

Enhanced heart disease prediction using SKNDGR ensemble Machine Learning Model

Basil Jacob ¹

¹Basil Jacob, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore – 632014, Tamil Nadu, India

Abstract – Cardiovascular disease (CVD) stands as a formidable global health challenge, exerting a significant impact on worldwide morbidity and mortality rates. Traditional monitoring methods often prove inadequate in capturing the intricate and dynamic nature of cardiovascular health, impeding healthcare professionals' ability to discern subtle patterns preceding cardiac events. This study seeks to revolutionize heart disease monitoring by leveraging the prowess of well-established machine learning algorithms, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Gradient Boosting, and Random Forest. Diverging from conventional single-model approaches, our work embraces ensemble learning, synergizing the strengths of standalone models—SVM, KNN, Naive Bayes, Decision Tree—with ensemble models like Gradient Boosting and Random Forest. Through the amalgamation of these models, we introduce a novel ensemble learning approach, named the SKNDGR model, which achieves an outstanding accuracy of 99.19%, surpassing the performance of all other models. This exceptional outcome is attributed to the model's capacity to establish diverse, robust, and non-linear decision boundaries, necessitating minimal hyperparameter tuning and ensuring computational efficiency. The research holds the promise of swift heart disease detection, facilitating timely treatment interventions, and showcases advantages such as high performance, accuracy rates, flexibility, and an elevated success rate. The proposed model provides a high degree of precision and recall, enabling researchers to obtain the most accurate results when diagnosing patients suffering from heart disease.

Key Words: Cardiovascular disease, Machine Learning Algorithms, Ensemble learning, SKNDGR, Hyperparameter Tuning, Computational efficiency

1. INTRODUCTION

In the relentless fight against cardiovascular diseases (CVD), where the lives of over half a billion individuals hang in the balance worldwide, the urgency is glaringly apparent. The World Health Federation's 2023 report paints a grim picture, revealing 20.5 million CVD-related deaths in 2021—nearly one-third of all global fatalities—an alarming surge from the estimated 121 million CVD deaths. This ominous trend, fueled by unhealthy lifestyles, heightened stress levels, and environmental factors, emphasizes the critical necessity for

prompt and accurate diagnostic solutions to enable timely interventions. Recognizing this imperative, the quest for a diagnostic system rooted in machine learning classifiers becomes paramount.

In the intricate tapestry of healthcare, precision becomes the linchpin between life and death. Here, data preprocessing and standardization emerge as silent architects, akin to a skilled detective organizing clues to solve a complex case. Healthcare professionals depend on the seamless harmony of standardized data, unlocking crucial insights for informed decisions in the pursuit of optimal patient care. Some commonly used standardization techniques, including StandardScaler (SS) and Min-Max Scaler, are effective in rescaling numerical features. However, they do not handle missing feature values directly, and it's necessary to address missing values through imputation or other methods before applying these standardization techniques.

Achieving peak performance in machine learning models demands the strategic utilization of balanced datasets during both training and testing phases. Furthermore, refining predictive capabilities hinges on the incorporation of relevant features, underscoring the critical importance of optimizing model performance through data balancing and meticulous feature selection. In this high-stakes landscape, the development of advanced diagnostic tools powered by machine learning classifiers stands as a beacon of hope, promising to revolutionize the battle against CVD and reshape the future of global healthcare.

In standard practice, doctors often prescribe multiple tests, creating delays in timely disease notification. This underscores the need for an approach enabling swift and timely predictions. Machine learning, a subset of artificial intelligence, operates on the premise that systems can learn from data, identify patterns, and make decisions with minimal human intervention. For heart disease prediction, researchers normally use the existing machine learning algorithms such as SVM, KNN, Decision Tree, Random Forest, Naïve Bayes, Logistic regression and so on. However, it is quite difficult to get predictions with a high degree of accuracy with these models functioning alone. Our research work aims to utilize the combined power of these models and generate a new model that returns results with a high degree of precision and accuracy.

In this paper, we proposed a machine learning classification model which combines the prediction capabilities of each of the following individual algorithms: SVM, KNN, Naïve Bayes, Decision Tree, Gradient Boosting and Random Forest to create a single powerful ensemble model capable of producing results with a high degree of accuracy and precision. Our work has been organized as follows: Section 2, a literature review of the various papers that work on machine learning classification for heart diseases. Section 3 Proposed methodology followed for our study. Section 4 Evaluation Metrics used to compare the models used. Section 5, Discusses the result obtained. Section 6, concludes the paper and Section 7, provides suggestions for future work.

2. LITERATURE SURVEY

This paper [1] proposes a heart disease prediction system using the Random Forest algorithm, emphasizing its efficiency in classification and regression. The application processes input from a CSV file, employs an ensemble of decision trees for accurate predictions, and facilitates early detection of heart conditions. Users can input their health data to assess the likelihood of heart disease, enabling informed decisions on whether to consult a doctor.

Ibomoie Domor Mienye et al. [2] have introduced an improved machine learning method for heart disease risk prediction. The approach utilizes mean-based dataset partitioning, employing classification and regression tree (CART) models. A homogeneous ensemble is formed using an accuracy-based weighted aging classifier ensemble. Experimental results on the Cleveland and Framingham datasets yield impressive classification accuracies of 93% and 91%, respectively, outperforming other algorithms and recent scholarly works. Comparative evaluations with k-nearest neighbor (KNN), logistic regression (LR), linear discriminant analysis (LDA), support vector machine (SVM), CART, gradient boosting, and random forest highlight the superior performance of the proposed method. This study establishes the proposed approach as an advanced and effective tool for heart disease risk prediction.

In this paper [3], the researchers conducted a rigorous comparative analysis of four prominent machine learning algorithms—k-nearest neighbor, decision tree, linear regression, and support vector machine (SVM)—utilizing the UCI repository dataset. The primary focus was to ascertain the algorithm exhibiting the highest test accuracy, a pivotal metric evaluated through a detailed examination facilitated by a confusion matrix. Notably, the study's conclusive findings underscored the exceptional performance of the k-nearest neighbor (knn) algorithm, boasting an impressive accuracy score of 87%.

M. Nikhil Kumar et al. [4] have introduced a novel model enhancing Decision Tree accuracy in identifying heart disease patients, leveraging various Decision Tree

algorithms. Utilizing the Waikato Environment for Knowledge Analysis (WEKA), the study preprocesses UCI repository data and underscores WEKA's pivotal role in machine learning. Focusing on decision tree classifiers, particularly J48, Logistic Model Tree, and Random Forest algorithms, the research employs reduced error pruning, confident factor, and seed parameters for heart disease diagnosis. Experimental results reveal J48 as the optimal classifier, achieving 56.76% accuracy and 0.04 seconds to build, outperforming LMT and Random Forest.

Min Chen et al. [5] in this study employ cutting-edge machine learning techniques for disease prediction, utilizing K-Nearest Neighbor (KNN) and Convolutional Neural Network (CNN) algorithms. They address incomplete data challenges through a latent factor model and propose the Convolutional Neural Network-based Multimodal Disease Risk Prediction (CNN-MDRP) algorithm. Remarkably, the CNN-MDRP achieves an impressive 94.8% prediction accuracy, outperforming benchmarks and demonstrating superior convergence speed compared to CNN-based unimodal predictions. This research, led by the authors, represents a significant leap in disease risk assessment, emphasizing both accuracy and efficiency in predictive modeling.

This research [6] compares the efficacy of Support Vector Machine (SVM) classifier and Linear Regression (LR) model in detecting heart disease using a dataset from the UCI machine learning repository. The study, with a sample size of 60, demonstrates that SVM achieves a 90.43% accuracy, outperforming LR with 78.56%. The results indicate the significance ($p=0.021$) and emphasize SVM's superior accuracy for heart disease detection.

This study [7] proposes a comprehensive disease prediction model based on patient symptoms, employing K-Nearest Neighbor (KNN) and Convolutional Neural Network (CNN) algorithms for accurate predictions. The model utilizes a dataset of disease symptoms, incorporating living habits and checkup information for precision. The CNN algorithm achieves an accuracy of 84.5%, surpassing KNN, with lower time and memory requirements. Post-prediction, the system assesses the risk associated with general diseases, categorizing them into lower or higher risk levels. This research highlights the effectiveness of CNN in enhancing accuracy and efficiency in general disease prediction.

This study [8] introduces the use of the Multi-layer Perceptron Neural Network with Back-propagation as a training algorithm to propose an innovative diagnostic system for heart disease prediction. Leveraging 14 significant attributes based on medical literature, the system surpasses alternative approaches, effectively predicting heart disease risk levels. Notably, the prediction system achieves a superior accuracy of 93.39% for 5 neurons in the hidden layer, with a swift runtime of 3.86 seconds.

Comparative analyses against various classification techniques, including Decision Tree, Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine, Generalized Linear Model, Gradient Boosted Trees, and Deep Learning, highlight the superior efficiency and accuracy of the proposed system in heart disease prediction.

The authors [9] have employed an ensemble approach using the Cleveland dataset, combining Majority Vote with MP, RF, BN, and NB. Through attribute selection, they achieved a peak accuracy of 85.48%, recommending this method for heart disease prediction.

Chunyan Guo et al. [10] made a substantial contribution to the field by employing an Recursion enhanced random forest with an improved linear model (RFRF-ILM). The integration of innovative feature combinations and categorization methods showcased their dedication to enhancing performance and accuracy.

This study [11] focuses on the application of the Random Forest data mining algorithm for predicting heart disease. The implementation resulted in a notable classification accuracy of 86.9%, coupled with a diagnosis rate of 93.3% utilizing the Random Forest algorithm.

In this paper [12], the authors have compared five machine learning algorithms—Random Forest classification, Support Vector Machine, AdaBoost Classifier, Logistic Regression, and Decision Tree Classifier—to identify the most accurate model. The Random Forest Classifier outperformed others with an accuracy of 85.22%. While alternative algorithms achieved accuracy levels above 50%, Random Forest emerged as the superior choice. This study lays the groundwork for further research in optimizing accuracy models.

Nadia Rubaiyat et al. [13] of this study, have applied Logistic Regression, Random Forest Classifier, and k-Nearest Neighbor algorithms to a dataset of 310 patients, revealing Random Forest as the most accurate at 89%. The study asserts that Random Forest surpasses the other models and suggests its potential in medical fields. However, due to the limited dataset, further processing and a larger population are needed for robust conclusions.

Aditi Gavhane et al. [14] propose developing an application for predicting heart disease vulnerability based on fundamental symptoms like age, sex, and pulse rate. Advocating for the implementation of neural networks due to their established accuracy and reliability, the authors position them as a central component of the proposed system. The presented method demonstrates promising results, achieving precision and recall scores of 0.91 and 0.89, respectively.

The authors [15] have explored diverse classifiers and the impact of data processing on predicting heart disease. Findings highlight Logistic Regression (86%) and Naive Bayes (79%) as effective with high-dimensional datasets, while Decision Tree (70%) and Random Forest (84%) excel with smaller dimensional datasets. Notably, Random Forest outperforms the Decision Tree Classifier due to its optimized learning algorithm. These insights underscore the critical role of data processing in achieving accurate predictions for heart disease.

This research [16] work, explores the impact of feature selection methods, specifically Correlation Based Feature Selection (CBFS) and Principal Component Analysis (PCA), when combined with eight classifier models. Performance metrics, including Accuracy, Precision, Recall, F1 Measure, and ROC, guide the assessment to identify the most effective classifier. The study reveals that the CBFS with MLP Classifier model stands out with the highest performance for the FHS dataset. Among the eight generated models, CBFS-MLP proves superior to the standalone MLP classifier, showcasing enhanced accuracy.

This research [17] paper predicts heart disease probability, with K-nearest neighbor yielding the highest accuracy. The study discusses and compares algorithms used for heart disease prediction, applying four classification techniques—K-nearest neighbor, Naive Bayes, decision tree, and random forest. Notably, K-nearest neighbor, Naive Bayes, and random forest show the best results, with the highest accuracy (90.789%) achieved by K-nearest neighbors ($k = 7$).

This research [18] focuses on predicting cardiovascular heart diseases using patient medical attributes. Utilizing a UCI repository dataset, the study employs 14 attributes, including gender, age, chest pain, and fasting sugar levels, and applies Logistic Regression, KNN, and Random Forest Classifier. The results demonstrate enhanced heart disease prediction with Logistic Regression and KNN, achieving an average accuracy of 87.5%, surpassing prior models at 85%. Notably, KNN stands out with the highest accuracy among the three algorithms at 88.52%, highlighting its effectiveness in this context.

This paper [19] introduces an innovative approach, NNNT (Neural Network and Decision Tree), leveraging Neural Network for model training and Decision Tree for testing classification, aiming to enhance heart disease prediction. Comparative analysis with Naive Bayes, Support Vector Machine, Neural Network, Voted Perceptron, and Decision Tree algorithms demonstrates superior accuracy and performance. The proposed model achieves remarkable precision and recall scores of 99.2 and 98.4, surpassing the average scores of around 84 in other methods. This study provides a novel perspective for researchers to analyze heart disease data, contributing to more effective predictions and better maintenance of human health.

In this research [20], the authors scrutinized a Kaggle dataset encompassing vital attributes linked to heart disease, such as age, gender, blood pressure, and cholesterol. Evaluation of machine learning techniques, including Support Vector Machines (SVM), K-Nearest Neighbor (KNN), and Decision Trees (DT), revealed suboptimal performance and accuracy when dealing with extensive datasets. To overcome this limitation, the study aimed at elevating prediction accuracy through the implementation of an Artificial Neural Network (ANN) using TensorFlow Keras. The outcome was striking, with the artificial neural network achieving an impressive accuracy of 85.24%.

In conclusion, the reviewed papers underscore remarkable advancements in heart disease classification through established machine learning architectures. Analyzing the performance metrics across these studies reveals the potential for further efficiency gains by leveraging the predictive capabilities of individual methods through strategic combinations.

3. PROPOSED METHODOLOGY

In our research on heart disease prediction, we systematically processed the dataset [21] through standardization and feature selection techniques to distill key aspects. The dataset underwent a strategic partitioning using the 'train_test_split' function, with 70% allocated for training and the remaining 30% for temporary data. This temporary data was further divided into a 40% validation set and a 60% testing set, all under a fixed random state of 42 for consistency. This meticulous approach ensures an effective evaluation of machine learning models, providing a solid foundation for learning, tuning, and assessing performance with reproducible results. The innovative SKNDGR (SVM, KNN, Naive Bayes, Decision Tree, Gradient Boosting, Random Forest) model was employed in the subsequent classification phase to predict heart disease based on the refined features. The model's effectiveness was comprehensively evaluated, comparing its performance with existing methods using diverse metrics. This systematic pipeline ensures robust training, validation, and assessment, contributing to the reliability and reproducibility of our research outcomes.

3.1 Data Source

Our investigation relies on a rich and meticulously curated heart disease prediction dataset sourced from Kaggle, dating back to 1988 and spanning four distinctive databases: Cleveland, Hungary, Switzerland, and Long Beach V. Although the dataset boasts a comprehensive 76 attributes, encapsulating crucial factors for predicting heart disease, our focus hones in on a strategically chosen subset of 14 attributes meticulously detailed in Table 1. This dataset, comprising a total of 1025 patient records, demonstrates an

equitable distribution, with 51.3% reflecting cases of heart disease and 48.7% without. For the meticulous training of our predictive model, we harness the power of the initial 13 attributes, while the last attribute functions as the pivotal target variable, defining the predicted class. This finely tuned subset not only propels the training phase but also undergoes rigorous evaluation against a dedicated test dataset, ensuring the robustness of our model. This comprehensive dataset offers a diverse array of patient characteristics and health indicators, providing a robust foundation for our heart disease prediction model.

Sl No	Attribute Name	Attribute Description
1	age	Age in years
2	sex	1 for Male, 0 for Female
3	cp	Chest pain Type (values 0,1,2,3)
4	trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
5	chol	serum cholestoral in mg/dl
6	fbs	fasting blood sugar (1 = True, 0 = False)
7	restecg	Resting electrocardiographic results (values 0,1,2)
8	thalach	Maximum heart rate achieved
9	exang	exercise induced angina (1 = yes, 0 = no)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	The slope of the peak exercise ST segment (values 0,1,2)
12	ca	Number of major vessels colored by flourosopy (values 0,1,2,3)
13	thal	Defect type (0 = normal; 1 = fixed defect; 2 = reversable defect)

Table -1: Input Parameters

The below flow diagram illustrates the workflow followed for the creation of our model:

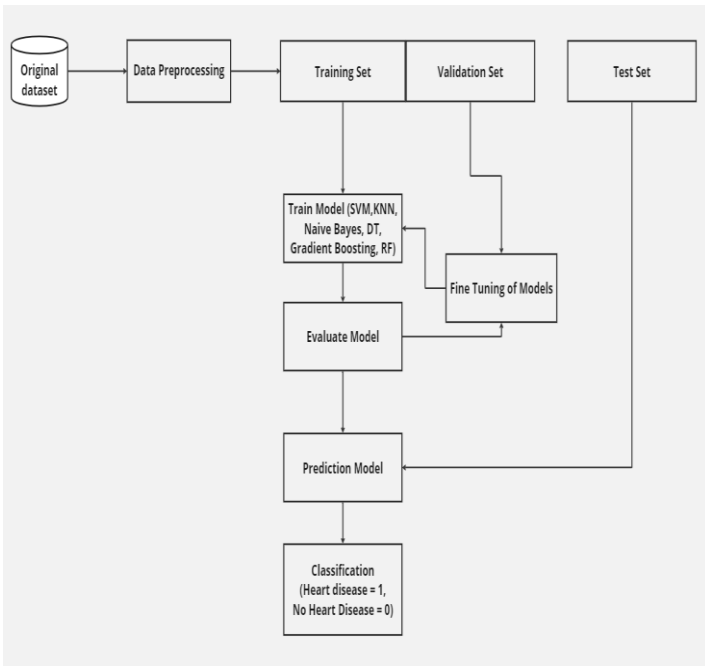


Fig -1: Workflow Diagram

3.2 Data Preprocessing

Data preprocessing constitutes a pivotal phase in classification problems, exerting a profound impact on the efficacy and dependability of machine learning models. The inherent quality of input data directly shapes a model's aptitude for pattern recognition and accurate predictions. In our research, we meticulously address data preprocessing through a two-step approach: Standardization and Feature Selection.

3.2.1 Standardization

This essential step involves transforming features to ensure a mean of 0 and a standard deviation of 1. This process is indispensable, particularly in machine learning models reliant on distance-based metrics, as it mitigates the undue influence of features with larger scales. To achieve this, we employ the `StandardScaler()` class from the scikit-learn Python library, ensuring a consistent and standardized representation of our data.

3.2.2 Feature Selection

A crucial aspect of our data preprocessing strategy is feature selection, yielding manifold benefits such as dimensionality reduction, improved model interpretability, and the potential for enhanced performance by focusing on salient features. Leveraging the `SelectKBest()` class from scikit-learn, we identify and retain the top 10 features out of the 13 trainable features in our dataset. These selected features, including age, sex, cp, chol, thalach, exang, oldpeak, slope, ca, and thal, are deemed pivotal for our classification task. This judicious

filtering ensures that our model is trained on the most pertinent and discriminative aspects of the dataset, fostering a more robust and insightful learning process.

3.3 Algorithm Implementation

The below are the machine learning and ensemble learning algorithms used for our study:

3.3.1 Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised classification algorithm deeply rooted in the principles of structural risk minimization (SRM), a concept pioneered by Vladimir Vapnik (Cortes & Vapnik, 1995). Initially tailored for binary classification tasks, SVM has undergone a transformative evolution, emerging as a versatile tool extensively applied in multiclassification scenarios. At its core, the algorithm endeavors to identify an optimal hyperplane within an N-dimensional space, strategically maximizing the margin between the closest points of distinct classes in the feature space. The adaptability of this hyperplane's dimensionality to the number of features underscores SVM's effectiveness in segregating data points into different classes while minimizing the expected generalization error. In our implementation, we initialize an SVM classifier with default parameters and proceed to train the model on the designated training set, paving the way for subsequent predictive analyses.

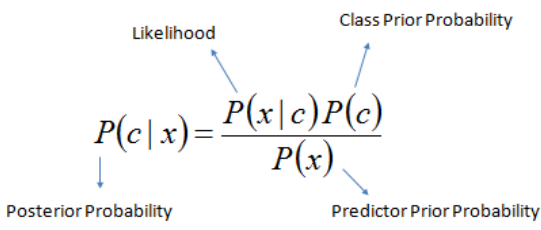
3.3.2 K-Nearest Neighbors (K-NN)

The K-nearest neighbors (K-NN) algorithm, a fundamental component of supervised classification in machine learning, distinguishes itself through its versatility and broad applicability. Recognized for its simplicity and ease of implementation, K-NN excels without requiring assumptions about the underlying data distribution. Its capacity to handle both numerical and categorical data renders it a flexible choice for a diverse range of datasets in classification and regression tasks. Operating as a non-parametric method, K-NN harnesses data point similarity for predictions, showcasing resilience to outliers compared to alternative algorithms. The algorithm identifies the K nearest neighbors to a given data point using a distance metric, such as Euclidean distance, and determines the point's classification or value through the majority vote or average of the K neighbors. In the realm of scientific research, K-NN emerges as a potent and accessible tool for data analysis and prediction. For our study, we initialized our training model with an initial value of k set to 15.

3.3.3 Naïve Bayes

Naive Bayes, a powerful family of probabilistic classifiers, thrives on the simplicity of Bayes' theorem, assuming strong independence between features. This seemingly naive

approach conceals robust predictive capabilities, surpassing even sophisticated methods and shining in large datasets. Particularly effective in classification problems, Naive Bayes emerges as a dynamic, reliable tool simplifying complexity for impactful results. Our code employs the Gaussian Naive Bayes classifier, poised to make predictions on new, unseen data, leveraging insights gleaned from the training set. Bayes' theorem offers a method for calculating the posterior probability, $P(c|x)$, using the probabilities $P(c)$, $P(x)$ and $P(x|c)$.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$


$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Fig -2: Bayes Theorem Formula

$P(c|x)$ represents the posterior probability of a class (target) given a predictor (attribute).

$P(c)$ is the prior probability of the class.

$P(x|c)$ is the likelihood, indicating the probability of the predictor given the class.

$P(x)$ is the prior probability of the predictor.

3.3.4 Decision Tree

A decision tree embodies a flowchart-like tree structure where each internal node signifies a feature, branches represent rules, and leaf nodes depict the algorithm's outcomes. As a versatile supervised machine-learning algorithm, it adeptly handles both classification and regression problems, showcasing its robustness. Notably powerful on its own, the decision tree also plays a pivotal role in Random Forest, contributing to its strength. In our implementation, the Decision Tree classifier is employed, with the parameter `random_state=42` set for reproducibility, ensuring consistent and replicable results across multiple executions.

Root Node

1. Represents the topmost node in the tree.
2. Symbolizes the entire dataset.
3. Initiates the decision-making process.

Decision/Internal Node

1. Represents a node that involves a choice based on an input feature.
2. Branches off to connect with leaf nodes or other internal nodes.

Leaf/Terminal Node

1. Represents a node without any child nodes.
2. Indicates the end of a branch.
3. Provides class labels or numerical values.

3.3.5 Gradient Boosting

Gradient Boosting is a widely utilized boosting algorithm in machine learning, tailored for both classification and regression tasks. Operating as an ensemble learning method, it sequentially trains models, with each new iteration dedicated to rectifying the errors of its predecessor. Through the amalgamation of numerous weak learners, Gradient Boosting amalgamates their collective strength, resulting in a robust and powerful learner that enhances predictive performance. In our implementation, the boosting model is trained on the training set, and the `random_state` parameter of the Gradient Boosting classifier is set to 42 for consistent and reproducible results.

3.3.6 Random Forest

The Random Forest Algorithm, a stalwart in the realm of supervised machine learning, garners widespread recognition for its effectiveness in handling both Classification and Regression tasks. Like a flourishing forest teeming with diverse trees, the strength of a Random Forest lies in its profusion of constituent trees. The algorithm's predictive accuracy and problem-solving prowess are intricately linked to the abundance of trees it encompasses. Functioning as a classifier, the Random Forest houses numerous decision trees, each operating on distinct subsets of the dataset. Through an astute aggregation of their outputs, the algorithm systematically elevates overall predictive accuracy. Anchored in the principles of ensemble learning, this approach entails the fusion of multiple classifiers to adeptly navigate complex problems and enhance the overall performance of the model.

3.3.7 SKNDGR Proposed Model

Our proposed novel SKNDGR model, integrates the earlier mentioned machine learning algorithms and ensemble methods as its core components, utilizing a soft voting mechanism as seen in Figure 3. This forward-thinking ensemble strategy is intricately crafted to enhance overall

model accuracy by synthesizing nuanced insights from a diverse array of state-of-the-art algorithms. Through meticulous evaluations of individual and ensemble accuracies, we present a comprehensive assessment of the model's capabilities on the test set, highlighting its unparalleled performance in cutting-edge data analysis and prediction.

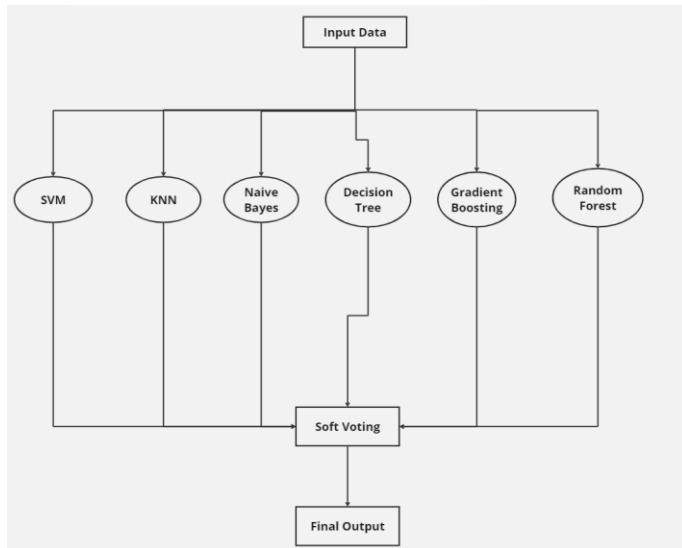


Fig -3: SKNDGR Model Architecture

3.3.8 Hyperparameter Tuning

In our code, hyperparameter tuning is performed for four different machine learning models—Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and Gradient Boosting—on validation data using the GridSearchCV method. For each model, a grid of hyperparameters is specified, and the model is trained with various combinations of these hyperparameters using cross-validation (cv=5). The hyperparameters with the best performance are determined based on the cross-validated results, and the respective models are re-instantiated with these optimal hyperparameters. This iterative process ensures that the models are fine-tuned to achieve the best performance on the validation set, enhancing their predictive capabilities for subsequent evaluation on unseen data. These fine tune models later served as the base of our proposed SKNDGR model. Fine-tuning hyperparameters in Naive Bayes is limited by its inherent simplicity and strong assumption of feature independence, with few parameters and robustness being key factors. Similarly, hyperparameter tuning in Random Forest may not yield substantial gains due to its built-in robustness against overfitting, aggregation of diverse tree predictions, and the algorithm's responsiveness to default settings. Our implementation abstains from fine-tuning both Naive Bayes and Random Forest, as their performance relies more on inherent characteristics and quality of features than intricate parameter adjustments, avoiding unnecessary computational expenses.

4. EVALUATION METRICS

In our proposed work, the training data is employed to generate the model, the validation data is used for fine-tuning, and the test data evaluates its performance. In the medical field, the evaluation of models emphasizes primary criteria for thorough assessments, encompassing factors such as:

4.1 Accuracy

Accuracy is a fundamental metric in machine learning that gauges the proportion of correctly predicted instances out of the total. It is calculated as the ratio of correct predictions to the overall predictions. While accuracy provides a straightforward interpretation and is suitable for balanced datasets, it has limitations. In imbalanced datasets, accuracy can be misleading as a model might achieve high accuracy by predicting the majority class, while performing poorly on the minority class. Additionally, accuracy doesn't consider the costs associated with different types of misclassifications. In such cases, other metrics like precision, recall, F1 score, or AUC may offer a more comprehensive evaluation of model performance.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

4.2 Precision

Precision, a key machine learning metric, gauges the accuracy of positive predictions by calculating the ratio of true positives to the sum of true positives and false positives. It reflects a model's ability to avoid false positives, crucial in fields like medical diagnoses or fraud detection. However, for a thorough evaluation, precision should be considered alongside metrics like recall and the F1 score, particularly in scenarios with imbalanced datasets.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

4.3 Recall

Recall is a measure assessing a model's capacity to detect and encompass all pertinent instances of a specific target class in a given dataset. The calculation involves determining the ratio of true positives to the sum of true positives and false negatives. Essentially, recall gauges the model's effectiveness in not overlooking positive instances, emphasizing sensitivity or the fraction of actual positive instances correctly identified by the model.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

4.4 F1 Score

The F1 score in machine learning is a single metric that combines both precision and recall. It is calculated as the

harmonic mean of precision and recall, providing a balanced measure of a model's performance, especially in situations with imbalanced class distribution.

4.5 ROC-AUC (Receiver Operating Characteristic - Area Under the Curve)

The AUC (Area Under the Curve) value in machine learning is a metric that quantifies the performance of a binary classification model based on its Receiver Operating Characteristic (ROC) curve. Specifically, it represents the area under the ROC curve, which illustrates the trade-off between true positive rate and false positive rate across different threshold values. A higher AUC value indicates better model discrimination, with 1.0 being perfect, 0.5 indicating random performance, and lower values suggesting poorer discrimination.

5. RESULTS AND DISCUSSION

The core objective of this research is to achieve a highly precise prediction of heart disease in patients. To realize this, the SKNDGR architecture was implemented leveraging the powerful sklearn machine learning library. This choice was driven by the accessibility of freely available inbuilt classes, facilitating seamless integration for researchers, ensuring efficient and rapid implementation and analysis. The robust training and testing procedures were executed on a Windows 10 system equipped with an Intel Core i7 9th generation CPU and 8 GB RAM, ensuring a solid foundation for the experimentation and evaluation phases.

In Table 2, we meticulously present the performance metrics of the deployed classification algorithms, elucidating crucial indicators such as accuracy, precision, recall, and the F1 score. Within the diverse array of machine learning models applied to classification, SVM and KNN exhibit commendable accuracy rates, standing at 85.48% and 87.90%, respectively. Notably, despite Decision Tree and Naïve Bayes registering lower overall accuracy, the former boasts a remarkable precision score of 91.23%, in close proximity to the precision scores of SVM (91.67%) and KNN (93.44%). Venturing into the realm of ensemble learning, Random Forest unequivocally outshines Gradient Boosting, securing an accuracy pinnacle of 98.39% coupled with flawless precision at 100%. This exceptional performance extends to the F1 score domain, where Random Forest claims the highest accolade, positioning it as the epitome of reliability within the existing literature. Our proposed SKNDGR model consistently outperforms its counterparts across accuracy, precision, recall, and the F1 score spectrum. A precision value of 1 accentuates the SKNDGR model's prowess in minimizing false positives, an imperative consideration in scenarios where precision is paramount. The elevated recall and F1 score further fortify the model's efficacy in discerning positive instances accurately. In light of these compelling results, the SKNDGR model emerges as the paragon of reliability and

robustness for the given classification task, charting a transformative course in the landscape of classification methodologies.

Table -2: Performance Metrics of different methods

Classification algorithm	Accuracy	Precision	Recall	F1 Score
SVM	0.8548	0.9167	0.8088	0.8594
KNN	0.8790	0.9344	0.8382	0.8837
Naïve Bayes	0.8145	0.8261	0.8382	0.8321
Decision Tree	0.8306	0.9123	0.7647	0.8319
Gradient Boosting	0.8145	0.8358	0.8235	0.8296
Random Forest	0.9839	1	0.9706	0.9851
SKNDGR Model	0.9919	1	0.9853	0.9926

Figure 4 vividly illustrates the comparative performance of various machine learning algorithms, shedding light on the noteworthy contributions of the proposed SKNDGR in this study. While the random forest algorithm exhibits a commendable accuracy of 98.39%, the proposed method stands out with an impressive accuracy of 99.19%. This improvement positions SKNDGR as a notable contender, showcasing its potential to enhance machine learning accuracy. The observed advancements underscore the significance of SKNDGR in contributing to the ongoing progress in the field of machine learning.

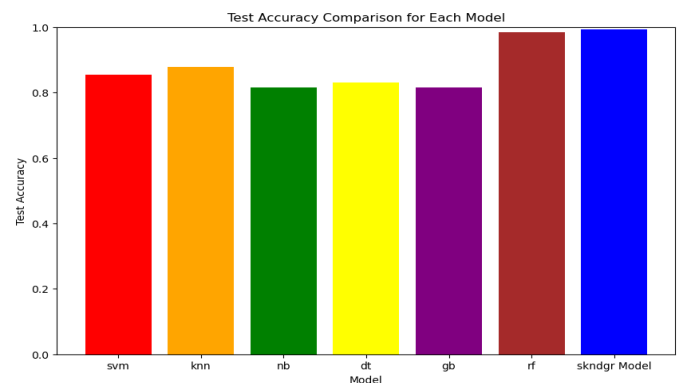


Fig -4: Test Accuracy Comparison of different methods

Figure 5 illustrates the Area Under the Curve (AUC) values for the various proposed algorithms. The ROC curve encapsulates the interplay between true positive and false

positive rates, offering a visual depiction of each model's discriminatory prowess. Notably, our novel SKNDGR model outshines all existing methods, as evidenced by its trajectory closely hugging the upper-left corner of the graph. This positioning signifies superior overall performance, underscoring the SKNDGR model's ability to achieve high true positive rates while keeping false positive rates to a minimum. The clear distinction in the AUC values further emphasizes the efficacy of the SKNDGR model, solidifying its standing as a leading solution within the evaluated algorithms.

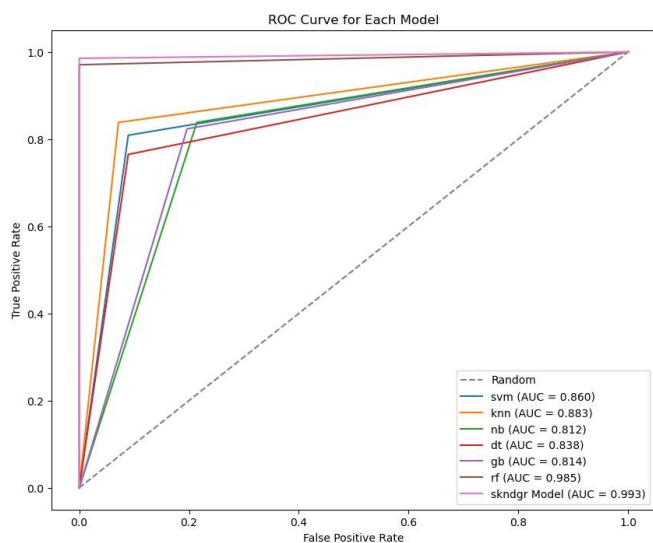


Fig -5: ROC curve of different methods

Figure 6 illustrates a confusion matrix, offering a concise summary of a classification model's performance by enumerating true positive, true negative, false positive, and false negative predictions. The proposed model demonstrates highly accurate classification for the given instances.

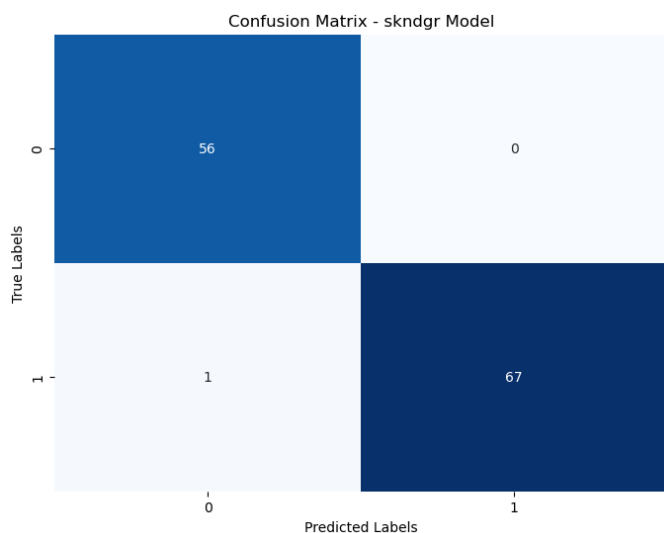


Fig -6: Confusion Matrix of proposed method

6. CONCLUSION

Heart diseases is very common among the elderly people and those exposed to a stressful environment. The early diagnosis and treatment can help avoid the patient from reaching critical or fatal conditions. The proposed machine learning approach enables quicker and accurate diagnosis of the patient's medical condition. Here, we have used the heart disease dataset to test the performance of the existing state of the art machine learning architectures. We found that the Random Forest algorithm performed the best independently from all the exiting machine learning algorithms. Our novel ensemble model combining the strength of all the individual models resulted in the best performance with an accuracy and F1 score of 99.19 % and 99.26% respectively.

7. FUTURE WORK

In the future, exploring the use of Artificial Neural Networks for handling classification problems is a promising avenue. Artificial Neural Networks (ANNs), especially deep neural networks, present distinct advantages over traditional machine learning models. They excel in capturing intricate, non-linear relationships in large datasets, automatically learning relevant features without the need for manual engineering. This makes ANNs highly effective for complex tasks and scenarios with abundant data, where traditional models might struggle. While traditional machine learning models may be more interpretable and computationally efficient in certain situations, the superior ability of ANNs to handle complexity and large-scale data makes them a compelling choice for a wide range of applications.

REFERENCES

[1] M.Snehith Raja, M.Anurag, Ch.Prachetan Reddy and NageswaraRao Sirisala, "MACHINE LEARNING BASED HEART DISEASE PREDICTION SYSTEM", in 2021 IEEE International Conference on Computer Communication and Informatics, 2021, pp. 1-5,doi: 10.1109/ICCCI50826.2021.9402653

[2] Ibomoiye Domor Mienye, Yanxia Sun and Zenghui Wang, "An Improved Ensemble Learning Approach for the Prediction of Heart Disease Risk", in Elsevier Informatics in Medicine Unlocked, vol.20, 2020, pp. 1-12, doi: https://doi.org/10.1016/j.imu.2020.100402

[3] Archana Singh and Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", in 2020 International Conference on Electrical and Electronics Engineering, 2020, pp. 1-6, doi: 10.1109/ICE348803.2020.9122958

[4] M. Nikhil Kumar, K. V. S. Koushik and K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools", in International

Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 3, 2018, pp. 1-12, doi: 10.13140/RG.2.2.28488.83203

[5] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", in IEEE Access, vol. 5, 2017, pp. 8869 - 8879, doi: 10.1109/ACCESS.2017.2694446

[6] B. Manoj Kumar and P S. Uma Priyadarsini, "Efficient Prediction of Heart Disease using SVM Classification Algorithm and Compare its Performance with Linear Regression in Terms of Accuracy", in JOURNAL OF PHARMACEUTICAL NEGATIVE RESULTS, 2022, pp. 1-8, doi: <https://doi.org/10.47750/pnr.2022.13.S04.171>

[7] Dhiraj Dahiwade, Gajanan Patle and Ektaa Meshram, "Designing Disease Prediction Model Using Machine Learning Approach", in 2019 3rd International Conference on Computing Methodologies and Communication, 2019, pp. 1-5, doi: 10.1109/ICCMC.2019.8819782

[8] K. Subhadra and Vikas B, "Neural Network Based Intelligent System for Predicting Heart Disease", in International Journal of Innovative Technology and Exploring Engineering, vol. 8, 2019, pp. 1-4

[9] C. Beulah Christalin Latha and S. Carolin Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", in Elsevier Informatics in Medicine Unlocked, vol. 16, 2019, pp. 1-17, doi: 10.1016/j.imu.2019.100203

[10] Chunyan Guo, Jiabing Zhang, Yang Liu, Yaying Xie, Zhiqiang Han and Jianshe Yu, "Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform", in IEEE Access, vol. 8, 2020, pp. 59247 - 59256, doi: 10.1109/ACCESS.2020.2981159

[11] Arpit Gupta, Ankush Shahu, Masud Ansari, Nilalohit Khadke and Ashwini Urade, "Heart disease classification using Random Forest", in International Research Journal of Engineering and Technology, vol.10, 2023, pp. 1-5

[12] Sai Bhavan Gubbala, "Heart Disease Prediction Using Machine Learning Techniques", in International Research Journal of Engineering and Technology, vol. 9, 2022, pp. 1-5

[13] Nadia Rubaiyat, Anika Islam Apsara, Abdullah Al Farabe and Ifaz Ishtiak, "Classification and prediction of Orthopedic disease based on lumber and pelvic state of patients", in 2019 IEEE International Conference on Electrical, Computer and Communication Technologies, 2019, pp. 1-4, doi: 10.1109/ICECCT.2019.8869540

[14] Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Kailas Devadkar, "Prediction of Heart Disease Using Machine Learning", in 2018 Second International Conference on Electronics, Communication and Aerospace Technology, 2018, pp. 1-4, doi: 10.1109/ICECA.2018.8474922

[15] Ching-seh Mike Wu, Mustafa Badshah and Vishwa Bhagwat, "Heart Disease Prediction Using Data Mining Techniques", in ACM Proceedings of the 2019 2nd International Conference on Data Science and Information Technology, 2019, pp. 7-11, doi: <https://doi.org/10.1145/3352411.3352413>

[16] Ambily Merlin Kuruvilla and N.V Balaji, "Heart disease prediction system using Correlation Based Feature Selection with Multilayer Perceptron approach", in IOP Conference Series, Annual International Conference on Emerging Research Areas on "COMPUTING & COMMUNICATION SYSTEMS FOR A FOURTH INDUSTRIAL REVOLUTION" (AICERA 2020) 14th-16th December 2020, Kanjirapally, India, vol. 1085, 2021, pp. 1-6, doi: 10.1088/1757-899X/1085/1/012028

[17] Devansh Shah, Samir Patel and Santosh Kumar Bharti, "Heart Disease Prediction using Machine Learning Techniques", in Springer Advances in Computational Approaches for Artificial Intelligence, Image Processing, IoT and Cloud Applications, vol. 1, 2020, pp. 1-6, doi: <https://doi.org/10.1007/s42979-020-00365-y>

[18] Harshit Jindal, Sarthak Agrawal, Rishabh Khara, Rachna Jain and Preeti Nagrath, "Heart disease prediction using machine learning algorithms", in IOP Conference Series: 1st International Conference on Computational Research and Data Analytics (ICCRDA 2020) 24th October 2020, Rajpura, India, vol. 1022, pp. 1-10, doi: 10.1088/1757-899X/1022/1/012072

[19] Mohammad Ali Hassani, Ran Tao, Marjan Kamyab and Mohammad Hadi Mohammadi, "An Approach of Predicting Heart Disease Using a Hybrid Neural Network and Decision Tree", in ACM Proceedings of the 5th International Conference on Big Data and Computing, 2020, pp. 84-89, doi: <https://doi.org/10.1145/3404687.3404704>

[20] Syed Nawaz Pasha, Dadi Ramesh, Sallauddin Mohmmad, A. Harshavardhan and Shabana, "Cardiovascular disease prediction using deep learning techniques", IOP Conference Series: International Conference on Recent Advancements in Engineering and Management (ICRAEM-2020) 9-10 October 2020, Warangal, India, vol. 981, 2020, pp. 1-6, doi: 10.1088/1757-899X/981/2/022006

[21] <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>