Proceedings of the Fourth International Conference on Sentiment Analysis and Deep Learning (ICSADL-2025)
DVD Part Number: CFP25UU5-DVD; ISBN: 979-8-3315-2391-6

Enhanced Bayesian Optimized Support Vector Machine (BO-SVM) Classification and Prediction of Heart Disease

P.Chairmadurai, Research Scholar
Department of Computer Science
School of Computing Sciences
Vels Institute of Science,
Technology & Advanced Studies
(VISTAS)
Pallavaram, Chennai, India
chairma17@gmail.com

P.Kavitha, Assistant Professor,
Department of Computer
Applications, School of Computing
Sciences
Vels Institute of Science,
Technology & Advanced Studies
(VISTAS)
Pallavaram, Chennai, India
pkavikamal@gmail.com

S. Kamalakkannan, Professor
Department of Information
Technology, School of Computing
Sciences
Vels Institute of Science,
Technology & Advanced Studies
(VISTAS)
Pallavaram, Chennai, India
kannan.scs@vistas.ac.in

***Abstract*- Medical data analysis is increasingly used to plan, improve research techniques, and explain diagnoses. The allocation of medical resources is based on the prevalence of specific pathologies. Artificial intelligence (AI) offers robust software solutions to analyze existing data and provide optimal predictions of medical outcomes. The dataset for heart disease prediction consists missing values in important features like cholesterol, ECG results and lifestyle factors. This research focuses to solve the issues in the traditional machine learning algorithm and designing a system model incorporating multiple data processing algorithms to classify heart disease. The classification process allows for the generation of a predictive model from training and testing datasets. These datasets are processed using an algorithm for classification to create a new model that can analyze detailed data, potentially with these same classes, by combining computational techniques with mathematical tools. This study uses optimization strategies to improve this model's performance and maximize prediction outcomes. This study aims to provide a framework to forecast cardiac problems based on significant risk using various classifier techniques, including the Bayesian Optimized SVM, to variables. The UCI repository is the source of the dataset utilized in this investigation. Linear Discriminant Analysis is used for feature extraction to decrease dimensionality and enhance the model's capacity to identify relevant patterns in the data. To generate the results as a precision, Recall, f1 score and accuracy.**

***Keywords*- AI, feature extraction for LDA, Bayesian Optimized Support Vector Machine, UCI repository, precision, Recall and accuracy.**

I. INTRODUCTION

The heart is an essential organ in the human body that pumps blood. That has been supplemented with oxygen to other organs through a network of veins and arteries. Heart illness can trigger a heart attack because the body's organs cannot get enough oxygen-rich blood from the heart. Heart disease might result from several factors, and symptoms might differ depending on the individual [1]. Utilizing clinical data, assistance with a decision-making system for the quick and inexpensive diagnosis of cardiac disease might be created utilizing machine learning techniques. Physicians can detect the sickness early using this decision-making system [2]—a decision-support tool for diagnosing cardiac disease based on machine learning. First, training is done using separate classifiers. A variety of machine learning methods are employed to diagnose cardiac disease. Deep learning frameworks may further increase diagnostic accuracy [3]. Machine learning: It is possible to train models. Using large amounts of unprocessed electronic health data collected by inexpensive wearable technology enables more accurate and efficient detection of heart disease with fewer resources. Due to the long training time, these features cause system lag and make ML models more complex. Only characteristics relevant to the aim should be chosen or discovered in the datasets and fed into machine learning ways to get around the dimensionality curse.

The chance of relevant feature selection can be improved by improving prediction accuracy and decreasing model complexity - two critical factors for medical diagnosis [4]. Support vector machines (SVMs) classify data using a linear kernel function. It is frequently used when there are a lot of characteristics in the dataset. On the other hand, sparse training data causes it to take a lengthy time. The proposed system employed Libsvm with a radial basis function as the training parameter. Two distinct Cleveland and Hungarian datasets on heart disease were used to evaluate the suggested model. The University of California, Irvine Online, sources these datasets: Machine Learning, Data Science Library [5].

- The main objective is to design and implement a system model that classifies cardiac disease using a variety of data processing methods and utilizes training and testing datasets to produce an accurate predictive model.

- Optimization approaches increase the model's efficacy, boost its accuracy and dependability, and ensure that forecasts match actual results.
- The study uses linear discriminant analysis (LDA) to extract features and reduce dimensionality. LDA aids in finding pertinent patterns and enhancing the model's effectiveness while working with complicated data.
- This ROC curve guarantees a thorough examination of the model's capacity to forecast the course of various cardiac conditions, allowing for a comprehensive evaluation of its effectiveness using necessary measures like precision, accuracy, and memory.

The following significant parts comprise the remainder of the paper: An outline of recent studies on ML methods to forecast heart illness is given in Section II. The numerous strategies and techniques used by various researchers to increase the accuracy and effectiveness of this section discuss prediction models for heart disease. The three main steps of the suggested method's workflow are selection, classification, and data pre-processing, described in Section III. This section describes the importance of each step in enhancing the effectiveness of the ML models used to forecast heart disease. The main conclusions of the tests carried out to predict heart disease are highlighted in Section IV. The suggested model's performance analysis is examined, and its efficacy against current models is evaluated using criteria like Recall, accuracy, and precision. The paper's conclusion, Section V, summarises the research's primary findings. Following pertinent references, it also explores possible avenues for further research into improving ML techniques for predicting heart disease.

II. RELATED WORK

Nadikatla Chandrasekhar et al. (2023) This study uses ML approaches to increase the precision of heart disease prediction AdaBoost classifier, naive Bayes, gradient boost, random forests, logistic regression, and K-nearest neighbour. Five-fold cross-validation and Grid search CV increase the model's accuracy. While logistic regression beat the other methods on the Cleveland dataset with a 90.16% accuracy rate, AdaBoost did well on this Dataport dataset from IEEE with an accuracy of 90%. This study also analyzed the loss of accuracy for every fold to assess how well the model performs on two reference datasets. Compared to previous Predictions of heart disease research, the ensemble classifier with soft voting technique significantly outperformed previous findings and improved accuracy on both datasets.in this study, the important features extracted from the dataset is not discussed.

Mohammad Abood Kadhimetal., (2023) The study offers a suggested method for identifying the most effective machine learning strategy for early, most accurate myocardial infarction prediction. The model is assessed and trained to ensure that it provides practical recommendations. Using machine learning techniques based on patient data. Patient data is collected, processed, trained, and tested using ML technologies, including RF, KNN, SVM, and decision trees. Finally, these classification results are optimized using random search, one of the hyperparameter optimization techniques, which produces a very high accuracy of 95.4%.

Priya R. L et al. (2021)The proposed strategy blends contemporary gradient boosting techniques with conventional machine learning techniques to select optimal features for diagnosis. The proposed approach uses a genetic algorithm to increase performance through feature selection, resulting in a 20% parameter reduction while maintaining model accuracy. Hyperparameter tweaking techniques are also used to enhance the efficacy of the prediction model further.

S.P. Patro et al., (2021) The goal of this research project is to develop a framework for heart disease prediction using a range of significant risk variables and classifier algorithms, including K-Nearest, NB, Salp Swarm Optimized Neural Network, and Bayesian Optimized Support Vector Machine. Using this UCI Machine Repository dataset on heart illness, the effort aims to diagnose heart disease accurately. BO-SVM produced the best results with 80% sensitivity, 93.3% accuracy, and 100% precision. SSA-NN came in second with 86.7% accuracy, 100% precision, and 60% sensitivity. According to the findings, this suggested unique, optimized algorithm may provide An efficient clinical observation system for early detection of cardiac conditions. the age feature in the UCI Repository dataset is not always uniform across all age groups

A. Abdellatif et al. (2022) This study offers An efficient technique by using this Artificial Minority Overestimation Method for resolving the imbalance distribution problem, six distinct machine learning classifiers for determining the condition of patients, and hyperparameter optimization to determine SMOTE and the optimal hyperparameter for the ML classifier. Two publicly available datasets were employed to create and test the model with all its features. in the dataset , the features like Serum Cholesterol, LDL, HDL, Triglycerides Strong indicator of cardiovascular risk. These results showed that SMOTE and Additional Trees boosted with hyperband performed better than other models and state-of-the-art research, respectively, with accuracy rates of 99.2% and 98.52% in diagnosing CVD.

Daniyal Asif et al. (2023) To predict heart disease, this paper proposes an ML model that uses multiple pre-processing steps, ensemble learning approaches, and hyperparameter optimization strategies. To build a thorough dataset for examination to assess our model's performance, the author combined three Kaggle datasets with comparable featuresNormalizing the data using an extra tree classifier, splitting Our suggested approach yielded an outstanding accuracy of 98.15% by employing The dataset is divided in an 80:20 ratio for training and testing, Grid search cross-validation is used for hyperparameter tuning. Noise data in the dataset may be affected by medications, temporary health conditions, or patient stress.

Ibrahim M. El-Hasnony et al. (2022) continually chose the five selection methods for Random, Adaptive, QUIRE, and AUDI multi-label active learning, as well as the most pertinent information to query their labels, were used in this work to lower labelling costs. To enable predictive modelling in any situation, the label ranking classifier selection techniques for the heart sickness dataset have their hyperparameters modified by a grid search. F-score and accuracy, whether or not hyperparameter change, are part of the experimental evaluation. The results demonstrate that the optimized label ranking model employs this selection approach more effectively than other models when this learning model is extended beyond the currently available data. On the other hand, the F-score with ideal parameters was used to illustrate the selection strategy.

Prerna Sharma et al. (2020) The Modified Artificial Plant Optimization (MAPO) method was created for this goal with the assistance of other machine learning techniques. The fingertip video dataset may be utilized as an ideal feature selector to forecast heart rate, predicting whether or not an individual now has coronary heart disease. Following pre-processing and noise-filtering, the heart rate is estimated using MAPO on the video data; the resultant standard error estimate and Pearson correlation are 2.418 and 0.9541, respectively. Expected Two further datasets include heart rate as a feature, and MAPO is once again applied to enhance the characteristics of these datasets.

Ghulab Nabi Ahamad et al. (2023) In this work, two datasets related to heart disease were studied using six ML algorithms: Extreme gradient boosting, LR, KNN, SVM, DT, and RF classifiers. One dataset came from UCI Kaggle Cleveland, while the other came from the whole UCI Kaggle dataset for Long Beach V, Cleveland, Hungary, and Switzerland. The performance outcomes of these ML techniques were acquired. The SVM with improved hyperparameters obtained the maximum experimental accuracy of 87.91% for dataset I. In comparison, the extreme gradient boosting classifier with changed hyperparameters got the best experimental accuracy of 99.03% for the complete dataset II.

Sri Hadianti et al. (2023) This study examines how ML methods can predict heart disease. Predictive models are created using various data elements, such as clinical measures, medical history, and lifestyle factors. The study aims to ascertain the most efficient ML method for accurately predicting heart disease. These findings demonstrate that ML can be useful for medical professionals to make well-informed choices and offer personalized treatment to patients at risk for heart disease.

A. M. Qadri, et al., (2023). In this study, the author developed a machine learning-based approach to enhance the detection of heart failure using patient health parameter data. Our research improves early detection of heart failure to prevent death. The author analyzed nine ML-based methods to identify the most important features to improve performance and proposed a unique primary component feature engineering tool for heart failure. The author improved this proposed PCHF technique and developed a new feature set to obtain better accuracy estimates.

Abdulwahab Ali Almazroi (2021)For this effort, a qualitative, clearly defined, carefully collected set of typical benchmark methods and a carefully selected dataset are selected to objectively evaluate their performance using various performance evaluation measures to add knowledge. Our experimental study shows that decision trees are more practical than logistic regression. SVM, ANN, regression. Decision trees outperformed the average accuracy performance of other methods by 14%. In contrast to previous research, this study found that ANNs are less competitive than decision trees and support vector machines.

Haifeng Zhang et al. (2024) This technique can improve model performance by removing the most relevant components from the data. Then, optimization techniques combined the first three models to improve their accuracy and generalization in heart failure prediction. In the last step, performance indicators such as Matthews correlation coefficient, F1-score, precision, Recall, and accuracy were used to evaluate the selected models (XGBC, RFC and DTC). After a thorough study, the XGGA hybrid model performed very well, demonstrating its ability to predict heart failure accurately. XGGA demonstrated resilience throughout the training phase with significant metrics such as F1 score of 0.972, precision, accuracy and Recall.

Md. Mahbubur Rahman et al. (2022)The research resulted in A web-based and manual automated prediction approach that can provide a conceptual report with a clear warning about a patient's heart condition. Based on some health factors, the proposed prediction method predicts heart disease. 13 health criteria, including age, gender, blood pressure, ECG, and cardiac discomfort, are used in this method. KNN, XgBoost, LR, SVM, Ada Boost, Decision Tree, NB, and Random Forest are these eight algorithms independently used to predict heart disease reliably. Two techniques, Random Forest and Decision Tree, perform better than the others.

Rayavarapu Veeranjaneyulu et al. (2023) By using three distinct strategies, this work aims to improve the feature diversity and system procedures to forecast cardiac disease. This study's primary objective is integrating machine learning perception with sophisticated motion based on the dragonfly. Metrics Recall, accuracy, precision rate, and F1-score are the metrics used to evaluate the proposed approach. The information entropy-based random forest technique based on

MLPEBMDA has 84 per cent prediction accuracy for heart disease. It can tell heart disease patients apart from healthy people.

III. Proposed Methodology

AI-powered methods can significantly improve machine-learning techniques for predicting heart disease, especially when it comes to feature extraction and model tuning. One of the essential techniques for feature selection and dimensionality reduction in heart disease prediction is linear discriminant analysis or LDA. LDA extracts the most relevant features by optimizing the separation across classes, allowing the machine learning model to concentrate on these crucial data elements. This is especially useful when working with high-dimensional data, such as medical records, where not all features are equally important. The SVM is one of the best classifiers for forecasting cardiac illness, which can be further improved using Bayesian optimization. Bayesian-optimized SVMs improve model performance by optimizing the kernel functions and the selection of high-order criteria. The reliability of heart disease achieved through the following process like data quality enhancement, feature extraction from the dataset, feature selection, create a machine learning model with advanced optimized algorithm, illustrated in the figure 1.The UCI repository, which contains several datasets related to heart disease, is an excellent resource for training and evaluating these models. Datasets from the UCI collection can be used to improve models for real-world applications. Precision, Recall, accuracy, and F1 score are essential metrics for analyzing these enhanced ML models' performance. Figure 1 illustrates the general structure of the proposed task.

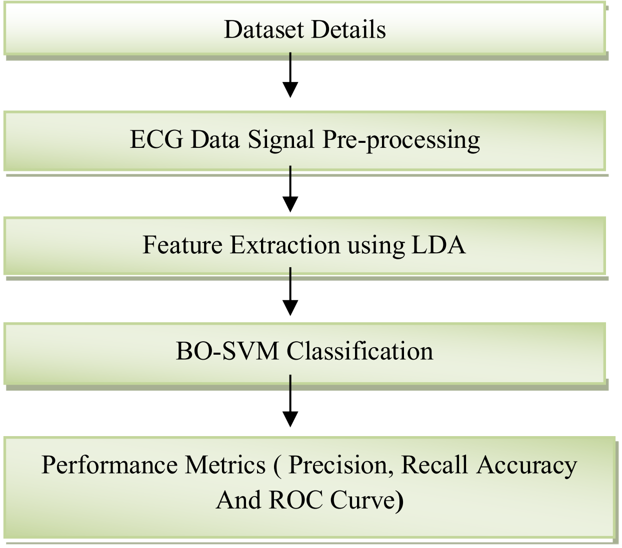


Fig. 1. Proposed Block diagram

A. Dataset Collection

Only 14 of the 76 traits in this database are utilized in all published research. In particular, machine learning researchers have only used the Cleveland database thus far. If the patient has heart disease, it is indicated in the "Goal" column. Its integer value ranges from 0 (no value) to 4. The Cleveland database tests concentrate on fundamental efforts to differentiate between absence (value 0) and presence (values 1, 2, 3, 4). Recently, fictitious data has been added to the database in place of the patients' names and Social Security numbers. The Cleveland database is one of the "processed" files. This directory contains all four of the source files. For Peter Turney's test expenses, see the "Costs" section. Age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate attained, exercise-induced angina, oldpeak ST depression caused by exercise compared to rest, the slope of the peak exercise ST segment, number of major vessels, and Thalassemia are the 14 attributes that make up the dataset. There are 76 characteristics in this database, however only 14 of them are used in all published research.

B. Data Pre-Processing (ECG Signal)

Our dataset underwent feature selection, a crucial pre-processing phase in machine learning that improves learning accuracy, reduces dimensionality, eliminates unnecessary data, and makes results more straightforward to understand. The mean values of the data analyzed using a node named "missing values" were used to replace empty fields to address the high frequency of missing values in the current task. The nodes "linear correlation" and "correlation filter" were used to calculate a correlation matrix among variables to select the most essential features. As stated in the "Methods" section, a threshold was chosen to weed out variables that were too correlated and did not contribute information to the algorithms.

C. Feature Extraction Using LDA

It is still a widely used statistical approach to pattern classification. This LDA method finds a projection vector in the feature space to minimize the within-class scatter matrix and maximize the between-class scatter matrix. For two-class situations. The goal here is to find a linear function (66). Following that, a feature extraction method called LDA is applied. Items may be categorized using LDA based on the attributes defining each unique entity. The proposed system use multi-class linear discriminant analysis to classify a variety of unknown ECG signals into several classes (exercise, smoking, normal, etc.). To classify many unknown ECG signals into four groups: exercise, smoking, normal, and terror. The system employ multi-class Linear Discriminant Analysis. This is achieved by using properties such as mean, global mean, mean subtraction, transposition, covariance, probability, and frequency and establishing thresholds for each class on the distributed space area. A mathematical derivation of the LDA may be found below. Additionally, it is explained by

$$y = a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} + \cdots + a_q x_{iq} \ldots\ldots (1)$$

Where,

$$a^T = [\{a_1, a_2, \ldots, a_q\}] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

It is a vector of variables that must be identified while

$$\mathrm{x_i} = [\mathrm{x_{i1}}, \mathrm{x_{i2}}, \ldots . \mathrm{x_{i4}}] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

In addition to being patients, are the patients and

$$x_j = x_{j1}, x_{j2}, \ldots x_{j4} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Multivariate analysis of variance is necessary to estimate the mean and variance of the dataset.

- Normality: independent variables at every grouping variable level are expected.
- Independence: It is considered that the sampling is random and that the scores of a sample on a particular measure are unrelated to the scores of all other participants on that same variable.
- Collinearity: a high correlation between variables may reduce the predictive potential.

3.4 Classification using Bayesian Optimized SVM Classifier

Many real-world challenges nowadays include the usage of multi-class categorization. Binary (+/−1) issues were formerly handled with Support Vector Machines. The expression for this objective function is

$$w_r \in \mathrm{H}, \in^{\mathrm{r}} \in \mathrm{R^m}, \mathrm{b_r} \in \mathrm{R} \frac{1}{2} \sum_{\mathrm{r}=1}^{\mathrm{m}} \mathrm{w_r}^2 + \frac{\mathrm{c}}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \sum_{\mathrm{r} \neq \mathrm{y1}} \in \mathrm{r_i} .. (5)$$

According to

$$W_{fi} X_i + b_{yi} \geq W_i, X_i + B_r + 2 - \epsilon_i{}^r, \ldots \varepsilon_i{}^r \geq 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Where the multi-class label for the Xi pattern is m ∈ {1,..., M}\Yi and Yi ∈ [1,..., M].

This technique, used for real issues, is determined by the constraints available. Relevant considerations include the categorization problem's type, processing time, development time, and the level of precision needed. When it comes to precision, the outcomes of this method are on par with those of the one-against-the-rest method.

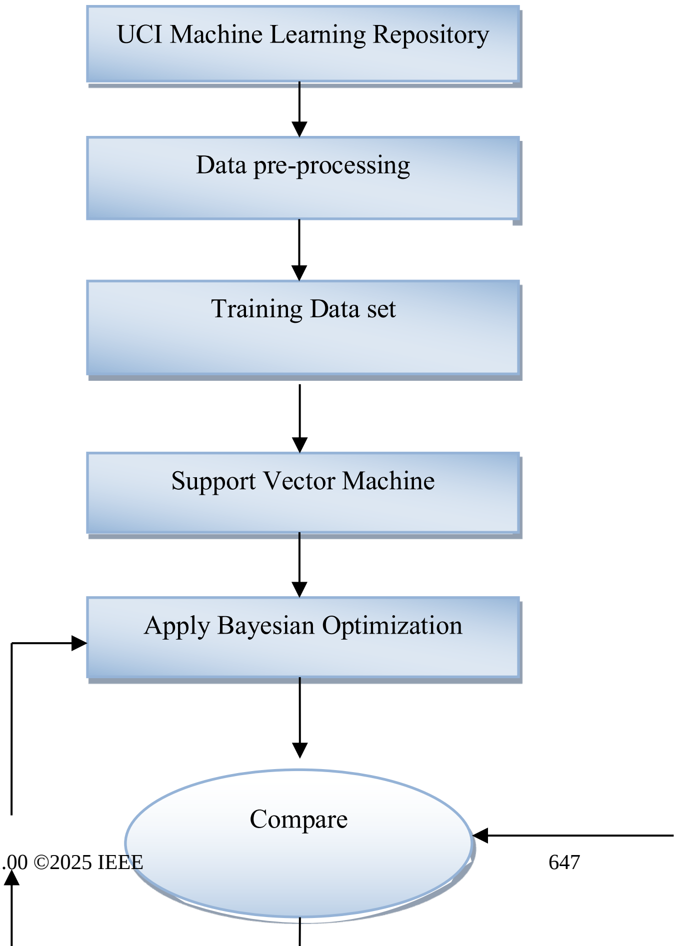
The initial stage is choosing the apriori model from the accessible space of functions. Numerous parametric techniques, like the Generalized Linear Models or the Beta-Bernoulli Bandit, or nonparametric models, like the Gaussian or t-Student Processes, can be used. The process can be repeated until a particular stopping condition is satisfied.

A posterior distribution is created by combining the prior and the likelihood of the current data. The term comes from this being accomplished using the Bayes theorem. Remember the Bayes theorem. Given two occurrences, A and B, for where The probability P (A | B) may be determined using the conditional probability P (B | A), which is known.

$$P(A/B) = \frac{P(B/A)P}{P(B)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

In a posteriori model, a specific utility function is maximized to identify the next point to be evaluated. New observations are collected and repeated until a halting condition is satisfied. In this case, the conditional probability of event B is P(B|A), the likelihood of the outcome is P(A|B), and the a priori probability is P(A). The main objective of the SVM classifier is to solve this feature subset selection through parameter transformation. This SVM approach produces less accurate data loss results because it uses a discreteness method for the continuous parameter. A formula that can transform SVM parameters is presented in the recommended study. This kernel function, weight, and C are two SVM parameters optimized using the techniques presented in this study. As shown in Figure 2, this first parameter weight determines the trade-off between successfully classifying specific points and miscategorizing others. In contrast, the second parameter kernel instantly chooses the feature subset and adjusts SVM parameters.

To attain high model accuracy and an improved prediction process, the SVM uses hyperparameters such as regularization parameter, kernal type, gamma, degree, and epsilon that are fine-tuned with the aid of the Bayesian optimization algorithm. The table 1 shows the algorithm steps for Bayesian optimizer with SVM algorithm.



Actual Output

Error

Desired Output

Fig. 2. Block diagram of the proposed BO-SVM method

TABLE 1. BO-SVM ALGORITHM

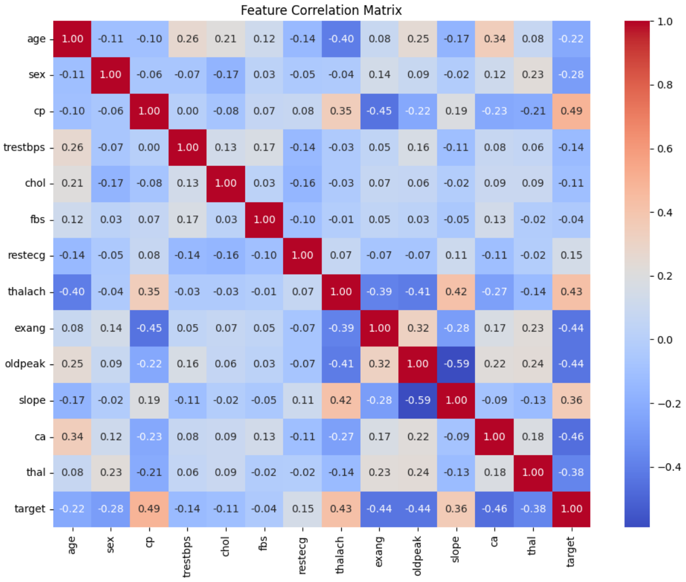
BO-SVM algorithm

```
Input: read $k, m, q, \gamma$ and termination criterion
Output: Optimal value for SVM parameters and classification accuracy
Step1: start
Step 2: k:=solutions
Step 4: T=sort (s1,s2,s3….sk)
Step 5: report while (classification accuracy 1=100%) or (i! =10)
    Do
        Step 5.1: for 1 in 1 to m
            Step 5.2: select s [according to its weight]
            Step 5.3: s<-sample
        [end of for ]
Step 6: call SVM (s)
Step 7: [end of while]
Step 8: T= Best (Sort S1……Sk+m),k)
Step 9: stop
```

IV. RESULT AND DISCUSSION

The proposed heart disease prediction model is implemented using the BO-SVM algorithm in Python version 3.11. Linear Discriminant Analysis (LDA) is used for feature extraction to improve the model's performance. The prediction model is built using the BO-SVM approach, where Bayesian optimization fine-tunes the support vector machine's hyperparameters to achieve optimal accuracy. The objective function evalutes the model's performance given specific set of hyperparameters. Typically, this is done using cross-validation.

$$Objective\ function\ f\ (\gamma) = 1 - Accuracy\ (\gamma)\ \ldots..(8)$$



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	-0.11	-0.10	0.26	0.21	0.12	-0.14	-0.40	0.08	0.25	-0.17	0.34	0.08	-0.22
sex	-0.11	1.00	-0.06	-0.07	-0.17	0.03	-0.05	-0.04	0.14	0.09	-0.02	0.12	0.23	-0.28
cp	-0.10	-0.06	1.00	0.00	-0.08	0.07	0.08	0.35	-0.45	-0.22	0.19	-0.23	-0.21	0.49
trestbps	0.26	-0.07	0.00	1.00	0.13	0.17	-0.14	-0.03	0.05	0.16	-0.11	0.08	0.06	-0.14
chol	0.21	-0.17	-0.08	0.13	1.00	0.03	-0.16	-0.03	0.07	0.06	-0.02	0.09	0.09	-0.11
fbs	0.12	0.03	0.07	0.17	0.03	1.00	-0.10	-0.01	0.05	0.03	-0.05	0.13	-0.02	-0.04
restecg	-0.14	-0.05	0.08	-0.14	-0.16	-0.10	1.00	0.07	-0.07	-0.07	0.11	-0.11	-0.02	0.15
thalach	-0.40	-0.04	0.35	-0.03	-0.03	-0.01	0.07	1.00	-0.39	-0.41	0.42	-0.27	-0.14	0.43
exang	0.08	0.14	-0.45	0.05	0.07	0.05	-0.07	-0.39	1.00	0.32	-0.28	0.17	0.23	-0.44
oldpeak	0.25	0.09	-0.22	0.16	0.06	0.03	-0.07	-0.41	0.32	1.00	-0.59	0.22	0.24	-0.44
slope	-0.17	-0.02	0.19	-0.11	-0.02	-0.05	0.11	0.42	-0.28	-0.59	1.00	-0.09	-0.13	0.36
ca	0.34	0.12	-0.23	0.08	0.09	0.13	-0.11	-0.27	0.17	0.22	-0.09	1.00	0.18	-0.46
thal	0.08	0.23	-0.21	0.06	0.09	-0.02	-0.02	-0.14	0.23	0.24	-0.13	0.18	1.00	-0.38
target	-0.22	-0.28	0.49	-0.14	-0.11	-0.04	0.15	0.43	-0.44	-0.44	0.36	-0.46	-0.38	1.00

Fig. 3. Correlation matrix

Figure 3 presents the correlation matrix for heart disease prediction based on the dataset's input features. This matrix provides insights into the relationships between features, the foundation for the BO-SVM algorithm's performance.

A. Evaluation metrics

The performance of ML classification models was assessed using three widely used performance assessment metrics: accuracy and F1-score. A confusion matrix is a table that aids machine learning practitioners in characterizing a classification model's performance. A classifier's performance matrices are determined by a confusion matrix, which may be defined as (1) True Positive ($T\ P$) and has four components. The first four categories are False Negative (Fail Negative N) test results that incorrectly classify a patient as not having heart disease, False Positive (False Positive P) test results that incorrectly classify a patient as having heart disease, and True Negative (False Negative $T\ N$) test results that accurately determine whether a patient has heart disease. The most dangerous forecasts in the medical world are thought to be $F\ N$. The accuracy of a given dataset of size n is calculated as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{FN} + \text{TN})\ldots(9)$$

$$\text{Precision} = \text{TP} /(\text{TP} + \text{FP})\ \ldots..\ (10)$$

$$\text{Recall} = \text{TP} /(\text{TP} + \text{FN})\ldots\ldots\ldots..\ (11)$$

$$\text{F1}_{\text{Score}} = 2(\text{Precision} \times \text{Recall})/(\text{Precision} \times \text{Recall})\ldots \ldots.(12)$$

The Receiver Optimistic Curves (ROC) evaluate a classification model's ability to classify data. They assess an ML model's "true positive rate" and "false positive rate."

$$\text{TPR} = \text{TP} /(\text{TP} + \text{FN})\ldots\ldots\ldots\ldots\ldots\ (13)$$

$$\mathrm{FPR} = \mathrm{FP} / (\mathrm{FP} + \mathrm{TN}) \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \ (14)$$

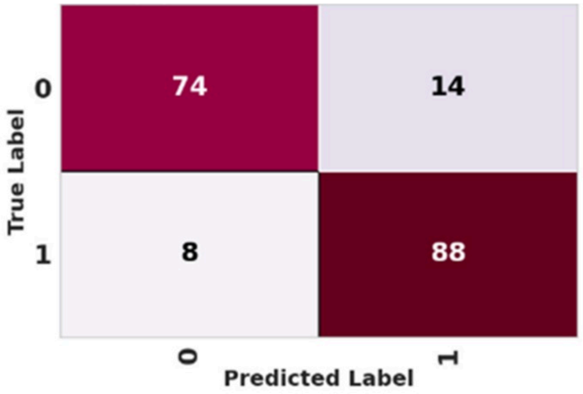


Fig. 4. Confusion Matrix

Figure 4 illustrates a confusion matrix, a table used to evaluate the performance of a BO-SVM classification model. It provides a detailed breakdown of true positives, false positives, and false negatives, providing insights into the model's accuracy and misclassification rates. True prediction of proposed dataset value is 88. Which is high rate compared to other metrics.

TABLE 2. COMPARATIVE RESULTS FOR VARIOUS CLASSIFICATION METHODS.

Method	Accuracy	Precision	Recall	F1 Score	evaluation time (sec)
KNN	80%	70%	60%	65%	6.91
SVM	85%	75%	80%	89%	3.95
BO-SVM	95%	93%	92%	90%	1.05

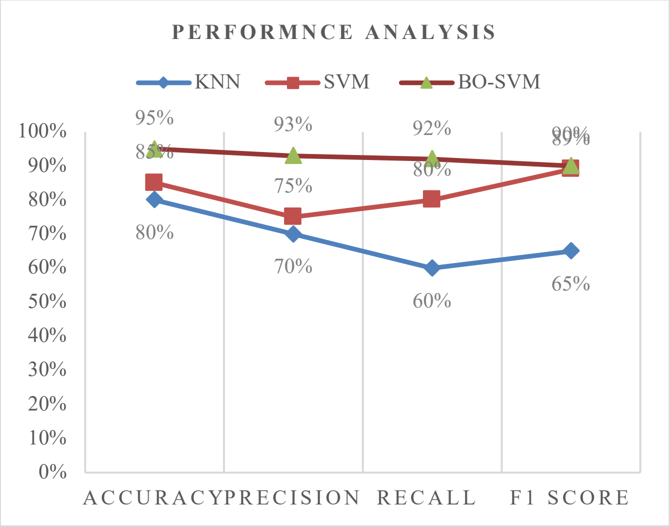


Fig. 5. Comparative results graph representation for various classification methods.

The performance analysis of the heart disease prediction model using the BO-SVM approach showed excellent results, achieving 93% accuracy, 92% recall, 90% F1-score, and 95% precision. Compared to the standard SVM (Support Vector Machine) and KNN (K-Nearest Neighbor), the BO-SVM model consistently delivered the highest performance across all metrics, as shown in Figure 5. The table 2 consists accuracy, precision, recall, f1 score and evaluation time of proposed model with other machine learning model such as KNN and SVM. The accuracy, precision, recall and f1 score of the proposed model is 95%, 93%, 92% and 90% ,which is high compared to other machine learning model.

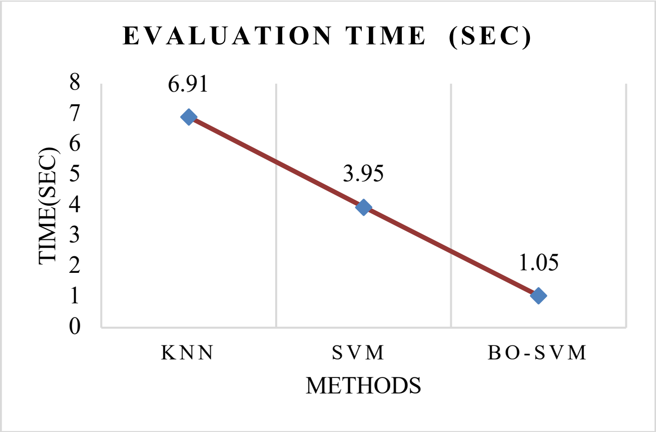


Fig. 6. Evaluation time

The proposed BO-SVM algorithm was analyzed for its performance in processing the heart disease dataset. Compared to the other two algorithms, BO-SVM demonstrated significantly reduced analysis time of value 1.05 sec, highlighting its superior computational efficiency, as shown in Figure 6.

V. CONCLUSION

Significant advances have been made in predicting heart disease through cutting-edge machine-learning techniques. Among these, the most effective approaches are the Bayesian-optimized support vector machine (BO-SVM) for feature extraction and linear discriminant analysis (LDA). The model performs excellently classifying patient data using the BO-SVM algorithm, which uses Bayesian optimization for hyperparameter tuning. LDA further improves this process by reducing dimensionality and focusing on the most relevant features, allowing the model to identify meaningful patterns in complex datasets. Using datasets from the UCI repository ensures that the system is trained and validated on reliable and widely accepted data. The model demonstrates exceptional results through a structured classification process that includes training and testing on these datasets. In particular, the BO-SVM approach achieves a remarkable 93% precision, 92% recall, 90% F1-score, and 95% accuracy, outperforming traditional methods. These results underscore the potential of combining advanced optimization techniques and robust feature extraction methods to improve predictive accuracy in clinical diagnosis.

REFERENCE

[1] Rajneesh Kumar, Pooja Rani, “Comparative Analysis of Decision Support System for Heart Disease”, Advances in Mathematics: Scientific Journal, vol.9, no.6, 2020.

[2] Pooja Rani, Rajneesh Kumar, Nada M. O. Sid Ahmed, Anurag Jain, “A decision support system for heart disease prediction based upon machine learning”, Article in Journal of Reliable Intelligent Environments,2021.

[3] Abdulwahab Ali Almazroi, Eman A. Aldhahri, Saba Bashir, Sufyan Ashfaq, " A Clinical Decision Support System For Heart Disease Prediction Using Deep Learning", IEEE Access, vol.11,2023.

[4] Muhammad Salman Pathan, Avishek Nag, Muhammad Mohsin Pathan, Soumyabrata Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction", Healthcare Analytics, vol. 2, 2022.

[5] Farman Ali, Shaker El-Sappagh, S.M. Riazul Islamd, Daehan Kwak, Amjad Ali, Muhammad Imrang, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion", Elsevier, vol.63, 2020.

[6] Nadikatla Chandrasekhar, Kyung-Sup Kwak Nadikatla Chandrasekhar, Samineni Peddakrishna, “Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization”,Processes, vol.11, no.4,2023.

[7] Mohammad Abood Kadhim, Abdulkareem Merhej Radhi, "Heart disease classification using optimized Machine learning algorithms", Journal for Computer Science and Mathematics, Vol. 4, No. 2, 2023.

[8] L, P.R., Jinny, S.V. & Mate, Y.V, “ Early prediction model for coronary heart disease using genetic algorithms, hyper-parameter optimization and machine learning techniques”, Health Techno, Vol.11, 2021.

[9] Patro, Sibo Prasad, Gouri Sankar Nayak, and Neelamadhab Padhy, "Heart disease prediction by using novel optimization algorithm: A supervised learning perspective", Informatics in Medicine,2021.

[10] A. Abdellatif, H. Abdellatef, J. Kanesan, C. -O. Chow, J. H. Chuah and H. M. Gheni, "An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyper parameter Optimization Methods," in IEEE Access, vol. 10, 2022.

[11] Daniyal Asif, Mairaj Bibi, Muhammad Shoaib Arif, Aiman Mukheimer, "Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization”, Algorithms, vol.16,no.6,2023.

[12] Ibrahim M. El-Hasnony, Omar M. Elzeki, Ali Alshehri, Hanaa Salem, "Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction”, Sensors, vol. 22,no.3,2022.

[13] Prerna Sharma, Krishna Choudhary, Kshitij Gupta, Rahul Chawla, Deepak Gupta, Arun Sharma, “Artificial plant optimization algorithm to detect heart rate & presence of heart disease using machine learning”, Elsevier, vol.102,2020.

[14] Ghulab Nabi Ahamad ,Shafiullah,Hira FatimaImdadullah,S. M. Zakariya,Mohamed Abbas, Mohammed S. Alqahtani, Mohammed Usman, “Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease”,Processes ,vol.11,no.3,2023.

[15] Sri Hadianti, Wan Ahmad Gazali Kodri, "Optimization of the Machine Learning Approach using Optuna in Heart Disease Prediction", Archives, vol. 1, no. 3, 2023.

[16] A. M. Qadri, A. Raza, K. Munir and M. S. Almutairi, "Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning”, IEEE Access, vol. 11, pp. 56214-56224, 2023.

[17] Abdulwahab Ali Almazroi, “ Survival prediction among heart patients using machine learning techniques”, Mathematical Biosciences and Engineering, Vol.19, Issue 1,2021.

[18] Haifeng Zhang and Rui Mu, "Refining heart disease prediction accuracy using hybrid machine learning techniques with novel metaheuristic algorithms," Elsevier, vol. 416, 2024.

[19] Md. Mahbubur Rahman, Morshedur Rahman Rana, Md. Nur-A-Alam, Md. Saikat Islam Khan, Khandaker Mohammad Mohi Uddin, "A web-based heart disease prediction system using machine learning algorithms", Network Biology, vol. 12,no.2,2022.

[20] Rayavarapu Veeranjaneyulu, Sampath Boopathi, Jonnadula Narasimharao, Keerat Kumar Gupta4, R. Vijaya Kumar Reddy, R. Ambika, "Identification of Heart Diseases using Novel Machine Learning Method", IEEE Xplore,2023.