ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2097

Enhancing Coronary Artery Disease Detection with a Hybrid Machine Learning Approach: Integrating K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) Algorithms

Abi Izang Igyem[1]; Fatima Umar Zambuk[2]; Badamasi Imam Yau[3]; Mustapha Abdulrahman Lawal[4]
Department of Mathematical Sciences
Abubakar Tafawa Balewa University Bauchi, Nigeria

Sandra Hoommi Hoomkwap[5]
Department of Computer Sciences
University of Jos Plateau, Nigeria

Fatima Shittu[6]
Department of Computer Sciences
Federal Polytechnic, Damaturu Yobe, Nigeria

Atiku Baba Shidawa[7]
National Institute for Policy and Strategic Studies (NIPSS), Kuru Plateau, Nigeria

Ismail Zahraddeen Yakubu[8]
Department of Computing Technologies
SRM Institute of Science and Technology Kattankulathur, Chennai, India, 60320

Abstract:- Recent studies have identified coronary artery disease (CAD) as a leading cause of death globally. Early detection of CAD is crucial for reducing mortality rates. However, accurately predicting CAD poses challenges, particularly in treating patients effectively before a heart attack occurs due to the complexity of data and relationships in traditional methodologies. This research has successfully developed a machine learning model for CAD prediction by combining K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) Classifier techniques. The model, trained and tested on a dataset of 918 samples (508 with cardiac issues and 410 healthy cases), achieved an accuracy of 82% for KNN, 84.3% for SVM, and 88.7% for the hybrid model after rigorous training and testing.

***Keywords**:- Coronary Artery Disease, Machine Learning and Heart Disease.*

I. INTRODUCTION

In today's fast-paced world, people are often overwhelmed with their daily routines and responsibilities, leaving little time for self-care. This lifestyle has led to a rise in stress, anxiety, and various health conditions among individuals. Heart disease, particularly coronary artery disease (CAD), stands out as a major concern, contributing significantly to global mortality rates according to the World Health Organization (WHO). CAD occurs when the coronary arteries, responsible for supplying blood, oxygen, and nutrients to the heart muscle, become narrowed due to inflammation and cholesterol buildup [1].

The prevalence of heart-related diseases, including angina and myocardial infarction (heart attacks), underscores the importance of early detection and preventive measures. The heart plays a vital role in maintaining overall bodily functions, and any dysfunction can lead to severe health consequences. Unfortunately, many risk factors associated with heart disease, such as high cholesterol and blood pressure, often go unnoticed in the early stages, making early detection challenging[2].

In recent years, advancements in machine learning (ML) and artificial intelligence (AI) have revolutionized disease detection and prediction, offering valuable insights into risk assessment and symptom forecasting. ML algorithms, such as the Neural Network Algorithm, have shown promising results in predicting CAD with high accuracy, as demonstrated in studies using data from multiple medical repositories.

Despite these advancements, there remain challenges in effectively detecting and preventing CAD, especially in high-risk patients. Existing hybrid models, such as the one proposed by Archana et al. (2022), combine machine learning techniques like random forest and naïve Bayes. However, these models may have limitations in terms of assumptions, computational complexity, and cost.[3]

To address these challenges and improve prediction accuracy for CAD, a more robust and efficient hybrid model is proposed, leveraging the strengths of various machine learning algorithms while overcoming their limitations. This enhanced hybrid model aims to enhance early detection, improve risk assessment, and ultimately reduce fatalities associated with coronary artery disease [4].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2097

The remaining part of this work includes section 2 related work section 3 method, section 4 findings and section 5 concludes the study.

II. RELATED WORK

The most prevalent kind of heart disease, coronary artery disease, develops gradually and frequently goes undetected until a heart attack strikes. Over the past few decades, CAD has been identified as one of the top causes of death globally (Dhar et al., 2018). The probability of death can be minimized with early detection of CAD. Artificial intelligence and machine learning are widely acknowledged to play an important role in the medical field for diagnosing the disease and classifying or predicting the outcomes. Research has been conducted using machine learning technology to identify heart disease from historical medical data to uncover correlations in data. Multiple studies have reported on the use of various algorithms to foresee heart issues. The table above demonstrates the need for additional study in heart failure, even though several researchers have used machine learning techniques. In the work of (Archana, K.S. et al., 2022), they recommended a hybrid machine learning prediction system that foresees the risk of rising heart disease.

Multiple approaches to machine learning have been used to accurately predict or identify various forms of cardiac disease. K-means and Artificial Neural Networks were used in a hybrid technique to increase accuracy, identify, and extract the unknown information of heart illness in the prediction of heart disease by [8]. These detected cardiac problems with a 97% accuracy rate. In order to improve the accuracy of coronary prediction in 2021, [9] implemented an AI technique to find relevant traits in a hybrid random forest linear model approach to predict heart disease. The model's accuracy in predicting heart disease was 88.7%. (Aravind, A. et al., 2021) developed predictive models utilizing various machine learning algorithms (Generalized linear model, Decision tree, Random forest, Support vector machine, neural network, and k-nearest neighbor) to aid clinicians in the early detection of coronary artery disease, with neural networks achieving the highest accuracy of 93%. [11] used machine learning to detect heart disease using historical medical records in order to find correlations in the data which greatly increase the correctness of prediction rates. The classifier techniques they employed, Modified Naive Bayes and Random Forests, yielded a 92% accuracy.

The majority of the corresponding literatures that were reviewed used supervised learning to recognize Coronary Disease; for this reason, this research will employ an unsupervised learning technique to address the topic at hand.

III. METHODOLOGY

The proposed system's architecture is presented in figure 1 below with six components which includes: Dataset, Data preprocessing, Feature selection, Train/Test data, hybrid algorithm and heart problem/absence of heart problem.

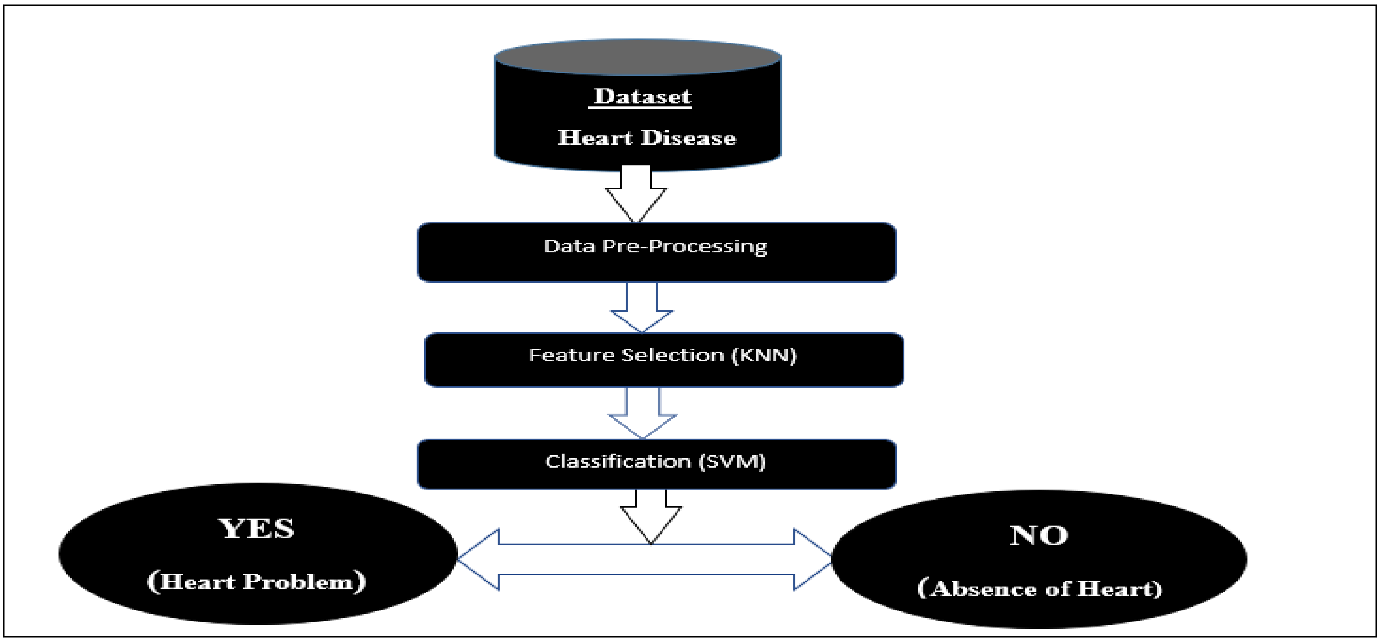

Fig 1: Architecture of the Proposed System

A. Advantages of the KNN Algorithm

It is a simple algorithm with a quick processing time. It suitable for both classification and regression problems KNN has high accurac8y; there is no need to compare to more effective supervised learning models. Lastly, no assumptions about data – no need to make additional assumptions, tune several parameters, or build a model.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2097

B. Advantages of the SVM

The SVM classifiers are excellent in the High Dimensional Space and have good spatial accuracy. Less memory is needed for SVM classifiers because they only consume a minimal quantity of training data. SVM functions admirably when there is a significant variation between the classes. Use of high density spaces is preferable for SVM. When the dimensions exceed the sample size, SVM is advantageous. The memory is well utilized by SVMs.

C. Dataset

This dataset was developed in September 2021 by integrating various datasets—Cleveland, Hungarian, Switzerland, Long Beach and Stalog—that were previously available separately. The final heart disease dataset consists of 918 samples with 410 instances for the class of people who are healthy and 508 cases for the class of people who have heart problems.

- Database: Number of instances:
- Cleveland: 303
- Hungarian: 294
- Switzerland: 123
- Long Beach VA: 200
- Stalog (Heart) Data Set: 270
- Total 1190

➢ *The dataset used is made up of 11 clinical features:*

- The patient's age,
- Sex,
- Type of chest pain (typical angina, atypical angina, non-anginal pain or asymptomatic),
- The resting blood pressure mmHg,
- The serum cholesterol (mm/dl),
- The fasting blood sugar (value 1 if FastingBS > 120 mg/dl, and value 0 otherwise),
- Resting electrocardiogram results (which can be Normal, ST if the patient has ST-T abnormalities or LVH if the patient shows probable ventricular hypertrophy),
- Mthe maximum heart rate (numeric value between 60 and 202),
- Exercise-induced angina which can be yes or no,
- The oldpeak (numeric value measured in depression) and finally, the slope of the peak exercise ST segment (Up, Flat, Down).
- The column number 12 contains the output class which can be 1 (heart disease) or 0 (normal).

D. Data Pre-Processing:

To clean up and extract more valuable data from the dataset, a preprocessing step will be performed. The age will be removed, and three new columns indicating different ages—young, adult, and elder—will be introduced in its place. The characteristic of sitting BP will also be changed into three new columns for low BP, medium BP, and high BP. Finally, the cholesterol feature will be changed into three separated columns that indicate how high, medium, and low the risk is. ChestPainType, RestingECG, and ST Slope are three features that will each be encoded using a single hot encoding technique. We have a dataset with 24 features at the end of this procedure. To eliminate inconsistency, a k-fold cross-validation will be performed using 10 folds for each trial.

E. Data Classification (KNN and SVM)

In the approach shown in figure 3.1 above, the entire training dataset is used as an input, and a process called feature augmentation is used to enhance the number of features. In order to determine whether the sample belongs to a positive class or not, the classification model uses this new dataset, which has more features than the previous one.

F. Choice of the Simulation Environment

Using WEKA 3.8 (Waikato Environment for Knowledge Analysis), this study's experimental analysis will be carried out. According to Hall et al. (2009), Waikato University in New Zealand created WEKA, an open source machine learning program, in Java.

G. Choice of metrics

➢ *Below are the Metrics we've Chosen:*

- Precision: This metric calculates the ratio of 'True Positives' to the sum of 'True Positives' and 'False Positives.' Put simply, it measures the accuracy of positive predictions.

➢ *The Mathematical Formula is:*

$$TP\ /\ (TP\ +\ FP)\ ...eqn(i)$$

- **Recall:** This is defined as the ratio of 'True Positives' to the sum of 'True Positive' and 'False Negative', it the fraction of positives that were correctly defined.

➢ *The Mathematical Formula is:*

$$TP\ /\ (TP\ +\ FN)\ ...eqn(ii)$$

- F1-Score: It is the value of weighted mean of 'Precision' and 'Recall'. This score would address the question of 'What percent of positive predictions were right?

➢ *The Mathematical Formula is:*

$$2 * (Recall * Precision)/(Recall\ +\ Precision)\ ...eqn(iii)$$

- **Accuracy:** is a percentage of accurately categorized data elements over the entire data occurrences.

$$\frac{TN + TP}{TN + FP + TP + FN} ...eqn(iv)$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2097

IV. EXPERIMENTAL SETUP AND RESULTS

This section presents the results and discussions of the proposed approach in predicting coronary artery disease using machine learning algorithms (KNN & SVM). It compares the performance of the classifiers with the existing system. The dataset used in this research work was developed in September 2021 by integrating various datasets of Cleveland, Hungarian, Switzerland, Long Beach and Stalog, which available separately. The final heart disease dataset consists of 918 samples with 410 instances for the class of healthy people and 508 cases for the class of people who have heart problems.

Table 1: Composition of Dataset

Database	Number of Instances
Cleveland	303
Hungarian	294
Switzerland	123
Long Beach VA	200
Stalog (Heart) Data Set	270
Total	**1190**

The dataset utilized comprises 11 clinical attributes: patient age, gender, chest pain type (typical angina, atypical angina, non-anginal pain, or asymptomatic), resting blood pressure (mmHg), serum cholesterol (mm/dl), fasting blood sugar (1 if FastingBS > 120 mg/dl, 0 otherwise), resting electrocardiogram results (Normal, ST-T abnormalities, or probable ventricular hypertrophy), maximum heart rate (numeric value between 60 and 202), exercise-induced angina (yes or no), old peak (numeric value measured in depression), and slope of the peak exercise ST segment (Up, Flat, Down). The 12th column represents the output class, which can be 1 (indicating heart disease) or 0 (representing normal). The experimentation was conducted using WEKA 3.9.6, an open-source machine learning scripting software.

The Figure 2 below shows the Home Page of the software with different features like Explorer, Experimenter, Knowledge Flow, Workbench and Simple CLI. Figure 2 shows the initial process of data training by loading the dataset in the WEKA machine learning environment after and loading then training the dataset; which consists of 918 samples with 410 instances for the class of healthy people and 508 cases for the class of people who have heart problems.

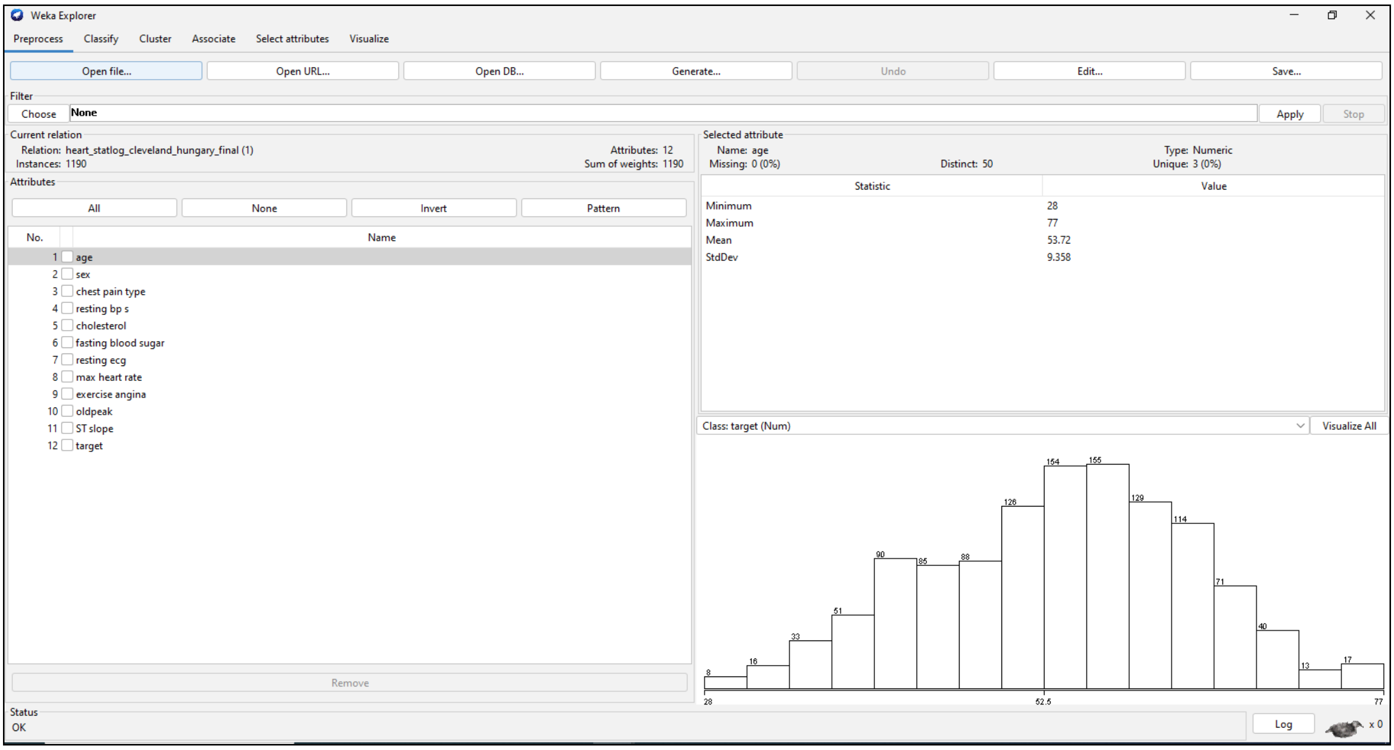

Fig 2: Loading the Dataset

A. *Output Results of the Proposed System (KNN & SVM)*

➢ *K-Nearest Neighbor Classifier*

Figure 3 shows the output of the KNN classifier with 75% correctly classified instances and 25% incorrectly classified instances.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2097

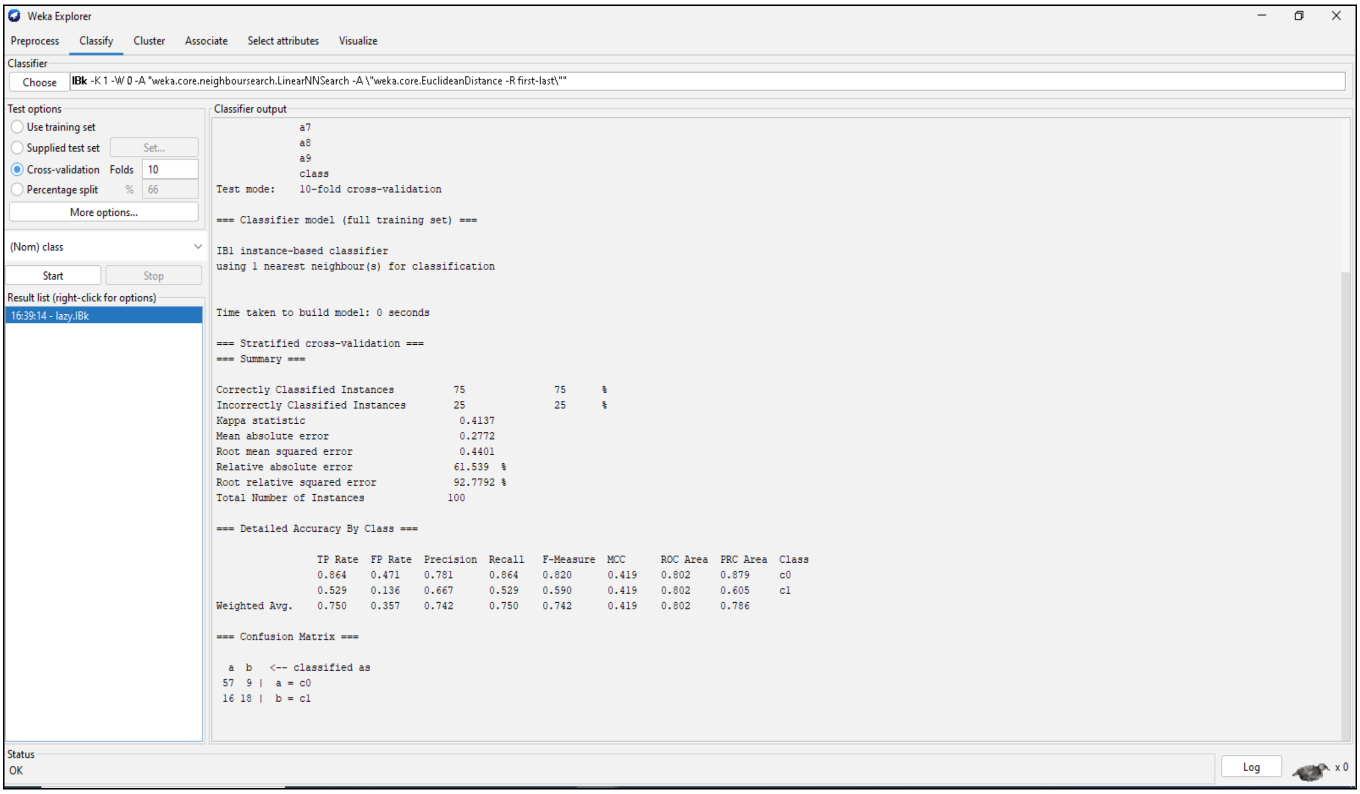

Fig 3: Output of KNN Classifier

B. Support Vector Machine Classifier

Figure 4 below shows the output of the KNN classifier with 74% correctly classified instances and 26% incorrectly classified instances.

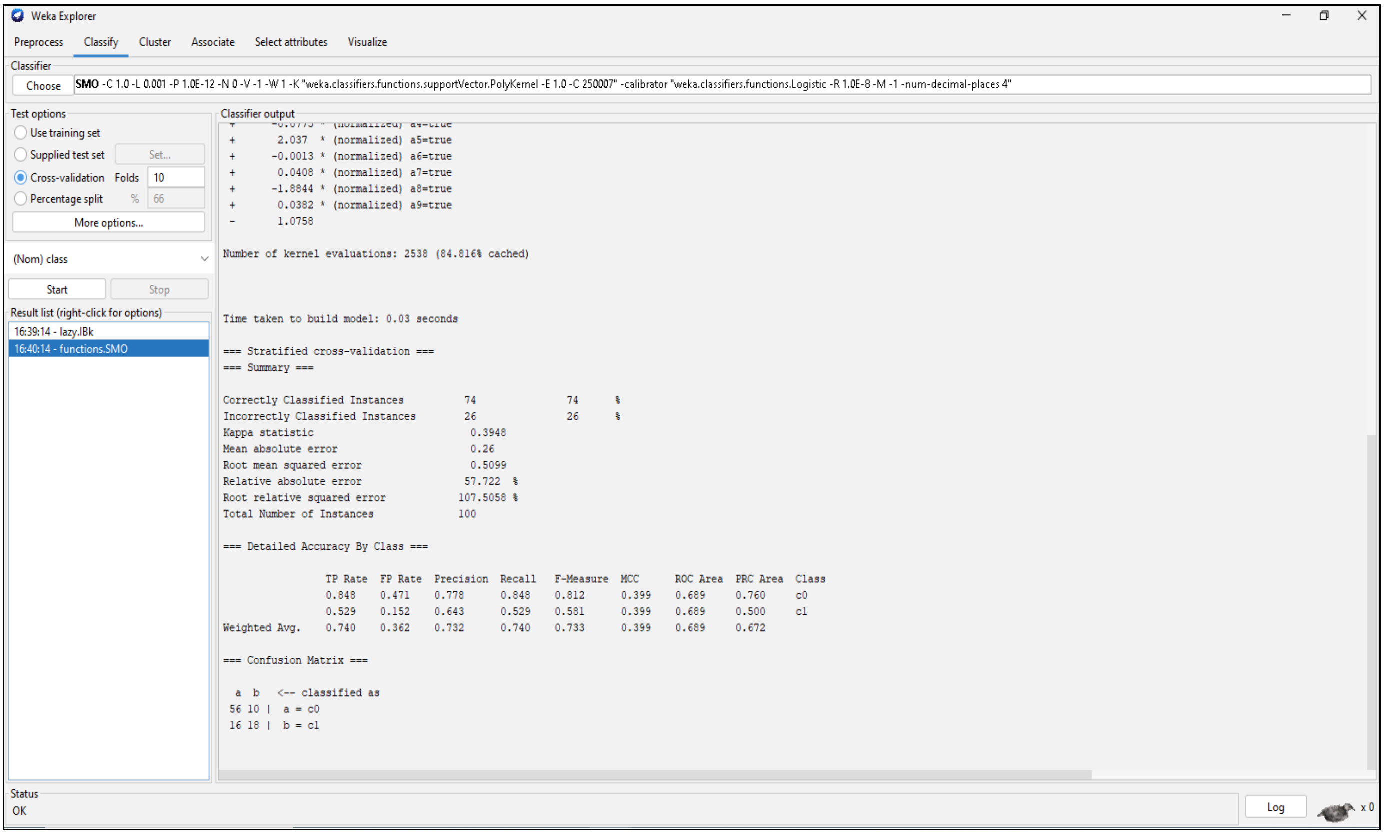

Fig 4: Output of SVM Classifier

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2097

From Figure 4 above, the two machine learning algorithms performed a classification task and it shows that the KNN algorithm has the highest precision in classifying the patients with coronary heart disease in the class label in the experiment.

Also, the results show that KNN algorithm has the highest detection accuracy (Recall). Finally, KNN classifier outperforms the other classifier in carrying out F-Measure in the experiment.

C. Performance Evaluation

A Coronary Artery Detection system (CAD) is assessed based on accuracy, detection rate, and F-measure. Precision indicates the proportion of correctly identified patients with coronary artery disease. It's calculated using the following formula:

Accuracy (Acc) is a widely used metric for classification performance, representing the ratio of correctly classified samples to the total number of samples. It's expressed as:

$$\text{Accuracy}=\frac{TP}{TP+FP} \quad \text{……………………………………….(1)}$$

Detection Rate or Recall is described as the number of attacks detected by the proposed technique to the total number of attacks truly there (Modi & Jain, 2016).

$$\text{Detection Rate (Recall)} =\frac{TP}{TP+FN} \quad \text{……………………….(2)}$$

$$\text{F-Measure} = \frac{2*Precision*Recall}{Precision+Recall} \quad \text{……………………….(3)}$$

- **True Positive (TP):** this is the number of patients with coronary artery disease that were correctly classified.
- **True Negative (TN):** this is the number of patients without coronary artery disease that were correctly classified.
- **False Positive (FP):** this is the number of patients with coronary artery disease that were incorrectly classified as normal.
- **False Negative (FN):** this is the number patients without coronary artery disease that were incorrectly classified.

Table 2: Percentage of Weighted Average of the Two Classifiers (Proposed system)

	KNN (%)	SVM (%)
Precision	78.1	77.8
Recall	86.4	84.8
F-Measure	82.0	81.2

Table 2 describes the percentages of the weighted average of the machine learning classifiers that were used to perform the experiment with a recall value of 86.4 and 84.8, precision value of 78.1 and 77.8 and F-Measure of 82 and 81.2 for KNN and SVM respectively.

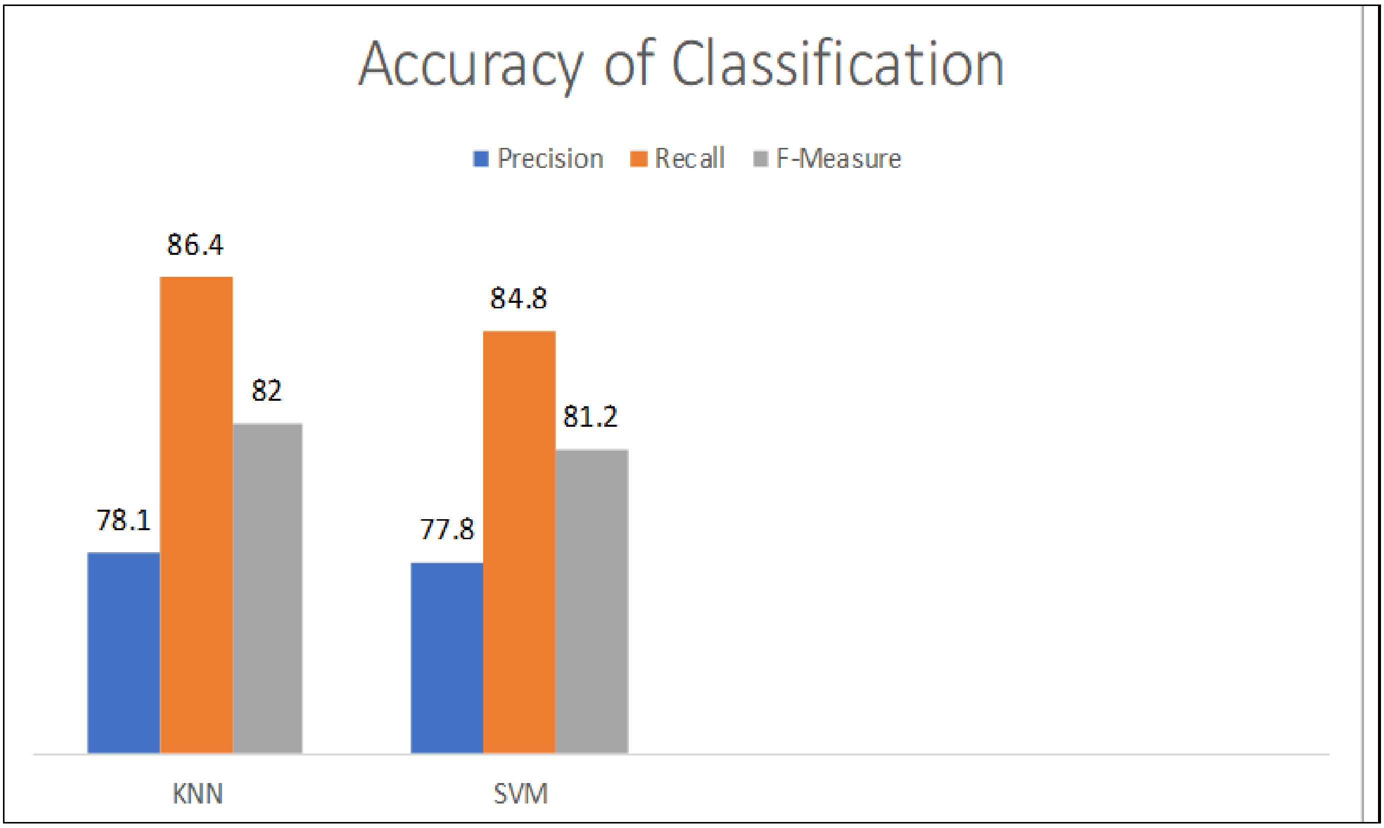

Fig 5: Graph of Weighted Average

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2097

The graph in Figure 5 is generated from Table 2 the X-axis denotes the percentage of performance while the Y-axis represents the Machine Learning Classifiers. The graph represents the percentage of the performance of the two (2) classifiers. The comparison shows that KNN algorithm outperforms the SVM algorithm in the level of Precision, Recall and F-measure.

D. *Comparison of the Existing System and the Proposed System*

In this research work, the existing system developed by Archana, et al. (2022) was implemented and the outputs are given below:

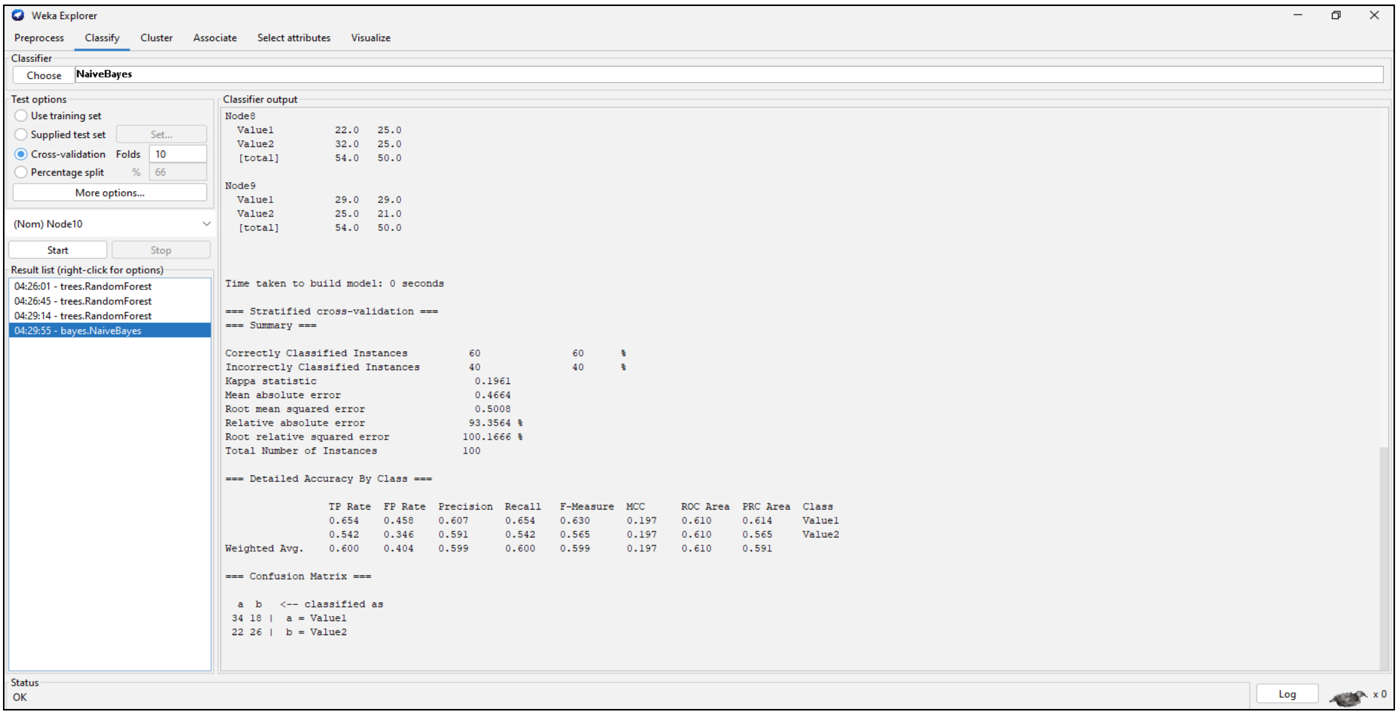

Fig 6: Output of Naïve Bayes (Existing System)

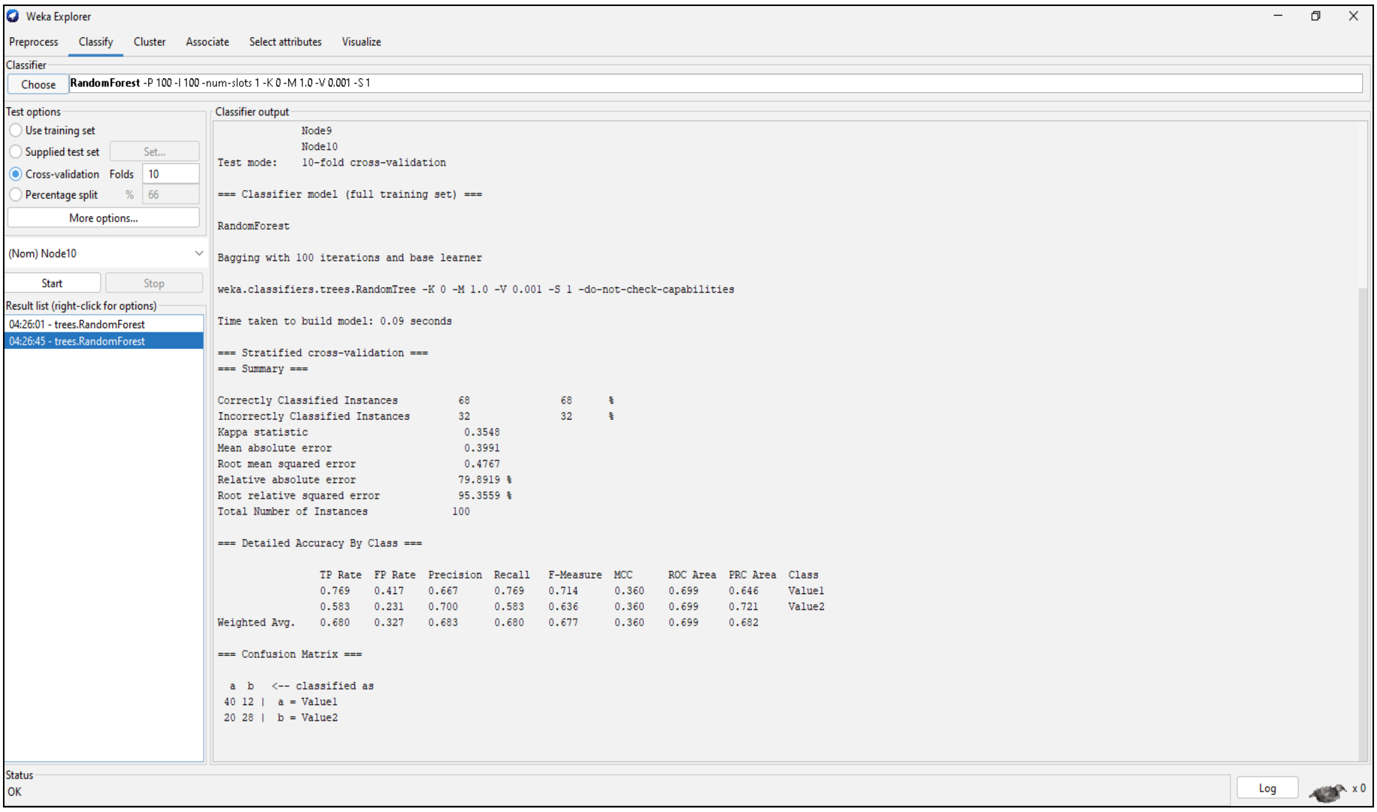

Fig 7: Output of Random Forest Classifier (Existing System)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2097

Table 3 gives the Precision, Recall and F-measure of the existing system on the two classifiers used (Naïve Bayes and Random Forest).

Table 3: Percentage of Weighted Average of the Two Classifiers (Existing System)

	Naïve Bayes (%)	**Random Forest (%)**
Precision	60.7	66.7
Recall	65.4	76.9
F-Measure	63.0	71.4

The Naïve Bayes algorithm has a percentage of 60% of correctly classified instances and 40% of incorrectly classified instances while Random Forest has a percentage of 68% of correctly classified instances and 32% of incorrect classified instances.

Table 4: Comparison of the Algorithms in the New System Against those in the Existing System

	New System		**Existing System (Archana, et al., 2022)**	
	KNN (%)	**SVM (%)**	**Naïve Bayes (%)**	**Random Tree (%)**
Precision	78.1	77.8	60.7	66.7
Recall	86.4	84.8	65.4	76.9
F-Measure	82.0	81.2	63.0	71.4

The Precision, Recall and F-measure of the new system outperform that of the existing system as shown in Table 4.

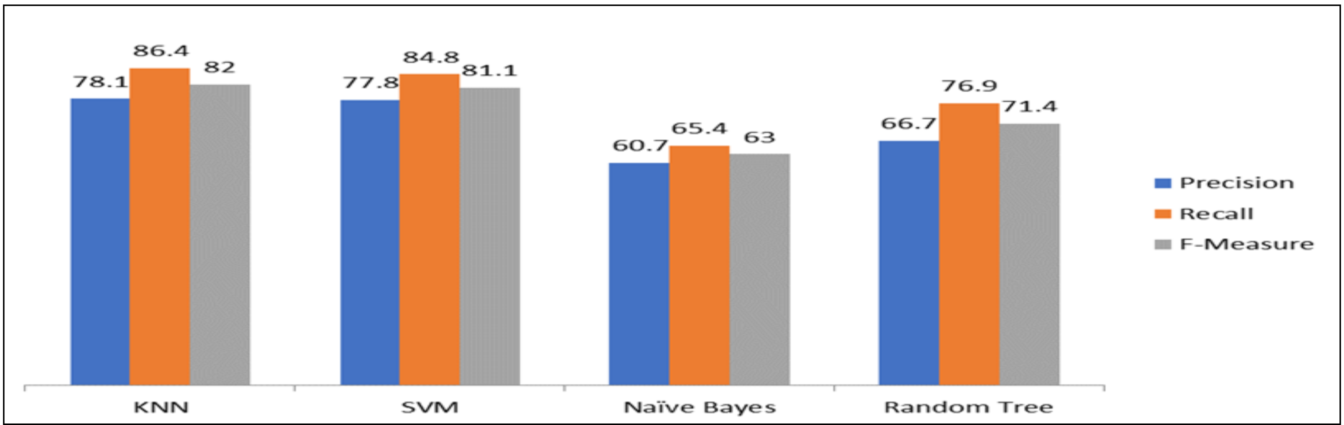

Fig 8: Graph Representing the New and Existing System

With the comparisons of both the existing and new systems in Figure 7, it is clear that the machine learning model used outperforms that of the existing system. It will be observed that precision, recall and the F-measure of the new system has more percentage compared to that of the existing system. Figure 7 shows a bar chart that shows the classifiers of both the proposed system (KNN & SVM) and the existing system (Naïve Bayes & Random Tree), the figure shows how the performance of the proposed system outperforms the existing system.

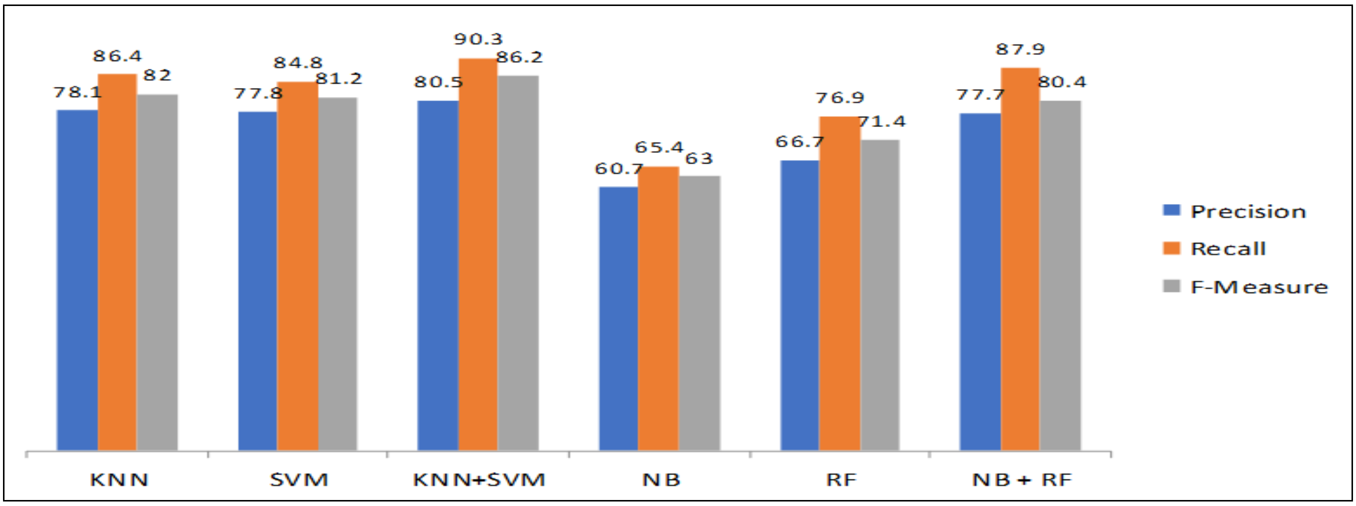

Fig 9: Hybridization of the Existing and Proposed Models Based on the Machine Learning Algorithm

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2097

With the above comparisons of the single algorithms, existing and the proposed hybrid algorithms in figure 8, it is obvious that the machine learning model of the proposed hybrid algorithms outperforms that of the single algorithms and the existing system. It will be observed that precision, recall and the F-measure of the proposed hybrid algorithms has more percentage compared to that of the existing system and the single algorithms. Figure 9 shows a bar chart that shows the classifiers of proposed hybrid algorithms, single algorithms (KNN & SVM) and the existing system (Naïve Bayes & Random Tree), it shows how the performance of the proposed hybrid algorithms outperforms the single algorithm and the existing system.

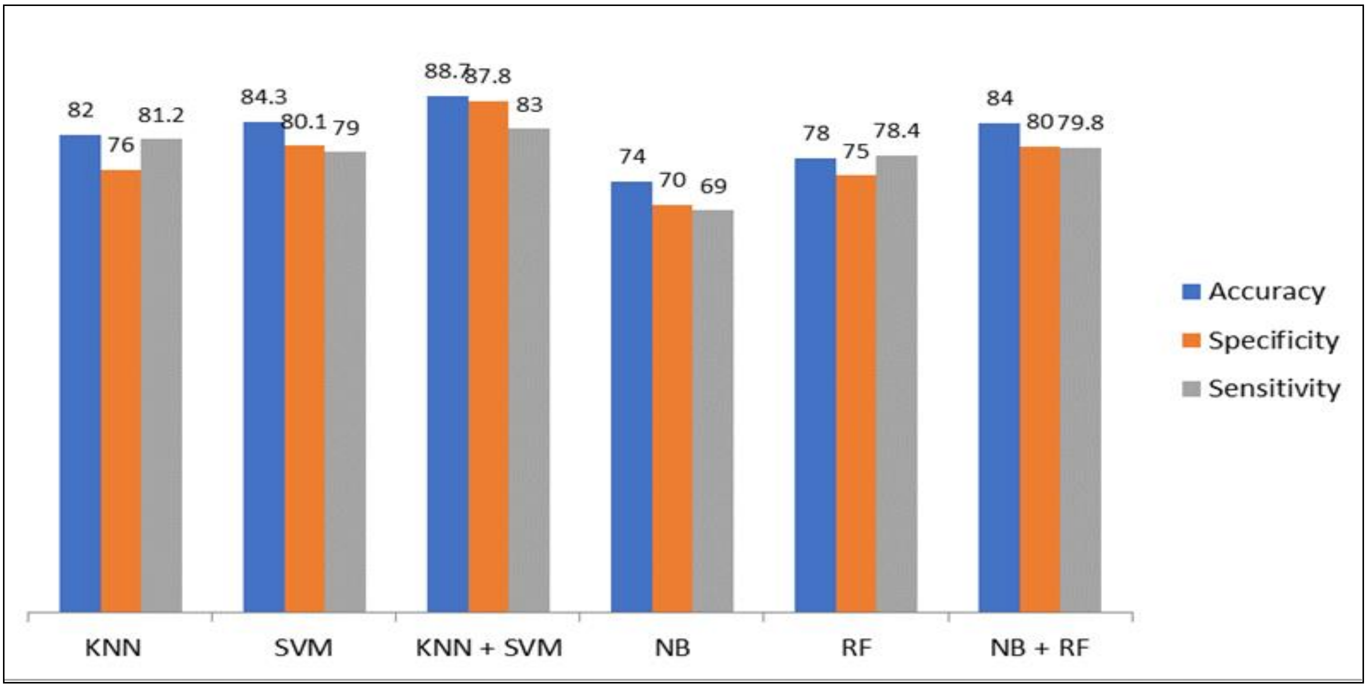

Fig 10: Graph Comparing the Accuracy, Specificity and Sensitivity of Proposed Hybrid System and Existing Hybrid System

V. CONCLUSION

According to the study's findings, machine learning has the potential to completely transform the healthcare industry. In the past, diagnosing diseases depended on routine processes and medical assessment, which was frequently limited and resulted in high costs. On the other hand, machine learning models represent an incredible breakthrough in healthcare diagnostics toward improved, scalable, and affordable approaches by providing a cost-effective method of diagnosing illnesses through the use of large datasets. Given the life-threatening nature of coronary artery disease and its broad impact on millions of people globally, the importance of early prediction in this condition cannot be stressed (Asadi et al., 2021).

This the successfully built a machine learning model for CAD prediction using the hybridized K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) Classifier techniques. The model was trained and tested using a dataset of 918 samples, which included 508 cases of people with cardiac problems and 410 cases of people in good health. After extensive training and testing, an accuracy of 82% and 84.3% for KNN and SVM respectively and 88.7% was attained for the hybrid model.

ACKNOWLEDGMENT

We wish to appreciate the expert support of our supervisors Dr. F. U. Zambuk and Dr. B. I Ya'u towards the success of this research.

REFERENCES

[1]. Ramalingam, V. Dandapath, V, A, & Karthik Raja M. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology. 2018;7(2.8):684–687. https ://doi.org/10.14419 /ijet.v7i2.8.10557

[2]. Mayo Clinic (2022). Coronary artery disease https://www.mayoclinic.org/diseases-conditions/coronary-artery- disease/symptoms-causes/syc-20350613

[3]. Swathy, M. & Saruladha, K. (2021). A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques, ICT Express (2021), doi: https://doi.org/10.1016/j.icte.2021.08.021

[4]. Mohan, S. Thirumalai, C. & Srivastava, G.(2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. Access IEEE. 2019;7:81542–54.

[5]. Fatima, M. & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9:1–16. https ://doi.org/10.4236/jilsa .2017.91001. Foresee Medical (2020) Benefits of Machine Learning in healthcare https://www.foreseemed.com/blog/machine-learning-in-healthcare

[6]. Weng, S.F. Reps, J. Kai, J. Garibaldi, J.M. & Qureshi, N. (2017). Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944

[7]. Awad, M. & Khanna, R. (2015). Support Vector Machines for Classification. https://www.researchgate.net/publication/300723807

[8]. Aditya, M. Prince, K. Himanshu, Arya, & Pankaj, K. (2015). EarlyHeart Disease Prediction Using Data Mining Techniques", CCSEIT, DMDB, ICBB, MoWiN, AIAP – 2014, pp. 53–59, DOI: 10.5121/csit.2014.4807

[9]. Animesh, H. Subrata, K. Mandal, A. Gupta, Arkomita Mukherjee & Asmita Mukherjee (2017). Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Advances in Computational Sciences and Technology, ISSN 0973-6107 Volume 10, Number 7 (2017) pp. 2137-2159

[10]. K. S. Archana, B. Sivakumar, Ramya Kuppusamy , Yuvaraja Teekaraman , & Arun Radhakrishnan (2021). Automated Cardioailment Identification and Prevention by Hybrid Machine Learning Models. *Computational and Mathematical Methods in Medicine* Volume 2022, Article ID 9797844, 8 pages https://doi.org/10.1155/2022/9797844