Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 32s (2025)
https://www.jneonatalsurg.com

A Versatile Approach for Improving Heart Disease Prediction Accuracy Via the LSKR Soft Voting Ensemble Model with Firefly Optimization- LSKR-SVE(FO)

Mr. Michael Raj.S[1], Dr. M.Mohamed Sirajudeen[2]

[1]Research Scholar, PG and Research Department of Computer Science, Nilgiri College of Arts and Science, Thaloor, The Nilgiri's.

[2]Associate Prof &Head, PG and Research Department of Computer Science, Nilgiri College of Arts and Science, Thaloor, The Nilgiri's.

Email ID: mdsirajudeen1@gmail.com

.Cite this paper as: Mr. Michael Raj.S, Dr. M.Mohamed Sirajudeen, (2025) A Versatile Approach for Improving Heart Disease Prediction Accuracy Via the LSKR Soft Voting Ensemble Model with Firefly Optimization- LSKR-SVE(FO). *Journal of Neonatal Surgery*, 14 (32s), 688-697.

ABSTRACT

In recent times, heart disease has emerged as a significant global health concern. The utilization of machine learning, deep learning, and other artificial intelligence (AI) tools to aid in medical diagnostics is steadily gaining traction.This work presents a unique LSKR soft voting ensemble model with Firefly optimization (LSKR-SVE(FO)), which consists of four different learners, to improve the prediction accuracy of heart disease. Here, PrincipalComponent Analysis and Lasso method have been utilized in feature extraction and feature selection respectively for enhancing the prediction accuracy of LSKR-SVE(FO). The weight value of SVE and overfitting problem can be minimized by utilizing Firefly optimization algorithm. This study investigates efficient heart disease diagnosis using the heart disease dataset from the UCI Machine Repository. Proposed LSKR-SVE(FO) achieved the highest performance (99.3% accuracy, 98.45% precision, 96.2% Recall), followed by SVM-AHP (96.3% accuracy, 98.5% precision, 88.3% recall). These findings suggest that our optimized algorithm offers an effective healthcare monitoring system for early heart disease prediction

Keywords: *Soft Voting Ensemble (SVE), Firefly optimization, LSKR(Logistic regression, SVM, K-nearest Neighbour(KNN), Random forest), Lasso method and SVM with Analytic Hierarchy Process (SVM-AHP)*

1. INTRODUCTION

The human heart encounters a multitude of variables on a daily basis, contributing to the development of various cardiac issues. Prompt identification and understanding of these disorders are imperative, particularly given the prevalent nature of cardiovascular diseases in contemporary society. In light of the escalating global mortality rates attributed to heart ailments, which account for a substantial portion of annual deaths worldwide, emphasis on cardiovascular health is paramount. The World Health Organization (WHO) highlights the alarming frequency of fatalities resulting from heart-related complications, underscoring the urgent need for comprehensive research and intervention strategies.

Among the myriad cardiovascular conditions, Congestive Heart Failure (CHF), Coronary Vascular Disease (CVD), Coronary Artery Disease (CAD), and Abnormal Heart Rhythms are particularly pervasive. Detection of cardiac abnormalities, especially in elderly individuals with co morbidities such as diabetes, presents a significant challenge due to varied symptomatology and subtle manifestations. This underscores the importance of accurate and efficient medical diagnosis, necessitating specialized expertise and advanced diagnostic modalities.

Advancements in technology, particularly machine learning algorithms and computer-based decision support systems, have revolutionized early detection and diagnosis of cardiac ailments. By harnessing vast datasets and leveraging intelligent algorithms, clinicians can now identify subtle pathological cues and predict disease onset with greater accuracy. Artificial neural networks, trained on historical data, have emerged as powerful tools for uncovering intricate disease dependencies and facilitating predictive modelling.

A crucial aspect of effective disease management lies in risk assessment and prognostication. A robust predictive model, fueled by accurate data, holds the potential to not only categorize individuals at risk of heart disease but also forecast the temporal onset of illness manifestation. Recognizing that heart disease encompasses a spectrum of cardiovascular disorders, including stroke, coronary heart disease, and heart attacks, underscores the multifaceted nature of cardiac pathology and the imperative for comprehensive risk evaluation.

In conclusion, the interdisciplinary integration of cutting-edge technology, clinical expertise, and robust data analytics holds immense promise in the quest to combat cardiovascular diseases and improve patient outcomes. By fostering collaboration between clinicians, researchers, and technology developers, we can usher in a new era of precision medicine aimed at mitigating the global burden of heart disease.

Here is an outline of the paper's structure: A brief summary of machine learning models and their use in the prediction of heart disease is provided in Section 2. The suggested method for detecting heart disease using machine learning techniques is presented in Section 3. In Section 4, the experimental setup is described in detail, and the findings are presented. In Section 5, the report culminates by proposing prospective research directions to improve the prediction of heart disease, highlighting areas that require additional investigation

2. LITERATURE SURVEY

Machine learning classification stands as a cornerstone and extensively employed method within the domain of machine learning. Its principal aim is to forecast categorical class labels of data instances based on their inherent features. Through the acquisition of patterns and relationships from labeled training data, classification models demonstrate the capability to render precise predictions on novel or unseen data, commonly referred to as test data. This procedure assumes particular significance when confronted with voluminous and intricate datasets, facilitating automated decision-making and pattern recognition. Within the realm of classification tasks, an array of algorithms is deployed, each characterized by its distinctive mathematical underpinnings and learning methodologies. Notable among these algorithms are logistic regression, support vector machines, decision trees, random forests, naive Bayes, k-nearest neighbors, and neural networks. The selection of an appropriate algorithm often hinges on the unique attributes of the dataset and the inherent complexities of the problem domain. Diverse algorithms present distinct strengths and weaknesses, underscoring the pivotal importance of selecting the most fitting one to attain optimal performance. Ensuring the efficacy of machine learning classification entails careful consideration of several key factors. Foremost among these is the paramount importance of high-quality data, which forms the bedrock for both training and testing the models. Additionally, feature selection emerges as a critical phase, necessitating the identification of the most pertinent and informative attributes from the dataset to enhance overall model performance.

Various model evaluation techniques, including precision, recall, accuracy, and F1-score, are employed to assess the performance of classification models. These metrics play a crucial role in quantifying the effectiveness of the model in accurately categorizing instances from the test data, thereby offering valuable insights into its strengths and limitations. The practical applications of classification span across numerous industries and sectors, exerting a profound impact on their operations. For instance, within the medical field, classification models play a pivotal role in disease diagnosis by analyzing patient symptoms and health records. In cybersecurity, classification algorithms aid in the detection of spam emails or malicious activities, bolstering security measures. Moreover, sentiment analysis relies on classification to discern the emotional tone conveyed in texts or social media posts, while image recognition extensively utilizes classification techniques to identify objects and patterns within images. In essence, machine learning classification emerges as a potent tool for facilitating data-driven decision-making, facilitating automated predictions, and pattern recognition across a myriad of real-world applications. Through the strategic utilization of appropriate algorithms, data quality, and evaluation methodologies, classification models furnish invaluable insights and bolster the efficacy of decision-making processes across diverse domains.

In a study [1], researchers applied altered K-means and Naïve Bayes algorithms to predict heart disease occurrence. Notably, their Naïve Bayes model achieved an impressive accuracy of 93%. The healthcare domain is rich with extensive datasets, making it an ideal arena for leveraging data science to glean valuable insights and enhance heart attack prediction accuracy. Within this research domain, datasets are meticulously structured based on pertinent medical parameters [2]. Among the diverse algorithms employed for accurate heart problem estimation, Naïve Bayes and Decision Trees have emerged as notable contenders [3][4], with Naïve Bayes being particularly favored for heart attack prediction. In another study centered on heart problem prediction [5], an exhaustive analysis of multiple risk factors is undertaken to evaluate their correlation with patient-related heart disorders. Subsequently, various algorithms are scrutinized to ascertain the most effective predictor of heart problems while mitigating false negatives. The overarching objective is to develop a robust and reliable heart problem prediction system poised to positively impact patient outcomes. Numerous studies have delved into the application of diverse machine learning (ML) techniques for heart disease identification.

In a particular investigation, a dataset comprising 14 parameters, such as age, blood pressure, and cholesterol, was gathered from the UCI Machine Learning repository. Artificial Neural Networks (ANN) demonstrated the highest accuracy rate of 96%, succeeded by Logistic Regression at 88%, Random Forest at 83%, Decision Tree at 83%, SVM at 70%, and K-NN at 68% [6]. Furthermore, machine learning (ML) models were harnessed to predict heart disease in a separate study. Among the three methodologies explored, KNN achieved the highest accuracy rate of 83%, while SVM exhibited the lowest accuracy at 65%, and Naïve Bayes achieved 80% accuracy [7]. Another investigation employed ML algorithms to analyze cardiovascular disease, encompassing Decision Tree, Logistic Regression, Random Forest, and Naive Bayes. The respective accuracy rates recorded were 81%, 85%, 90%, and 85% [8]. Notably, Artificial Neural Networks (ANNs) draw parallels to neurons in the human brain due to their interconnected layers of nodes. Intriguingly, ANN has been employed as a non-invasive means for the analytical examination of ischemic heart diseases and myocardial ischemia in preceding eras [9,10].

Recently, Artificial Neural Networks (ANN) have been employed as a supplementary approach to analyze medical and longitudinal data of heart failure (HF) patients, achieving notable accuracy rates of 77.8% for assessing HF severity and 84.73% for determining HF type [11]. Several studies have leveraged the UCI Machine Learning Repository database, housing pertinent medical data including age, blood pressure, cholesterol level, resting heart rate, and gender, to develop ANN-based Machine Learning models for analytical diagnostics of HF. These ANN methodologies have demonstrated proficient diagnostic accuracies ranging from 85% to 90% [12]. This underscores the potential of ANNs in contributing to effective and precise diagnostic assessments within the realm of heart failure. These investigations underscore the potential of machine learning techniques in accurately predicting heart disease and furnish valuable insights into the efficacy of diverse algorithms for this crucial medical application.

Cloud computing plays a pivotal role in evaluating the necessity and significance of extensive data in Internet of Things (IoT) scenarios. It has been postulated that IoT environments generate vast volumes of data, necessitating the application of big data analytics for its efficient management, storage, and evaluation. A CloudT project was delineated, which utilizes IoT sensors to disseminate information regarding various available services to residents of Japanese cities. Furthermore, a storage framework for Internet of Things systems, underpinned by cloud computing, was proposed. This approach could leverage the Hadoop file system to manage both structured and unstructured data, with an initial evaluation showcasing its efficacy. It was contended that optimal prediction performance is attained through the amalgamation of geographical proximity and catchment factors[13]. Additionally, research has demonstrated the utility of fuzzy models in establishing unknown correlations among crucial hydrological parameters, such as river flow and rainfall. The assertion was made that pervasive computing concerning flood warning systems would enhance their supervision[14]. Utilizing an adaptive neuro-fuzzy inference technique[15], water level projections were conducted for three distinct locations, with adjustments made based on flow conditions along the mainstream. Notably, cloud and cluster computing emerged as valuable and efficient technologies for flood monitoring systems, even in challenging circumstances. Nonetheless, while the hydrometeorological system effectively warned of imminent severe flooding, it encountered limitations in predicting the magnitude and timing of peak events[16].

To sum up, while weighted voting ensembles commonly utilize classifier-level weights derived from training results or metaheuristic methods like the Genetic Algorithm, the utilization of an enhanced firefly algorithm to generate class-level weights has not been previously documented, to the best of our knowledge. This paper elucidates the research undertaken on this particular technique.

3. PROPOSED METHODOLOGY

The techniques used to predict heart disease using the available dataset are described in this section. We then present our proposed approach, which uses machine learning algorithms to effectively predict heart diseases ahead of time. The system architecture shows how the proposed model operates and attempts to find correlations between the features of the data from the heart disease dataset.

3.1 Dataset Collection

The UCI Machine Learning Repository [28] provides the heart disease dataset, which consists of four databases from Cleveland, Hungary, Switzerland, and the VA Long Beach. The 303 entries and 14 attributes of the Cleveland database dataset were used in this investigation. 13 predictor variables and 1 independent class variable are present in each occurrence.

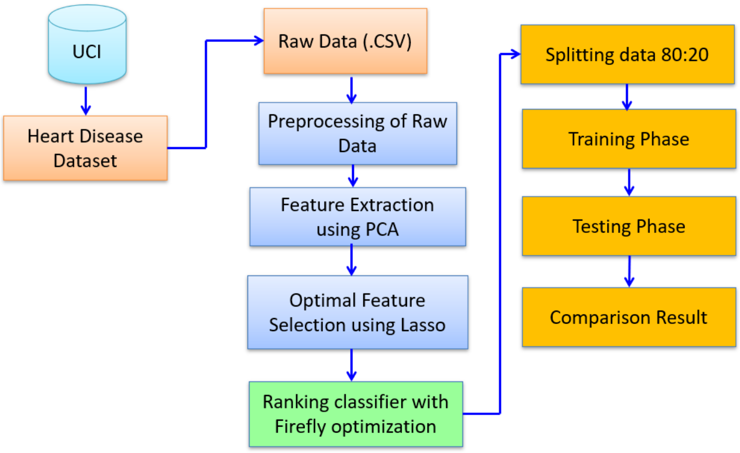
UCI
Heart Disease Dataset
Raw Data (.CSV)
Preprocessing of Raw Data
Feature Extraction using PCA
Optimal Feature Selection using Lasso
Ranking classifier with Firefly optimization
Splitting data 80:20
Training Phase
Testing Phase
Comparison Result

Figure1: System Architecture of Heart Disease Prediction System using LSKR-SVE(FO)

The proposed system's working model involves several steps:

Step 1 Sourcing the Heart Disease Prediction dataset from the UCI repository.

Step 2 Preprocessing the data samples by eliminating null values, applying denoising filtering, and removing outliers.

Step 3 Extracting features using PCA and selecting the most relevant attributes for heart disease prediction through Lasso method.

Step 4 Selecting efficient soft voting classifier with base algorithms such as (LR, SVM, KNN and RF) with Firefly optimization.

Step 5 This has been utilized for minimizing the overfitting problem in SVE and to categorize the chosen optimal features for predicting the efficient accuracy.

Step 6 Finally, evaluating the performance efficiency of Heart disease prediction and conducting comparisons with existing classifiers.

3.2 Preprocessing

Data preprocessing is widely recognized as a crucial step in applying data mining techniques to enhance prediction accuracy. By utilizing cleaned data and eliminating irrelevant or partially relevant features from the dataset, data preprocessing significantly contributes to improving prediction performance. This step involves the removal of unneeded, irrelevant, repeated data, as well as addressing spelling errors and unrealistic data that could potentially impact prediction accuracy or decrease overall performance.

It was discovered that there were no missing data entries in the cardiovascular disease databases used in this investigation. The total number of records in the dataset was decreased by identifying and removing duplicate records that had identical data values. In order to maintain uniformity, characteristics with different scales within the datasets were also rescaled, using a defined range of 0 to 1. Preprocessing data has a number of benefits. First off, it prevents judgments based on noise and lessens overfitting by removing redundant data. Second, because algorithms learn more quickly with fewer data points, it cuts down on training time. Clinical elements from various medical exams make up the remaining traits.

3.3 Feature Extraction using PCA

Principal Component Analysis (PCA) is a valuable tool in exploratory data analysis and predictive modeling. Its goal is to retain as much variability as possible while streamlining a dataset. PCA proves effective in visualizing genomic proximity and population correlations. The process involves normalizing the initial data, calculating the covariance or correlation matrix, and performing eigenvalue decomposition on the correlation matrix. [27][29]

3.4 Feature Selection using Lasso method

Predictive models require less time and money to train when redundant andunnecessary features are eliminated, which is made possible by feature selection.Before choosing a model and training it with the selected features, preprocessing data is done using feature selection. The model is trained using the complete dataset in order to make feature selection decisions. However, this approach may add bias in favor of particular attributes, which could inflate the model's performance in comparison to other models that have been examined.

Robert Tibshirani came up with the concept of LASSO, which stands for Least Absolute Shrinkage and Selection Operator, back in 1996. Regularization and feature selection are its two main functions. Through the imposition of an upper bound on the total absolute value of the model parameters, LASSO enables a process of shrinking that penalizes the coefficients of regression variables, forcing some of them to zero. Variables that have non-zero coefficients after downsizing are added to the model during feature selection. Reducing prediction error is the main goal of this process.[18]

The tuning parameter, which controls the penalty's severity, is very important. When big enough, it forces coefficients to exactly zero, which reduces dimensionality. Moreover, a larger percentage of coefficients shrink towards zero as the parameter's size increases.Obtain the Lasso regression coefficient of absolute values. Select the feature for the ideal feature subset that has a nonzero value.

$$LR^{lasso}(\hat{\beta}) = \sum_{i=1}^{k}\left(y^{i} - x'^{i}\hat{\beta}\right)^{2} + \boldsymbol{\lambda}\sum_{j=1}^{l}\left|\hat{\beta}^{j}\right| \qquad \text{---------------(1)}$$

The above equation(1) shows the calculation of penalty values.The penalty term, represented by alpha (α), denotes the extent of shrinkage or constraint imposed in the equation. A value of alpha greater than zero penalizes the optimization function, whereas a value of zero replicates the linear regression model (equation 1). As a result, Lasso regression reduces coefficients, which helps to lower multicollinearity and model complexity.

3.5 Feature Classification using Voting classifier with Firefly optimization

3.5.1 Voting Classifier:

A voting classifier is a model that consolidates the ensembles of other models to generate predictions based on the highest probability. It merges outputs from multiple classifiers and determines the outcome through majority voting. Two voting methods are utilized: soft voting, which averages the votes, and hard voting, which relies on the majority of votes. Due to its lower overfitting and low error rate, ensemble learning is the preferred choice.

The voting procedure is unique which is most used classifier-fusion technique. Consider a composite classifier which consists of Kseparate classifiers, v^1 through v^K. From a label set $\{x^1, x^2, \ldots, x^m\}$, each learner, represented by v^i, anticipates the sample label. To determine the final label, the judgment results from the K classifiers are combined in the fusion process. Assume that the classifier v^i produces a N-dimensional vector $(v_i^1(z), v_i^2(z), \ldots, v_i^N(z))K$ for a given input sample z. The output for the p^{th}class is denoted as $v_i^p(z)$. The types of these $v_i^p(z)$output values can range from binary labels to probability labels.

Weighted voting or majority voting can be used to aggregate binary labels [19].

The following is how the majority vote technique works:

$$V(z) = arg_j \max\left(\sum_{i=1}^{K}\left(v_i^p(z)\right)\right), v_i^p(z) \in \{0,1\} \quad \text{------- (2)}$$

The weighted voting can be calculated as:

$$V(z) = arg_j \max\left(\sum_{i=1}^{K} w^i \times \left(v_i^p(z)\right)\right), v_i^p(z) \in \{0,1\} \quad \text{---- (3)}$$

Here, $w^i \in [0, 1]$ signifies the weight assigned to classifier v^i, with a higher weight indicating more robust classification performance.

We used ensemble learning to apply the growth and pruning strategy for model optimization. Pruning is the process of removing specific members from a fully defined ensemble, whereas growing is the process of adding models to the ensemble to improve accuracy. The goal of this pruning procedure is to reduce computational complexity or the ensemble's model while maintaining performance. So, we introduce the optimization technique called improved firefly optimization for minimize the overfitting problem in voting classifier.

3.6 Firefly Algorithm:

Yang introduced the Firefly Algorithm [20], a metaheuristic approach inspired by the flashing behavior of fireflies and their bioluminescent communication. This algorithm operates based on several assumptions proposed by Yang. Fireflies, being unisexual, are attracted to each other regardless of gender.

Attractiveness correlates with brightness; thus, less bright fireflies are drawn to brighter ones, with attractiveness diminishing as distance increases.Fireflies with equal brightness move randomly.

New solution generations are produced through a combination of random walk and attraction among fireflies [21-24]. The brightness of fireflies corresponds to the objective function of the problem being addressed. Their attractiveness enables them to form smaller groups and swarm around local models, rendering the Firefly Algorithm suitable for optimization tasks[25,26].

Fireflies are naturally drawn to brighter counterparts. Additionally, brightness diminishes with distance following the inverse square law, as depicted in Eq. (4)

$F^I \propto \frac{1}{a^2}$ ----------- (4)

Suppose we have n fireflies in a scenario, and q^i is the answer for firefly i. Each firefly's brightness, represented by the I, has a correlation with the objective function f(q^i). According to Eq. 5, this brightness, I, indicates the fitness value or objective function of recent position, f (q).

$F^I = \mathrm{f}(q^i)$ ------------- (5)

Less bright (more attractive) fireflies are drawn towards brighter ones and subsequently move towards them. Each firefly possesses a specific attractiveness value, denoted as β. However, this attractiveness value β is relative and depends on the distance between fireflies. The attractiveness function of the firefly is defined by Eq. 6

$\beta(a) = \beta_0 e^{-\gamma a^2}$ ------------------- (6)

Firefly i, positioned at q^i, moves towards a brighter firefly j located at q^j, guided by their respective brightness levels [6] in equation [7].

$q^i(t+1) = q^i(t) + \beta_0 e^{-\gamma a^2}(q^i - q^j) + \propto \varepsilon^i$ -------------(7)

The system compares a new firefly position's attractiveness to the previous one. The firefly moves to the new spot if it receives a better attractiveness rating; if not, it stays where it is now. Either a predetermined fitness value or a predefined number of iterations dictate the Firefly Algorithm's termination criteria.

The brightest firefly undergoes random movement according to the conditions outlined in Eq. 8.

$q^i(t+1) = q^i(t) + \propto \varepsilon^i$ ----------------- (8)

3.7 Basic Algorithms Used:

a) K-Nearest Neighbor(KNN): One of the first machine learning algorithms, K-Nearest Neighbors (KNN), is categorized as supervised learning. As it sorts input data into the proper groupings within the existing dataset, it makes a distinction between new and old data. Although it is applicable to problems involving both regression and classification, it is usually used in classification processes. KNN can be applied to a variety of datasets because it is a non-parametric approach and does not make strong assumptions about the data.

b) Logistic Regression (LR). The maximum likelihood estimation concept forms the basis of logistic regression. It is anticipated that observed data will be the most likely under this paradigm. An activation function that can map values between 0 and 1 is applied to the weighted total of the inputs. The curve produced by this function, which is frequently referred to as a sigmoid function, is known as the sigmoid curve or S-curve.

c) Support Vector Machine (SVM): Significant accuracy is demonstrated by Support Vector Machine (SVM), which uses very little processing power. Researchers in machine learning have a strong belief in it. SVM is a flexible algorithm that can be used to solve regression and classification issues. To differentiate between classes in classification tasks, it uses hyperplanes.

d) Random Forest classifier (RF): The Random Forest classifier boosts the projected accuracy of a dataset by averaging the results of multiple decision trees trained on diverse subsets of the dataset. Instead of relying solely on one decision tree, the random forest aggregates predictions from many trees and determines the outcome based on majority votes. Increasing the number of trees in the forest enhances accuracy and mitigates overfitting.

4. RESULTS AND DISCUSSION

In this study, Google Colab was utilized on an HP computer equipped with a 3200-H processor and 8 GB of RAM. Initially, the dataset consisted of approximately 18,000 rows and 14 attributes, but after cleaning and preprocessing, it was reduced to around 9,000 rows and 10 attributes. As all attributes were categorical, outlier removal was conducted to enhance model efficiency. The algorithms employed in this research included Voting classifier with Firefly optimization. The dataset was split into two segments: 80% for training the model and 20% for testing. Various performance metrics, including precision, sensitivity, accuracy, and F1 score, were evaluated.

Performance Metrices:

To assess the effectiveness of the trained classifiers, diverse performance metrics are applied to the test set. These essential evaluation criteria include accuracy, precision, recall, and F1-score. To gain deeper insights into these metrics, we will explore their definitions, calculations, and potential interrelations. True Positives (TP), True Negatives (TN), False Positives

(FP), and False Negatives (FN) represent the foundational components of these metrics.

The accuracy score can be calculated as

$$accuracy = \frac{T^p+T^n}{T^p+T^n+F^p+F^n} \qquad (9)$$

The recall can be measured as

$$recall = \frac{T^p}{T^p+F^n} \qquad (10)$$

The precision can be computed using

$$precision = \frac{T^p}{T^p+F^p} \qquad (11)$$

The F-score can be calculated as

$$F_{Score} = 2 \times \frac{precision \times recall}{precision + recall} \qquad (12)$$

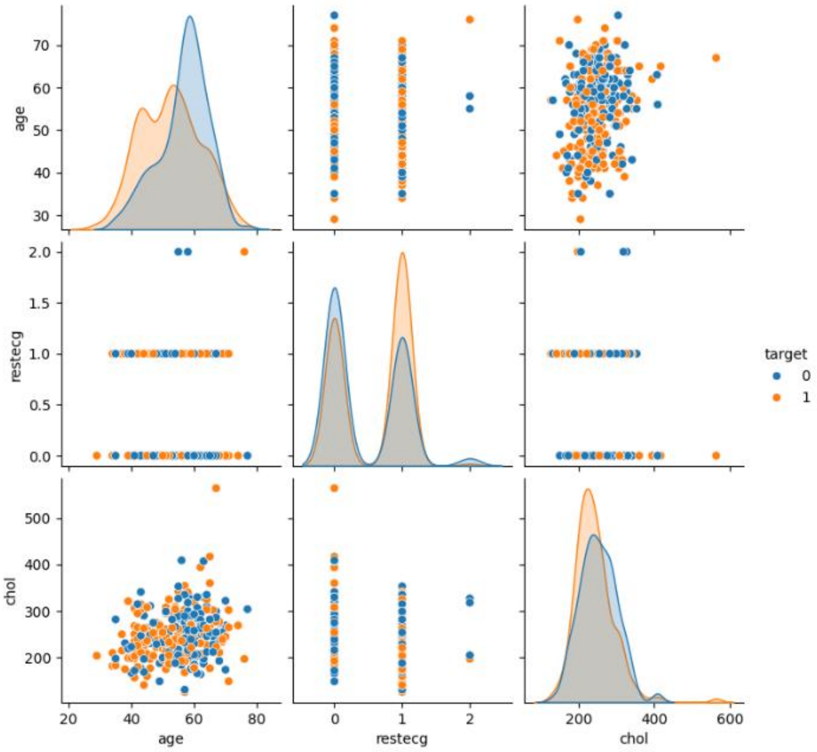

Figure 2: Attribute mapping (Cholesterol and RestECG position) for Heart Disease Prediction

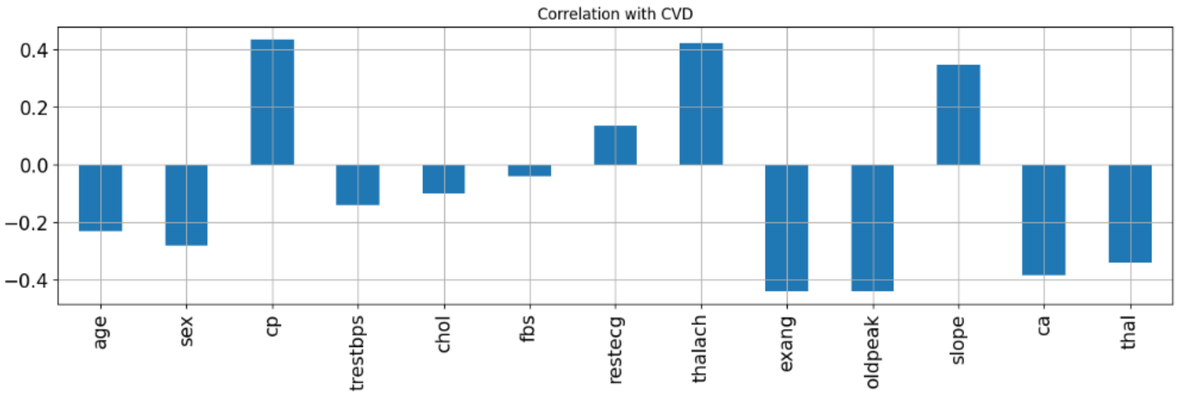

Figure 3: Bar chart for Attribute correlation with Heart Disease Prediction

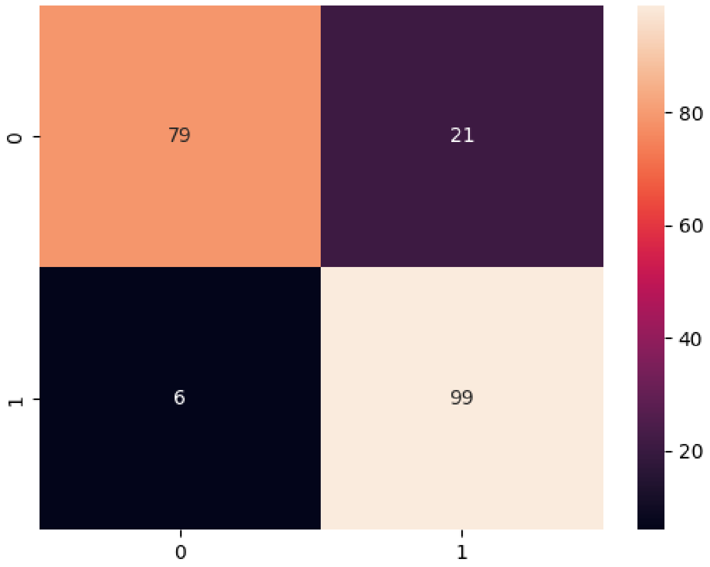

Figure 4: Feature Selection using Lasso

Table 1: Evaluation Table for AnticipatedLSKR-SVE(FO)

Performance Measures (in %)	Proposed LSKR-SVE(FO)	SVM-AHP	SVM	KNN	Adaboost
Accuracy	99.3	96.3	91.2	88.69	89.36
Recall	96.2	88.3	78.6	73.45	76.35
Precision	98.45	98.5	93.8	89.56	83.75
F-Score	94.85	89.24	81.3	79.65	80.17

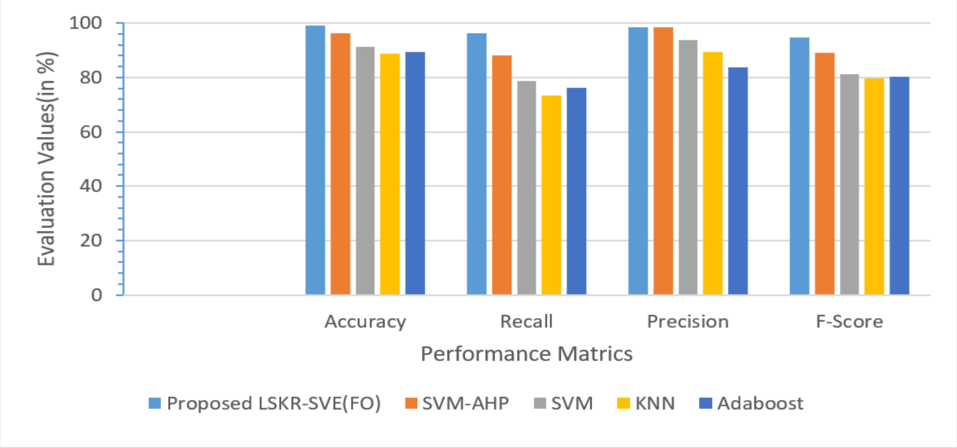

Figure 5: Comparison Result of Proposed LSKR-SVE(FO)

In Figure 2, depicting Attribute mapping (Cholesterol and RestECG position) for Heart Disease Prediction, relevant features such as Cholesterol and RestECG position were utilized alongside age to predict heart disease. Figure 3 presents a bar plot illustrating the distribution of attributes with respect to the target value. Based on this visualization, pertinent features for heart disease prediction were selected. Following Principal Component Analysis (PCA), the relevant features were assessed for correlation with the target values, resulting in a 28% reduction in features using the Lasso method. This reduction is demonstrated in the Correlation matrix depicted in Figure 4.

In this proposed methodology, multiple machine learning models are amalgamated using the ensemble approach to yield a collective outcome that surpasses the accuracy of any individual algorithm. Specifically, voting ensembles aggregate the predictions from 4 ML models through a voting mechanism.

After examining Figure 5, it is apparent that the LSKR-SVE(FO) classifier consistently outperforms individual ML classifiers in the dataset, achieving 99.3% accuracy. While individual classifiers reach a maximum accuracy of only 92%, as observed with the SVM classifier, the LSKR-SVE(FO) classifier effectively leverages the strengths of the four individual classifiers, resulting in enhanced accuracy. This underscores the potential of ensemble methods for improving performance in heart disease prediction tasks. Tables 1 shows the comparison result of previous research with the results of our proposed model on the dataset, demonstrating that our model yields higher accuracy compared to prior work.

The proposed method shows similar classification accuracy to other classifiers: SVM-AHP, SVM, KNN, and Adaboost are outperformed by 3%, 8.1%, 10.6%, and 9.9%, respectively, by LSKR-SVE(FO). Moreover, LSKR-SVE(FO) has a higher classification recall than SVM-AHP, SVM, KNN, and Adaboost, with respective advantages of 7.9%, 17.6%, 22.75%, and 19.85%. Furthermore, our suggested model improves classification precision in comparison to SVM, KNN, and Adaboost by 5%, 8.9%, and 14.75%, respectively. The classification precision of SVM-AHP is also somewhat lower than that of LSKR-SVE(FO). Finally, SVM-AHP's classification F-score performs 4.65%, 8.89%, and 14.7% better than SVM, KNN, and Adaboost. To sum up, the suggested LSKR-SVE(FO) outperforms current techniques in terms of classification performance.

5. CONCLUSION

Heart disease stands as a prominent global cause of mortality, emphasizing the importance of early diagnosis to impede its progression. This research introduces an effective approach for heart disease prediction, yielding promising outcomes in risk classification. Leveraging PCA, features are extracted to streamline dataset dimensionality, with relevant features further selected via the Lasso method. Classification employs SVE with base learners including Logistic Regression, SVM, KNN, and Random Forest, albeit mindful of potential misclassification errors and overfitting, which are addressed through weighted parameter tuning in SVE via FO optimization, facilitating informed diagnostic decisions. Results showcase notable enhancements in accuracy (99.3%), precision (98.45%), recall (96.2%), and F1-Score (94.85%) compared to other classifiers in Cleveland datasets. Moreover, feature reduction from 14 to 4 features slashes computational load by 48%. Future endeavors entail integrating real-time patient data, refining feature selection methods, and developing a deep learning-based system for early heart disease detection. Deploying this ML model into practical applications, such as web or mobile platforms, enables real-time heart disease risk prediction, offering invaluable support in clinical settings for patient care

REFERENCES

[1] Choudhary, G., & Singh, S. N. (2020, October). Prediction of heart disease using machine learning algorithms. In 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE) (pp. 197-202). IEEE.

[2] Saravanakumar, S., & Saravanan, T. (2022). An effective convolutional neural network-based stacked long short-term memory approach for automated Alzheimer’s disease prediction. Journal of Intelligent & Fuzzy Systems, 43(4), 4501-4516.

[3] Jayasudha, R., Suragali, C., Thirukrishna, J. T., & Santhosh Kumar,B. (2023). Hybrid optimization enabled deep learning-based ensemble classification for heart disease detection. Signal, Image and Video Processing, 1-10.

[4] PR, S. (2023). An effective healthcare monitoring system in an IoMT environment for heart disease detection using the HANN model. Computer Methods in Biomechanics and Biomedical Engineering, 1-10.

[5] Saravanan, T., & Venkatesan, D. (2022, December). Predicting Consumer Intention using Logistic regression by analyzing social media data. In 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) (pp. 512-516). IEEE.

[6] Eknath, K. H., Bhanudas, K. D., Jalindar, D. B., &Gunjal, S. N. (2023, August). An effective heart disease predication using machine learning techniques. In AIP Conference Proceedings (Vol. 2790, No. 1). AIP Publishing.

[7] Solomon, D. D., Khan, S., Garg, S., Gupta, G., Almjally, A., Alabduallah, B. I., ... & Abdallah, A. M. A. (2023). Hybrid Majority Voting: Prediction and Classification Model for Obesity. Diagnostics, 13(15), 2610.

[8] Ramesh, B., & Lakshmanna, K. (2023). Multi Head Deep Neural Network Prediction Methodology for High-Risk Cardiovascular Disease on Diabetes Mellitus. CMES-Computer Modeling in Engineering & Sciences, 137(3).

[9] Thakur, A., Kaur, H., Goel, N., Paul, P., Asopa, P., Goswami, S., & Das, M. K. (2023). A Hybrid Approach for Heart Disease Detection using K-Means and K-NN Algorithm. American Journal of Electronics & Communication, 4(1), 14-21.

[10] Saravanan, T., Saravanakumar, S., Dandu, S., Vinotha, D., Kadhim, A. K., & Al-Chlidi, H. (2023, May). Prediction of Infant Growth using the Random Forest Algorithm. In 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1435-1439). IEEE.

[11] Ahmed, R., Bibi, M., & Syed, S. (2023). Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative study of SVM and KNN Algorithms. International Journal of Computations, Information and Manufacturing (IJCIM), 3(1), 49-54.

[12] Omankwu, O. C., & Ubah, V. I. (2023). Hybrid Deep Learning Model for Heart Disease Prediction Using Recurrent Neural Network (RNN). Journal of Science and Technology Research, 5(2).

[13] Bajpai, A., Sinha, S., Yadav, A., & Srivastava, V. (2023, June). Early Prediction of Cardiac Arrest Using Hybrid Machine Learning Models. In 2023 17th International Conference on Electronics Computer and

[14] Computation (ICECCO) (pp. 1-7). IEEE.

[15] Saravanabhavan, C., Saravanan, T., Mariappan, D. B., Nagaraj, S., Vinotha, D., &Baalamurugan, K. M. (2021, March). Data Mining Model for Chronic Kidney Risks Prediction Based on Using NBCbH. In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1023-1026). IEEE.

[16] Houssein, E. H., Mohamed, R. E., & Ali, A. A. (2023). Heart diseaserisk factors detection from electronic health records using advanced NLP and deep learning techniques. Scientific Reports, 13(1), 7173.

[17] Lu, Y. (2023). Heart Disease Prediction Model based on Prophet.Highlights in Science, Engineering and Technology, 39, 1035-1040.

[18] Chen, Q., & Huang, L. (2020). Research on Prediction Model of Gas Emission Based on Lasso Penalty Regression Algorithm. In Artificial Intelligence in China (pp. 165-172). Springer, Singapore.

[19] Cao, J.; Kwong, S.; Wang, R.; Li, X.; Li, K.; Kong, X. Class-specific soft voting based multiple extreme learning machines ensemble. Neurocomputing 2015, 149, 275–284

[20] Yang, X. S. (2008). Nature-Inspired Metaheuristic Algorithms. Frome: Luniver Press. ISBN 1-905986-10-6.

[21] Xin She Yang. (2011). Optimization Algorithms. Comput. Optimization, Methods and Algorithms, SCI 356. pp. 13–31.

[22] Yang, X.S. (2011). Metaheuristic and Optimization: Algorithm Analysis and Open Problems. Lecture Notes in National Physical Laboratory, UK.

[23] Yang, X.S. (2011). Metaheuristic Optimization. Scholarpedia. 6(8):11472.

[24] Yang, X. S. (2010), Firefly Algorithm, Stochastic Test Functions and Design Optimization. Int. J. Bio-Inspired Computation. 2, No. 2, pp.78–84.

[25] Yang, X. S. (2010). Nature-Inspired Metaheuristic Algorithms. Luniver Press. Second Edition.

[26] Yang, X.S. (2010). Firefly Algorithm for Multimodal Optimization. In: Stochastic Algorithms: Foundations and Applications, SAGA 2009, Lecture Notes in Computer Science., Vol. 5792, pp. 169-178.

[27] M. A. Hambali and R. G. Jimoh, “Performance Evaluation of Principal Component Analysis and Independent Component Analysis Algorithms for Facial Recognition,” J. Adv. Sci. Res. Its Appl., vol. 2, pp. 47 – 62, 2015.

[28] https://archive.ics.uci.edu/dataset/45/heart+disease

[29] Dr.N.AnandhaKrishnan, https://bpasjournals.com/library-science/index.php/journal/article/view/2329.