. Using data streams in real time from wearable devices and electronic health records, it could create dynamic personalized risk scoring. Also, hybridisation using older algorithms with newer deep learning architecture would balance interpretability with computational efficiency. Collaboration with clinicians and experts in the field would go a long way in refining models for embedding in diagnostic workflows and elucidating ethical issues such as data privacy and discriminatory bias. This will eventually lead to developing large-scale comprehensive systems that improve health systems and promote the prevention and detection of heart diseases worldwide.

REFERENCES

[1] Mohan, S., Thirumalai, C. and Srivastava, G. (2019, Jan.). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542–81554. doi: 10.1109/access.2019.2923707. Available: https://doi.org/10.1109/access.2019.2923707

[2] Pahwa, K. and Kumar, R. (2017, Oct.). Prediction of heart disease using hybrid technique for selecting features. 2017 4th IEEE Uttar Pradesh Section International Conference on

Electrical, Computer and Electronics (UPCON). doi: 10.1109/upcon.2017.8251100. https://doi.org/10.1109/upcon.2017.8251100
[3] Bouma, B. J., and Mulder, B. J. (2017). Changing landscape of congenital heart disease. Circulation Research, 120(6), 908–922. https://doi.org/10.1161/circresaha.116.309302
[4] Katarya, R., and Srinivas, P. (2020). Predicting heart disease at early stages using machine learning: A survey. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 302–305. https://doi.org/10.1109/icesc48915.2020.9155586
[5] Gavhane, A., Kokkula, G., Pandya, I., and Devadkar, K. (2018). Prediction of heart disease using machine learning. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 1275–1278. https://doi.org/10.1109/iceca.2018.8474922
[6] Thuraka, B. (2021). Machine learning, advanced health informatics, and diagnostic improvement opportunities. Interdisciplinary Journal of African & Asian Studies (IJAAS), 7(2), 1-10. ISSN: 2504-8694.
[7] Spencer, R., Thabtah, F., Abdelhamid, N. and Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. DIGITAL HEALTH, 6. doi: 10.1177/2055207620914777.
[8] Saqlain, M., Hussain, W., Saqib, N. A. and Khan, M. A. (2016, Aug.). Identification of heart failure by using unstructured data of cardiac patients. 45th International Conference on Parallel Processing Workshops (ICPPW). doi: 10.1109/icppw.2016.66. Available: https://doi.org/10.1109/icppw.2016.66
[9] Haq, U., Li, J. P., Memon, M. H., Nazir, S. and Sun, R. (2018, Dec.). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 1–21. doi: 10.1155/2018/3860146. Available: https://doi.org/10.1155/2018/3860146
[10] Nourmohammadi-Khiarak, J., Feizi-Derakhshi, M.-R., Behrouzi, K., Mazaheri, S., Zamani-Harghalani, Y. and Tayebi, R. M. (2019, Dec.). New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. Health and Technology, 10(3), 667–678. doi: 10.1007/s12553- 019-00396-3. Available: https://doi.org/10.1007/s12553-019-00396-3
[11] Selvarajan, N. G. P. (2019, Sep.). Integrating machine learning algorithms with OLAP systems for enhanced predictive analytics. World Journal of Advanced Research and Reviews, 3(3), 062–071. doi: 10.30574/wjarr.2019.3.3.0064. https://doi.org/10.30574/wjarr.2019.3.3.0064
[12] Zhang, H., Wang, J., Sun, Z., Zurada, J. M. and Pal, N. R. (2019, Jan.). Feature selection for neural networks using group lasso regularization. IEEE Transactions on Knowledge and Data Engineering, 32(4), 659–673. doi: 10.1109/tkde.2019.2893266. https://doi.org/10.1109/tkde.2019.2893266
[13] Kramer, O. (2017). "Genetic algorithms," In Studies in computational intelligence (11–19). https://doi.org/10.1007/978-3-319-52156-5_2
[14] Tao, P., Sun, Z., and Sun, Z. (2018). An improved intrusion detection algorithm based on GA and SVM. IEEE Access, 6, 13624–13631.https://doi.org/10.1109/access.2018.2810198
[15] Wang, L. (2019). Research and implementation of machine learning classifier based on KNN. IOP Conference Series Materials Science and Engineering, 677(5), 052038. https://doi.org/10.1088/1757-899x/677/5/052038
[16] Elbadawi, M., Gaisford, S., and Basit, A. W. (2020). Advanced machine-learning techniques in drug discovery. Drug Discovery Today, 26(3), 769–777. https://doi.org/10.1016/j.drudis.2020.12.003

[17] Chen, R., Dewi, C., Huang, S., and Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00327-4
[18] Hayaty, M., Muthmainah, S. and Ghufran, S. M. (2021, Jan.). Random and synthetic over-sampling approach to resolve data imbalance in classification. International Journal of Artificial Intelligence Research, 4(2). doi: 10.29099/ijair.v4i2.152. Available: https://doi.org/10.29099/ijair.v4i2.152
[19] Silva, L. et al. (2019, Jan.). "Outliers treatment to improve the recognition of voice pathologies. Procedia Computer Science, 164, 678–685. doi: 10.1016/j.procs.2019.12.235. Available: https://doi.org/10.1016/j.procs.2019.12.235
[20] Halim, K. N. A., Jaya, A. S. M. and Fadzil, A. F. A. (2020, Jan.). Data pre-processing algorithm for neural network binary classification model in bank tele-marketing. International Journal of Innovative Technology and Exploring Engineering, 9(3), 272–277. doi: 10.35940/ijitee.c8472.019320. Available: https://doi.org/10.35940/ijitee.c8472.019320
[21] Emmert-Streib, F. and Dehmer, M. (2019, Jan.). High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection. Machine Learning and Knowledge Extraction, 1(1), 359–383. doi: 10.3390/make1010021. Available: https://doi.org/10.3390/make1010021
[22] Varotsis, G. (2017, Oct.). The plot-algorithm for problem-solving in narrative and dramatic writing. New Writing, vol. 15, no. 3, pp. 333–347. doi: 10.1080/14790726.2017.1374414. Available: https://doi.org/10.1080/14790726.2017.1374414
[23] Chung, H., and Shin, K. (2019). Genetic algorithm-optimized multi-channel convolutional neural network for stock market prediction. Neural Computing and Applications, 32(12), 7897–7914. https://doi.org/10.1007/s00521-019-04236-3
[24] Muthukrishnan, R., and Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. IEEE. https://doi.org/10.1109/icaca.2016.7887916
[25] Pisner, D. A., and Schnyer, D. M. (2019). Support vector machine. In Machine learning (pp. 101–121). https://doi.org/10.1016/b978-0-12-815739-8.00006-7
[26] Moldagulova, A., and Sulaiman, R. B. (2017). Using KNN algorithm for classification of textual documents. IEEE, 665–671. https://doi.org/10.1109/icitech.2017.8079924
[27] Idhammad, M., Afdel, K., and Belouch, M. (2018). Semi-supervised machine learning approach for DDoS detection. Applied Intelligence, 48(10), 3193–3208. https://doi.org/10.1007/s10489-018-1141-2
[28] Cai, J., Luo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077
[29] Thuraka, B. (2022). Impact of technological innovations on healthcare service delivery in US: Medicare and Medicaid as facilitators. Interdisciplinary Journal of African & Asian Studies (IJAAS), 8(1), 1-10. ISSN: 2504-8694.
[30] Brzezinski, D., and Stefanowski, J. (2017). Prequential AUC: properties of the area under the ROC curve for data streams with concept drift. Knowledge and Information Systems, 52(2), 531–562. https://doi.org/10.1007/s10115-017-1022-8
[31] Pasupuleti, V., and Inyang, L. (2022). Mitigating pancreatic cancer through data-driven AI techniques, holistic health record, iHELP and integrative systems. *International Journal of Health and Pharmaceutical Research*, 7(2), 63-77. E-ISSN 2545-5737 P-ISSN 2695-2165. www.iiardjournals.org
[32] Shawana, T. A. (2022). Carbon emissions as threats to environmental sustainability: Exploring conventional and technology-based remedies. African Journal of Environmental Sciences and Renewable Energy, 8 (1). https://publications.afropolitanjournals.com/index.php/ajesre/article/view/736

[33] Pasupuleti, V. (2021). AI-based multimedia security in combating adversarial attacks, deepfakes, and ethical concerns. Interdisciplinary Journal of African & Asian Studies (IJAAS), 7(1), (ISSN: 2504-8694), 1-15.