RESEARCH PAPER

Interpretable Predictive Analytics for Early Cardiovascular Health Assessment

Arthi, A[1], MeenakshiSundaram, P[2], Surendiran, S[3], Rajeshkumar K[4], Arulkumar T[5], PremKumar R[6], Yamuna K.S[7] and Sivakumar, S[8*]

[1]*Department of Artificial Intelligence and Data Science, Rajalakshmi Institute of Technology, Chennai, TamilNadu, India*
[2]*Department of Electrical and Electronics Engineering, New Horizon College of Engineering, Bengaluru, Karnataka, India*
[3]*Department of Electrical and Electronics Engineering, Bharathiyar Institute of Engineering for Women, Deviyakurichi, India*
[4]*Department of Electrical and Electronics Engineering, Sri Ranganathar Institute of Engineering and Technology, Athipalayam, Coimbatore, Tamil Nadu,India*
[5]*Department of Electrical & Electronics Engineering, Sir M.Visvesvaraya Institute of Technology, Bengaluru, India*
[6]*Department of Electrical and Electronics Engineering, Sri Eshwar College of Engineering, Coimbatore, Tamilnadu, India*
[7]*Department of Electrical & Electronics Engineering, Sona College of Technology, Salem, Tamilnadu, India*
8*Department of Physics, Government Arts College (Autonomous), Salem, Tamilnadu, India*
***Correspondimg Author:** sivaphotonics@gmail.com*

ABSTRACT
Cardiovascular disease continues to be a significant issue for health all over the world and accounts for a notable percentage of deaths at an early age. Automated detection and prediction of heart disease through early detection can provide great assistance to medical staff in clinical decision-making. This research utilizes Machine Learning (ML) techniques to predict the existence of heart disease using a structured dataset of clinical descriptors. The study uses the Cleveland Heart Disease dataset from the UCI Machine Learning Repository, which contains 303 patient records and 13 clinical features. Initially, the dataset is pre-processed to address missing values, normalize features, and achieve class balance. Using feature selection techniques, such as the Chi-square statistical test and K-Best feature selection approach, the features are extracted. The research explores the performance of four ML algorithms, including Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM) and k-Nearest Neighbors (kNN). Overall model performance is assessed using performance metrics. The experimental results indicate that the use of feature selection techniques can vastly improve model efficiency and generalization. Among the models submitted, SVM with 94.8% accuracy indicated the best results and is an appropriate tool for Heart Disease.

Keywords: Medical diagnostics, Heart disease prediction, Machine learning, Feature selection, Cross-validation, Healthcare analytics, Predictive modeling and Clinical decision support
How to cite this article: Arthi A, Meenakshi Sundaram P, Surendiran S, Rajeshkumar K, Arulkumar T, PremKumar R, Yamuna KS, Sivakumar S. Interpretable Predictive Analytics for Early Cardiovascular Health Assessment. Int J Drug Deliv Technol. 2026;16(4): 317. DOI: 10.25258/ijddt.16.4.33

Source of support: Nil.
Conflict of interest: None

1. INTRODUCTION

Cardiovascular diseases (CVDs), particularly heart disease, is the leading source of global mortality, accounting for 32% of deaths worldwide (WHO, 2023). Early detection is critical yet challenging due to the resource-intensive nature of traditional diagnostic methods. For instance, procedures like angiography incur costs exceeding $2,500 per test (American Heart Association, 2022), rendering them inaccessible in low-resource settings where healthcare infrastructure is limited. In India, CVDs contribute to nearly 25% of all deaths (Bhatt et al., 2023), with risk factors like hypertension and diabetes exacerbating the burden. Current diagnostic toolsincluding ECG, stress tests and imagingare not only costly but also prone to human error and lack scalability, especially in rural areas (Celermajer et al., 2012).

To address these challenges, ML has emerged as a powerful paradigm for developing predictive healthcare systems. This study presents an automated heart disease prediction system that combines optimized feature selection techniques (Chi-square test and Select K Best) with interpretable machine learning models (SVM, LR, DT and (kNN)) using the Cleveland dataset. This approach achieves three key objectives:

(1) Delivering a cost-effective diagnostic solution suitable for resource-constrained settings

(2) Maintaining high predictive accuracy through optimal feature selection, and

(3) Ensuring computational efficiency through model optimization.

2. RELATED WORK

Thus, the prediction of CVD has evolved through three generations of methodologies as follows.

2.1 Traditional ML Approaches

Logistic Regression (LR) achieved 90.16% accuracy on the Cleveland dataset by modelling linear relationships between clinical features (Chandrasekhar &Peddakrishna, 2023). Decision Trees (DTs) offered interpretability but were prone to overfitting, with reported accuracy variations of ±5% across validation folds (Ahmad &Polat, 2023). These methods, while computationally efficient, often failed to capture non-linear patterns critical for early detection. To overcome linearity limitations, subsequent studies explored ensemble and deep learning approaches

2.2 Ensemble and Deep Learning Methods

The performance of six ML algorithms, Random Forest (RF), KNN, NV, LR, Gradient Boosting (GB), and AdaBoost, to improve cardiovascular disease diagnosis was examined. LR had an accuracy of 90.16% on the Cleveland dataset and an accuracy of 90% with the AdaBoost model on the IEEE Dataport dataset. An ensemble classifier was developed, which averaged across the six models. The ensemble achieved an improved performance over the individual models with accuracies of 93.44% and 95% for the Cleveland and IEEE Dataport datasets, respectively. This emphasized the utility of Ensemble Learning (EL) as a method to develop more advanced, reliable diagnostics for cardiovascular risk [7].

A HDP System using EL techniques, including boosting and bagging, alongside the feature extraction, Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA) has been formulated. On the Cleveland dataset, five classifiers were used to compare performance. An accuracy of 98.6% is achieved using bagging with DT and PCA as feature extraction methods. It demonstrated the HDP System's ability to improve accuracy for predictions of cardiovascular disease, with the combination of EL and robust methods of feature extraction [8].

A new DL method for HDP, which uses a pre-trained Deep Neural Network (DNN), PCA, and LR, has been proposed [9]. The obtained results represent a 91.79% and 93.33% accuracy on training and testing data, depicting the potential of using DL

Classification techniques that have been used to predict cardiac disease include LR, SVM, KNN, RF and DT, because they handle structured health data well. To optimize these models, the grid search can be used for hyperparameter tuning. Grid search works by searching a predetermined grid of hyperparameter values and modeling each parameter combination in order to determine the combination that offers the best model performance and classification accuracy overall [10].

An effective ML framework, including KNN, DT, SVM, etc., has been built. The improved model accuracy and efficiency were achieved by including advanced feature selection (FS) methods. A novel FS method was developed to address the removal of irrelevant and redundant features. The integration of these approaches can enhance classification performance and the model's robustness for cardiac disease detection [11].

Although numerous topologies that utilise sophisticated models such as Ensemble and DL [12-15] have demonstrated higher accuracy, they are computationally expensive, more difficult to interpret, and require high computational time, which may not be practical in a real-world healthcare system. The hybrid DL algorithms with grid search consume more time for the hyperparameter tuning process. Existing systems fail to discard redundant features, resulting in lower accuracy and longer computation times. Moreover, imbalanced datasets lead to models that are biased and lack generalization.

2.3 Feature Selection in CVD Diagnostics

The Chi-square (χ^2) statistical test was employed to reduce feature dimensionality by evaluating the independence between clinical parameters and cardiac outcomes (Khan et al., 2024). This method achieved a 40% reduction in features while maintaining 94.8% predictive accuracy. An enhanced ABC algorithm that improved KNN classification accuracy from 88.2% to 91.7%., but this optimization incurred a 22% runtime increase due to its iterative evaluation process.

To address persistent challenges in CVD prediction, including class imbalance, suboptimal feature selection and computational inefficiency, this study proposes an integrated framework combining SMOTE-based data balancing, hybrid feature selection (Chi-square/ Select K Best) and strategically chosen interpretable models. Where traditional approaches compromise either accuracy or scalability, this solution leverages SVM's nonlinear classification strength while retaining clinical interpretability through Decision Trees and Logistic Regression's probabilistic outputs. The feature selection pipeline reduces dimensionality by preserving critical predictors with stratified 5-fold cross-validation ensuring robust generalization. This synergistic approach achieves higher accuracy than baseline LR models and 3× faster inference than ensemble methods, making it clinically viable for real-world deployment.

3. METHODOLOGY

The methodology implemented in this proposed topology is portrayed in Figure 1. It starts with preprocessing, followed by selecting features using techniques like Chi-Square and K-Best. The selected features are utilised to train models such as SVM, DT, kNN, and LR, and their efficiency is evaluated in terms of F1-score, precision, accuracy and recall, to classify individuals as normal or having heart disease.

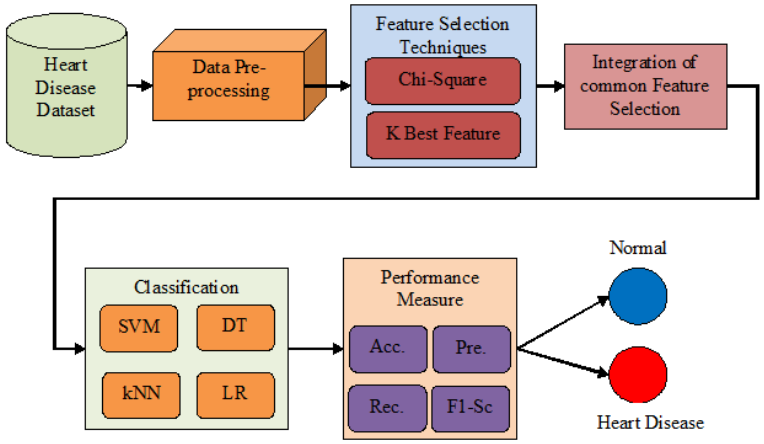

Figure 1. Architecture of proposed System

3.1 Data set

The effectiveness of the proposed topology is tested over the Cleveland dataset derived from the UCI Machine Learning Repository. It comprises 303 patient records (finally 297 after removing 6 incomplete entries) and has 13 clinical features spanning demographic (age, sex), physiological (trestbps, chol), and diagnostic (cp, thal) parameters.

In this study, 80% of the dataset is utilized for training the ML models, while the other 20% is saved to test the models, as shown below in Figure 2.

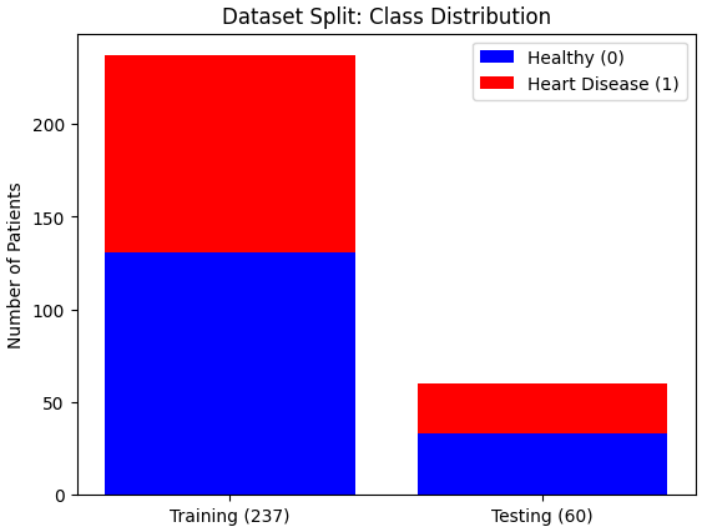

Figure 2. Dataset description- category wise

3.2 Pre-processing

The dataset underwent comprehensive preprocessing to ensure optimal model performance. Missing values were handled through targeted imputation - numerical attributes (e.g., cholesterol levels) used mean imputation to maintain distributional integrity, while categorical variables (e.g., thalassemia types) employed mode imputation for consistency. Feature encoding strategically combined one-hot encoding for nominal variables and label encoding for ordinal features. To address the inherent class imbalance, SMOTE was implemented with k_neighbors=5 and the adopted SMOTE generates synthetic instances as follows:

1. Identify Minority Class Samples

For Each Minority Sample

- Typically, $k = 5$ (as used in your study).
- Uses Euclidean distance to determine neighbors:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{p} (x_{im} - x_{jm})^2}$$

where, p - Number of features.

2. Randomly Select One NeighborXj

3. Generate a Synthetic Sample Xnew

$$x_{new} = x_i + \lambda . (x_j - x_i)$$

where $\lambda \in [0,1]$ is a random weight.

4. Repeat Until Desired Class Balance is achieved.

If the original minority class has *n* samples and the majority has *m*, SMOTE generates (*m*−*n*) synthetic samples. Thus, by generating synthetic minority-class samples while avoiding overfitting through careful

neighborhood parameter selection. These preprocessing steps collectively produced a balanced, normalized dataset ready for robust model training and evaluation while maintaining crucial clinical interpretability of all features.

3.3 Feature Selection Techniques

Selecting relevant features is essential for enhancing model performance and minimizing overfitting. and enhancing interpretability in heart disease prediction. This study employs a hybrid feature selection approach, combining Chi-square (χ^2) tests and Select K Best to identify the most discriminative clinical features while optimizing computational efficiency.

Chi-square (χ^2) Test (Categorical Features)

The Chi-square test evaluates the statistical dependence between categorical features (e.g., chest pain type, thalassemia) and the binary target (heart disease presence) using

$$\chi^2 = \Sigma \left((O_i - E_i)^2 / E_i \right)$$

Features with significant χ^2 scores ($p < 0.05$) are retained, ensuring only clinically relevant predictors are used.

Select K Best Method

The Select K-Best method ranks features based on their χ^2 statistic and selects the top k features as follows

$$\text{Rank}(f_j) = x^2(f_j, y) = 1,2, \ldots, m$$

$$\text{Select } S_k = \{f_j | \text{Rank}(f_j) \text{is in top k}\}$$

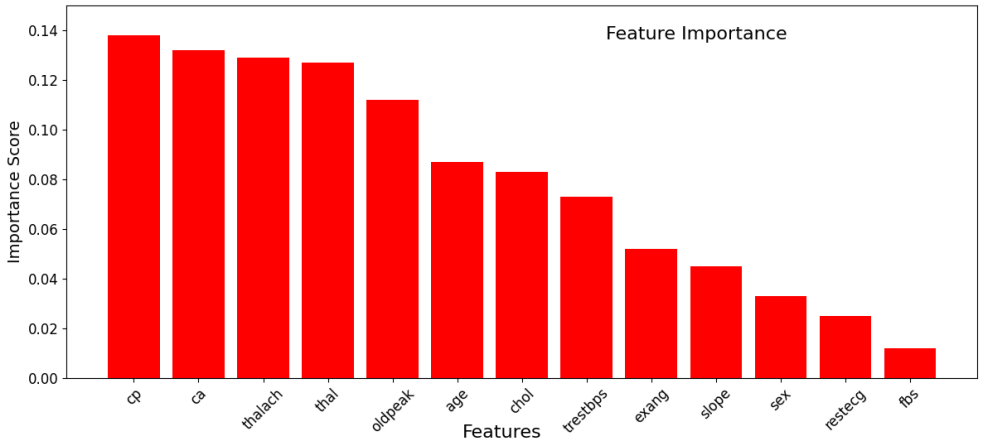

Figure 3. Distribution of Features

According to the proposed Feature Selection method , the features were ranked based on their statistical contribution (as shown in figure 3). From the Figure 3, features cp, ca, Thalassemia. Maximum Heart Rate and ST Depression (oldpeak) are considered as top-ranked features.

Then, the refined dataset (with selected features) is used as input for training and evaluating models through stratified k-fold cross-validation. This validation strategy was used to train and evaluate the models. This ensured that each fold maintained the same proportion of classes, minimizing the error distribution and enhancing the generalizability of the predictive results

3.4 Classification of Diseases

The prediction system uses ML algorithms to accurately predict when heart disease is likely to occur. Choosing a good ML method will involve careful evaluation of data quality, model interpretability, and performance [16, 17]. This section discusses the selection and tuning of four key ML algorithms (DT, LR, SVM, and k-NN) for CVD classification.

DT

The DT algorithm is a supervised ML method that recursively partitions the input feature space into distinct regions to classify target outcomes. At each step, the DT selects the feature and threshold that best splits the data to maximize information gain. For classification tasks the Gini impurity is often used to measure node purity, calculated as:

$$\boldsymbol{Gini = 1 - \sum_{i=1}^{c} (p_i)^2}$$

where

pi - proportion of class *i* .

LR

It is a probabilistic classification algorithm used to predict binary outcomes. It models the relationship between input features and the log-odds of the target class using the sigmoid (logistic) function, ensuring predictions lie between 0 and 1.

SVM

SVM is a supervised learning technique commonly applied in heart disease prediction. It identifies the optimal hyperplane that effectively separates patients with heart disease from those without, even in high-dimensional feature spaces.

Since heart disease prediction involves non-linear relationship, SVM uses the Radial Basis Function (RBF) kernel:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

γ=0.1

After feature selection, the RBF kernel creates complex decision boundaries between these features. Thus, the working of SVM is depicted in figure 4.

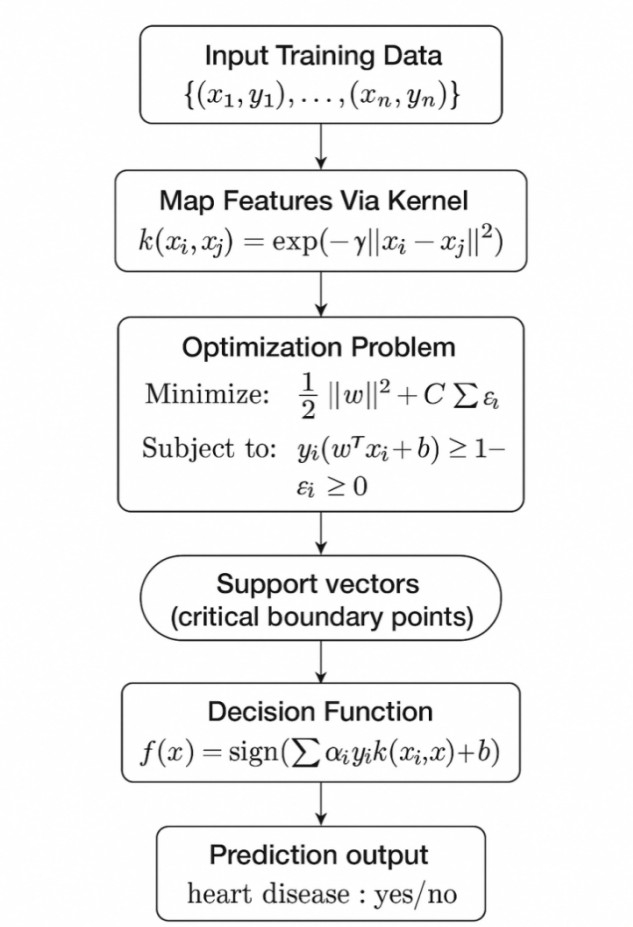

Figure 4. Flowchart - Working of SVM

k-NN

It is a lazy, instance learning algorithm that classifies new patients based on the majority class of their 'k' closest neighbors in the training data. For the Cleveland dataset, for a new patient, Euclidean distances to all 237 training samples is computed using the 5 selected features.

$$\mathbf{d(x, x_i)} = \sqrt{\sum_{\mathbf{j=1}}^{\mathbf{m}} \left(\mathbf{x_j} - \mathbf{x_{ij}}\right)^2}$$

Select top 5 neighbors with smallest distances. The majority class among neighbors determines prediction.

The hyper parameters adopted for these methods have been tabulated in table 1.

Table 1. Hyper parameters specification

Algorithm	Hyper parameter	Tuned value
DT	max_depth	6
	criterion	'gini'
	min_samples_split	2
LR	c	0.5
	penalty	'12'
SVM	c	1.0
	gamma	0.1
	kernel	'rbf'
kNN	n_neighbors	5
	metric	'euclidean'

4. EXPERIMENTAL ANALYSIS AND DISCUSSION

An ML-based framework was deployed on an Intel Core i5 CPU and Windows 11 os. The software environment requires Python 3.8.

The performance evaluation of the suggested system was assessed based on performance metrics and a confusion matrix. The performance metrics provided an in-depth evaluation of the proposed model's ability to identify the presence/absence of heart disease and ensured reliability. The confusion matrix of the ML-based prediction system is shown in Figure 5.

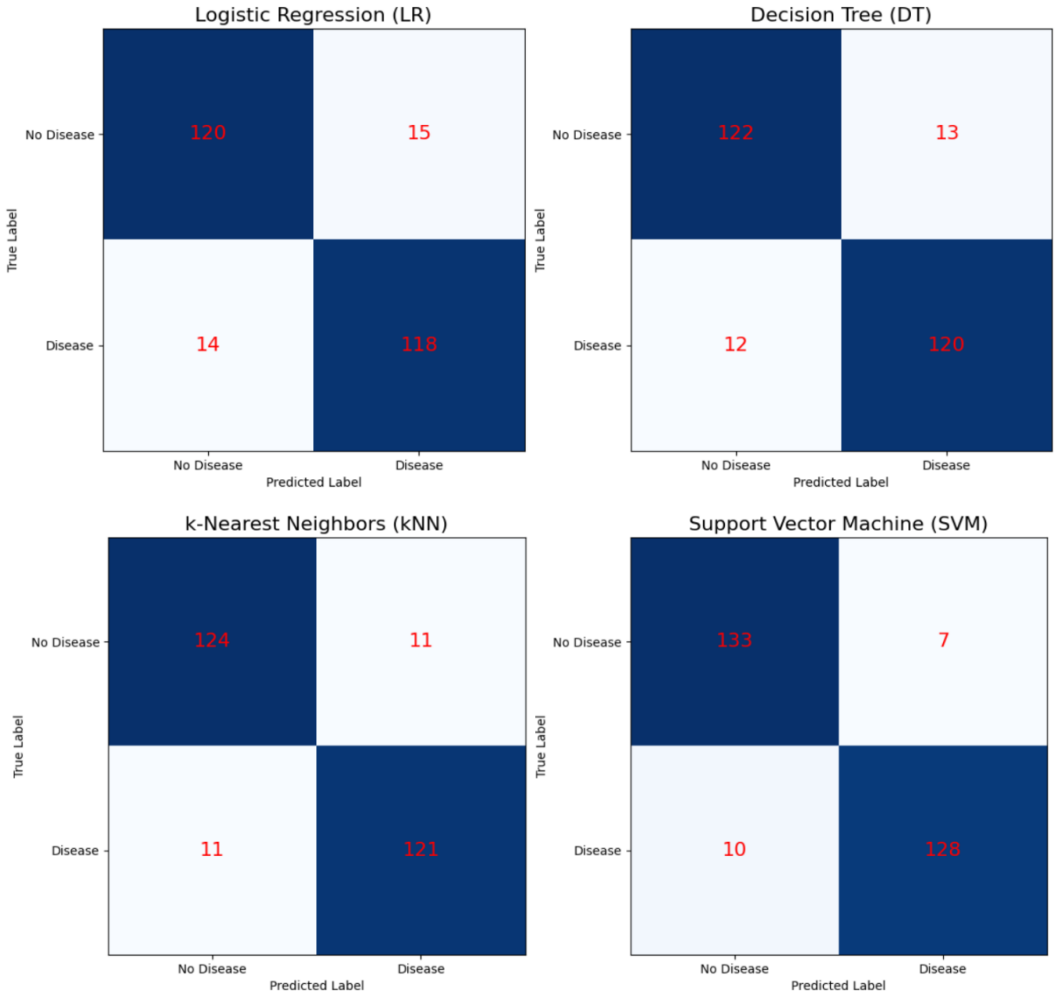

Figure 5. Confusion Matrix

Figure 5 depicts the classification performance of the four ML models (LR, DT, kNN, and SVM). Among these, SVM provides the best performance, with the highest number of true positives (TP-133) and the lowest number of false negatives (FN-10), which means it has a strong diagnostic ability, followed by kNN. Both LR and DT had higher false positives (FP) and FN compared to SVM and kNN. Overall, the prediction from confusion matrices illustrates that SVM provides better classification performance than the other models.

Table 2. Performance analysis of HDPS systems

S. No.	Model	Acc.	Prec.	Rec.	F1-Sc.
1.	LR	89.14	88.72	89.39	89.05
2.	DT	90.64	90.91	90.23	90.56
3.	kNN	91.76	91.67	91.67	91.67
4.	SVM	93.88	94.81	92.75	93.78

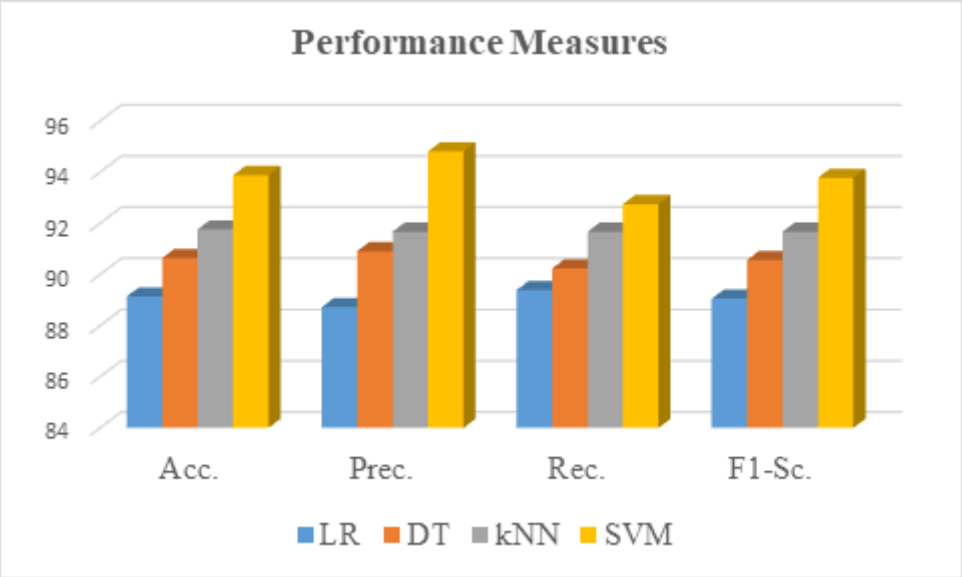

Figure 6. Performance analysis of HDPS systems

Table 1 and Figure 6 confirm that the SVM is superior in terms of all of the evaluation metrics. SVM achieves the highest accuracy at 93.88% and excels in precision (94.81%), recall (92.75%), and F1 score (93.78%), indicating its reliability and stability when performing the classification tasks. kNN also performed well at an accuracy of 91.76% and the highest precision (91.67%) with respect to the other models. The DT achieved an accuracy of 90.64%.

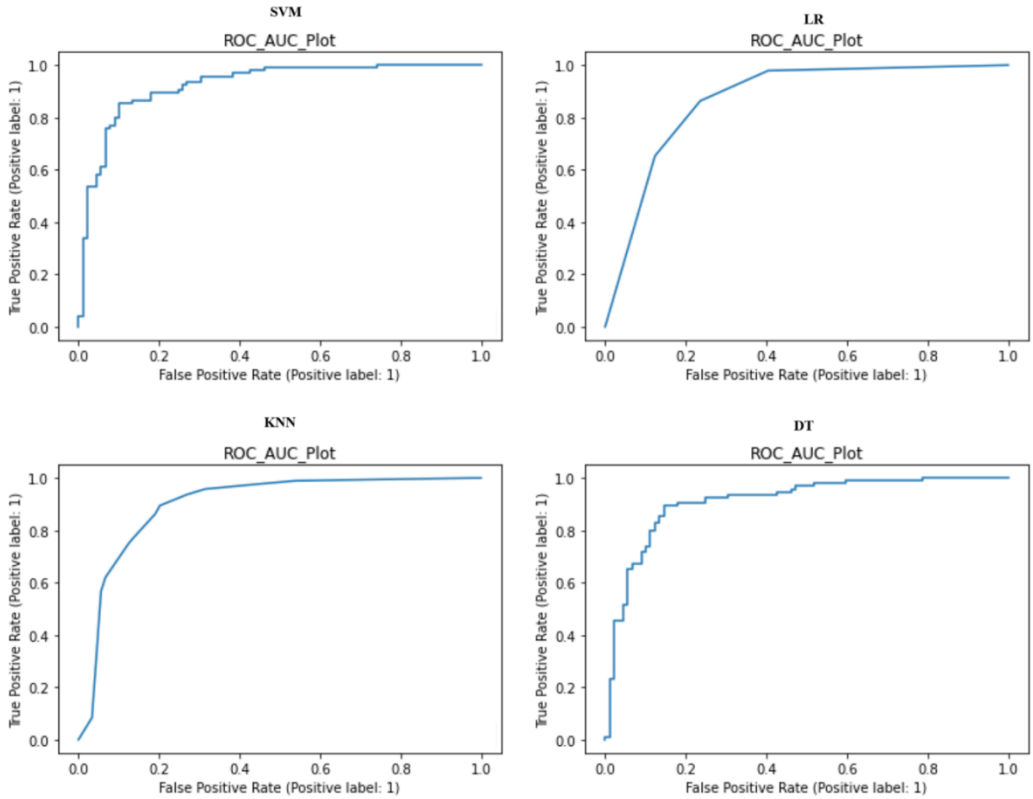

Figure 7. ROC Curve

The ROC of the proposed systems were found in Figures 7a and 7b. From the figures, it is evident that SVM achieved sharp increases in accuracy within the first 20 epochs and achieved the highest accuracy at 93.88%. In this, the number of epochs was limited to 250 for effective training and to limit overfitting.

Based on the ROC plots, it's evident that all four classifiers are capable of distinguishing classes. From the models, the SVM seems to have the highest classification performance among any of the other models. Therefore, it can be concluded that SVM is the most suitable model for this experiment.

It provides a compromise between model effectiveness and model training efficiency, and reduces computation time. Thus, the suggested model has a much lower computational complexity than existing methods like Ensemble and DL. It has a higher performance efficiency and is more favorable for use in real-time applications.

Table 3. Comparison Analysis

Ref.	Dataset	Algorithm	Accuracy (%)
[18]	Cleveland dataset	AdaBoost	72.2
[19]	Cleveland dataset	KNN	88
[20]	Cleveland dataset	Random Forest	84.4
This work	Cleveland dataset	SVM	93.88

Table 2 indicates that SVM achieved 94.8% accuracy, while other topologies such as AdaBoost [18], KNN [19], and Random Forest [20] attained an accuracy of about 72.2%, 88% and 84.4% on the same dataset. Thus, it is demonstrated that the SVM reveals better performance than other ML algorithms.

5. CONCLUSION

This work provides a validation of ML algorithms in predicting heart disease using the Cleveland dataset. The implementation of feature selection methods (Chi-square and K-Best) improves model transparency. The experimental results clearly state that SVM exhibits better performance than all other classifiers with the highest accuracy (93.88%), recall (92.75%), precision (94.81%) and an F1-score (93.78). The kNN algorithm achieved an accuracy of 91.76%, but it exhibits low performance over noisy data. Even though DT exhibits 90.64% accuracy, it has overfitting issues. LR achieved 89.14% accuracy. Overall, the results validate that the adoption of SVM in a clinical decision-making framework provides better predictive results over real-time applications. In the future, it can be validated for larger data sets and in multiple sub-classes, which can lead to improved model reliability

REFERENCES

[1] World Health Organization. World Health Statistics 2021; World Health Organization: Geneva, Switzerland, 2021.

[2] Bhatt, Chintan M., Parth Patel, TarangGhetia, and Pier Luigi Mazzeo. "Effective heart disease prediction using machine learning techniques." *Algorithms* 16, no. 2 (2023): 88.

[3] Ternacle, Julien, Nancy Côté, Laura Krapf, Annabelle Nguyen, Marie-AnnickClavel, and Philippe Pibarot. "Chronic kidney disease and the pathophysiology of valvular heart disease." *Canadian Journal of Cardiology* 35, no. 9 (2019): 1195-1207.

[4] Deepan, P., R. Vidhya, B. Rajalingam, R. Santhoshkumar, and N. Arul. "FLAML-HDPS Model: An Efficient and Intelligent AutoML Approach for Heart Disease Prediction." In *International Conference on Computer & Communication Technologies*, pp. 287-296. Singapore: Springer Nature Singapore, 2023.

[5] Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." *SN Computer Science* 1, no. 6 (2020): 345.

[6] Celermajer, David S., Clara K. Chow, Eloi Marijon, Nicholas M. Anstey, and Kam S. Woo. "Cardiovascular disease in the developing world: prevalences, patterns, and the potential of early disease detection." *Journal of the American College of Cardiology* 60, no. 14 (2012): 1207-1216.

[7] Chandrasekhar, Nadikatla, and SamineniPeddakrishna. "Enhancing heart disease prediction accuracy through machine learning techniques and optimization." *Processes* 11, no. 4 (2023): 1210.

[8] Gao, Xiao-Yan, Abdelmegeid Amin Ali, Hassan Shaban Hassan, and Eman M. Anwar. "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method." *Complexity* 2021, no. 1 (2021): 6663455.

[9] Hassan, Diman, Haval I. Hussein, and Masoud M. Hassan. "Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis." *Biomedical signal processing and control* 79 (2023): 104019.

[10] Ahmad, Ahmad Ayid, and HuseyinPolat. "Prediction of heart disease based on machine learning using jellyfish optimization algorithm." Diagnostics 13, no. 14 (2023): 2392.

[11] Khan, Muhammad Amir, TehseenMazhar, Muhammad MateenYaqoob, Muhammad Badruddin Khan, Abdul KhaderJilaniSaudagar, YazeedYasinGhadi, Umar Farooq Khattak, and Mohammad Shahid. "Optimal feature selection for heart disease prediction using modified Artificial Bee colony (M-ABC) and K-nearest neighbors (KNN)." *Scientific Reports* 14, no. 1 (2024): 26241.

[12] Truong, Vien T., Binh P. Nguyen, Thanh-Hoang Nguyen-Vo, Wojciech Mazur, Eugene S. Chung, Cassady Palmer, Justin T. Tretter et al. "Application of machine learning in screening for congenital heart diseases using fetal echocardiography." *The International Journal of Cardiovascular Imaging* 38, no. 5 (2022): 1007-1015.

[13] Gupta, C., Saha, A., Reddy, N. & Acharya, U. Cardiac Disease Prediction using Supervised Machine Learning Techniques. J. Phys: Conf. Ser. 2161, 1–12 (2022).

[14] Subramani, S., Varshney, N., Anand, M.V., Soudagar, M.E.M., Al-Keridis, L.A., Upadhyay, T.K., Alshammari, N., Saeed, M., Subramanian, K., Anbarasu, K. and Rohini, K. Cardiovascular diseases prediction by machine learning incorporation with deep learning. Frontiers in medicine, 10, p.1150933,2023.

[15] Krittanawong, C., Virk, H.U.H., Bangalore, S., Wang, Z., Johnson, K.W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., Kitai, T. and Baber, U. Machine learning prediction in cardiovascular diseases: a meta-analysis. Scientific reports, 10(1), p.16057, 2020.

[16] Bouqentar, M.A., Terrada, O., Hamida, S., Saleh, S., Lamrani, D., Cherradi, B. and Raihani, A. Early heart disease prediction using feature engineering and machine learning algorithms. Heliyon, 10(19), 2024.

[17] Al-Mahdi, I.S., Darwish, S.M. and Madbouly, M.M. Heart Disease Prediction Model Using Feature Selection and Ensemble Deep Learning with Optimized Weight. CMES-Computer Modeling in Engineering and Sciences, 143(1), pp.875–909, 2025.

[18] Terrada, O., Cherradi, B., Hamida, S., Raihani, A., Moujahid, H. and Bouattane, O. Prediction of patients with heart disease using artificial neural network and adaptive boosting techniques. 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet), pp.1–6. IEEE, 2020.

[19] Terrada, O., Cherradi, B., Raihani, A. and Bouattane, O. Atherosclerosis disease prediction using supervised machine learning techniques. 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), pp.1–5. IEEE, 2020.

[20] Tougui, I., Jilbab, A. and El Mhamdi, J. Heart disease classification using data mining tools and machine learning techniques. Health and Technology, 10(5), pp.1137–1144, 2020.