Review Article

ACCENTS Transactions on Information Security, Vol 9(34)
ISSN (Online): 2455-7196
http://dx.doi.org/10.19101/TIS.2023.829004

A hybrid decision tree and support vector machine approach for heart disease classification

Mukesh Kumar[1*] and Mohan Kumar Patel[2]
Department of Computer Science and Engineering, Patel College of Science & Technology, Bhopal, Madhya Pradesh, India[1]
Assistant professor, Department of Computer Science and Engineering, Patel College of Science & Technology, Bhopal, Madhya Pradesh, India[2]

Received: 20-October-2023; Revised: 10-December-2023; Accepted: 20-January-2024

Abstract
Heart disease remains a leading cause of morbidity and mortality worldwide, necessitating accurate and early diagnostic methods. This study proposes a hybrid model combining decision trees (DT) and support vector machines (SVM) to enhance heart disease classification. The hybrid DT-SVM model leverages DT's interpretability and SVM's accuracy, processing a comprehensive dataset from the UCI machine learning repository. Data preprocessing, including feature selection and scaling, ensures quality inputs for model training. The DT segments the data hierarchically, while SVM classifiers handle non-linear patterns within each segment. The model's performance, validated through k-fold cross-validation and metrics such as precision, recall, F1-score, and accuracy, demonstrates superior predictive capabilities. The hybrid approach consistently outperforms traditional models, achieving an accuracy of 98%, indicating its potential in classification to improve patient outcomes.

Keywords
Heart disease, Machine learning, Decision tree, Support vector machine, Hybrid model.

1.Introduction
The prevalence of heart disease, a leading cause of morbidity and mortality worldwide, indicates the need for accurate and early diagnosis to improve patient outcomes [1, 2]. Heart disease encompasses a range of conditions affecting the heart, including coronary artery disease, heart failure, and arrhythmias [3, 4]. The complexity and variability of these conditions pose significant challenges for traditional diagnostic methods [5, 6].

Recent advancements in advanced computational learning approaches offer transformative potential in this domain. Data mining involves extracting valuable insights and patterns from large datasets, a process known as knowledge discovery in databases (KDD) [7–10]. This includes data cleaning, pattern recognition, visualization, and rule generation.

Advanced computational techniques such as linear regression, logistic regression, support vector machines (SVM), naïve Bayes (NB), decision trees (DT), k-nearest neighbors (kNN), clustering methods, random forests (RF), and association rule mining (Apriori) have demonstrated remarkable capabilities in analyzing complex and voluminous datasets [11–14]. These methods can uncover meaningful patterns and make highly accurate predictions, often surpassing the diagnostic capabilities of traditional methods [12–16]. The integration of data-driven approaches in medical diagnostics aims to enhance accuracy, objectivity, and the utilization of extensive health data to enable personalized medicine.

The main objective of this paper is to apply the hybridization of decision trees with support vector machines (DT-SVM) for the classification of heart disease data. This hybrid approach leverages the strengths of both methods: decision trees' interpretability and SVM's high accuracy. By combining these techniques, the hybrid model aims to enhance the predictive performance and reliability in diagnosing heart disease. The interpretability of decision trees helps in understanding the decision-

*Author for correspondence

making process, while SVM's robustness contributes to improved accuracy, making this method particularly effective for medical diagnostics. This approach holds promise for better clinical decision support and improved patient outcomes.

The rest of the paper is organized as follows. Section 2 discusses the literature review. Section 3 covers the methods used and the dataset for experimentation. Results are illustrated in Section 4. Finally, conclusions are presented in Section 5.

2.Literature review

In this section our focus is to discuss and analyzes the related work to explore the methodological interventions along with the advantages and disadvantages.In 2023, Singh et al. [17] discussed the global health concern of cardiovascular disease and proposed machine learning models for early detection and precise prediction. Using classifiers like DT, NB, multilayer perceptron (MLP), and logistic regression on patient data, the DT achieved the highest accuracy of 98.04%, followed by MLP at 95.51%.

In 2023, Lakshmi and Devi [18] highlighted cardiovascular disease as the leading global cause of death. Their study used the Framingham dataset from Kaggle, applying an enhanced whale optimization algorithm for feature selection. They implemented and evaluated various machine learning classifiers on the dataset, focusing on accuracy, precision, recall, and F1-score for efficient heart disease prediction. In 2023, Hemalatha et al. [19] emphasized on the critical importance of accurately diagnosing heart disease to prevent serious health issues. The study compares and analyzes cardiovascular illness using various machine learning and deep learning methods, indicating the potential of these technologies in improving diagnostic accuracy and patient outcomes. In 2020, El et al. [20] addressed heart disease, a major global health concern, by utilizing machine learning models. The study employed stacking ensemble learning and K-fold validation with DT, kNN, and SVM. The RF model achieved the highest accuracy of 99.02%, highlighting the potential in early heart disease detection.

In 2023, Rustagi and Vijarania [21] explored using machine learning to predict coronary artery diseases and chronic heart failure, major causes of heart attacks. With rising global casualties, reliable systems are needed. They applied techniques like SVM, ANN, and RF to medical datasets, aiming to improve diagnosis and prediction of heart conditions, aiding clinicians in decision-making.

In 2023, Srivastava et al. [22] addressed the "black box" stigma in machine learning, which hinders its acceptance in medicine. They explored explainable machine learning, crucial for decisions like ICU admissions. Using boosting methods for heart disease detection, they highlighted how understanding model decisions can improve clinical reasoning and enhance prediction accuracy.

In 2023, Jadhav et al. [23] highlighted the rising deaths due to heart disease, affecting over 26 million people globally. They emphasized the need for accurate and timely diagnosis. The paper reviews recent research on using data mining and machine learning techniques to extract valuable insights from unstructured healthcare data for heart disease prediction.

In 2024, Tyagi and Jain [24] emphasized heart disease as a leading cause of death, noting the challenges of clinical data analysis in its prediction. They emphasized the efficacy of machine learning algorithms in making accurate predictions, aiding early diagnosis, and mitigating heart disease effects. Their study explores innovative learning approaches to enhance prediction accuracy by identifying critical features and employing various classification techniques. In 2023, Selvakumar et al. [25] highlighted the critical need for early detection and treatment of heart attacks to reduce mortality rates. Using machine learning (ML) techniques, they aimed to predict heart attack risks by analyzing features such as age, gender, and cholesterol levels. Their predictive model, trained on various datasets, seeks to enhance the accuracy of heart attack risk prediction, and ultimately reduce heart attack-related deaths.

In 2023, Aburayya et al. [26] addressed the rising global issue of heart disease and emphasized the need for early diagnosis. They developed a machine learning-based system to predict heart disease, considering factors like age, chest discomfort, blood pressure, gender, cholesterol, and heartbeat. Using algorithms like logistic regression, kNN, DT, NB, RF, and SVM, they employed cross-validation, feature selection, and metrics like accuracy, specificity, and sensitivity. Their system effectively distinguishes between healthy individuals and those with heart disease, enhancing diagnostic accuracy.

In 2024, Dibaji and Sulaimany [27] introduced a novel approach to enhance heart disease classification accuracy by integrating graph-based techniques and community detection with machine learning. Using community detection to capture complex relationships within datasets and one-hot encoding to enrich features, they achieved a 94% accuracy. This study highlights the potential of graph-based methods in improving heart disease classification and provides insights into future research opportunities and limitations.

3.Methods

In this paper a hybrid algorithm combining DT with SVM (DT-SVM) for the heart disease classification has been presented. This hybrid model leverages the hierarchical data partitioning capability of DT, which helps in managing the complexity and variability of medical data. Each segment created by the tree is then handled by a specifically trained SVM, allowing for detailed and nuanced classification based on the characteristics of that segment. This not only improves classification accuracy but also handles non-linear patterns effectively. The combination thus provides a powerful tool for medical diagnostics, capable of adapting to varied and complex datasets typically in heart disease cases.

In the hybrid DT-SVM model for heart disease classification, the process starts with data preprocessing, which includes dataset selection (*Figure 1*). For our approach, the Statlog and Cleveland heart disease datasets from the UCI machine learning repository were considered [28, 29]. The Statlog dataset contains 270 records, while the Cleveland dataset includes 303 records. Both datasets comprise male and female patients. Each dataset has 270 samples characterized by 13 attributes, with the class distribution being the 14th attribute. The dataset is then cleaned to handle missing values and outliers, improving its quality. Finally, the feature values are scaled to prepare for SVM processing.

DT construction starts with building a DT using Gini index to choose the best splits. This tree is not fully developed; instead, growth is stopped prematurely (pre-pruning) to leave broader segments. Parameters are set to define the maximum depth of the tree and the minimum samples at leaf nodes, ensuring the subsets are substantial enough for effective SVM classification. SVM integration at leaves, an SVM model is trained for each leaf of the DT. The SVMs are trained exclusively on the data points that reach each leaf. An appropriate SVM kernel, such as linear, polynomial, or radial basis function, is selected based on the nature of data segmentation and distribution at each leaf. SVM parameters like the regularization parameter (C) and the kernel coefficient (gamma) are optimized using grid search with cross-validation within each subset. Model Validation involves performing k-fold cross-validation to test the overall model's performance, assessing both the DT's segmentation ability and the SVM's classification power at each segment. Model optimization was performed based on a post-training reassessment of the impact of different features on the model's performance to refine the feature set. The tree and SVM parameters are continuously tuned based on performance feedback. The model is regularly updated and monitored to adapt to new data and emerging trends in heart disease diagnosis, ensuring its ongoing relevance and effectiveness in real-world applications.

Enhanced Hybrid DT-SVM Algorithm

Step 1: Dataset selection.

Step 2: Cleanse the data by removing missing values and outliers to enhance the quality.

Step 3: Select features using principal component analysis (PCA) where the selection criterion is based on eigenvalue analysis: $\lambda_i>1$ where λ_i is the eigenvalue of the i^{th} component.

Step 4: Standardize features to have zero mean and unit variance (Equation 1):

$$z = \frac{(x-\mu)}{\sigma} \quad (1)$$

where x is the feature value, μ is the mean, and σ is the standard deviation.

Step 5: Construct a DT using the Gini index to determine the best splits:

$$G = 1-\sum_{i=1}^{n} p_i^2$$

where p_i is the probability of class i at a given node.

Step 6: Define tree parameters such as maximum depth and minimum samples per leaf, ensuring effective segmentation.

Step 7: Train an SVM at each leaf, applying the data subset from the leaf.

Step 8: Choose an SVM kernel based on the segmented data's characteristics. Common choices include (Equation 2, 3 and 4):

Linear:

$$K(x, x') = x.x' \quad (2)$$

Polynomial:

$$K(x, x') = (\text{gamma}.\ x.x' + \text{coef0})^{\text{degree}} \quad (3)$$

Radial Basis Function (RBF):

$$K(x, x') = \exp(-\text{gamma}.\ ||x-x'||)^2 \quad (4)$$

Step 9: Optimize C and gamma using grid search. C controls the trade-off between achieving a low error

on the training data and minimizing the model complexity for better generalization. Gamma in kernels like RBF affects the influence of individual training examples on the learning process.
Step 10: Combine the DT's and trained SVMs into a cohesive model framework.
Step 11: Perform k-fold cross-validation, partitioning the data into k equally sized segments and rotating each as a validation set while others form a training set.
Step 12: Analyze feature importance and model parameters continuously, refining the setup to maximize performance.
Step 13: Update the feature vector by comparing each iteration.
Step 14: End.

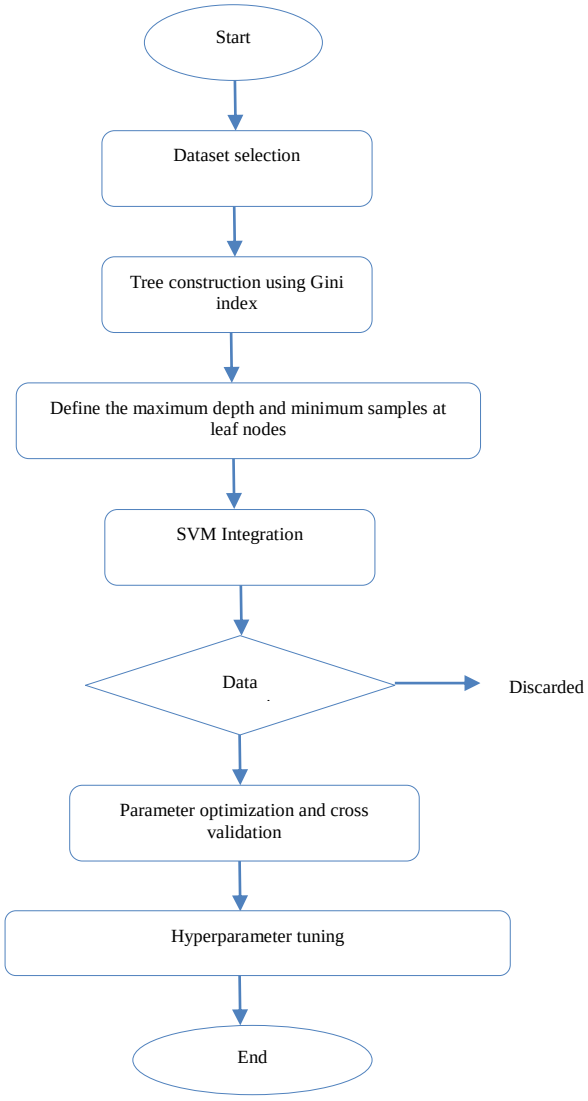

Figure 1 Flowchart depicting the process of DT-SVM algorithm

4.Results

Figure 2 shows the performance of various machine learning models—kNN, SVM, DT, and a hybrid SVM-DT—across different epochs in case of Statlog dataset. The best-performing model is the SVM-DT, consistently achieving the highest accuracy, peaking at 0.97 at 500 epochs, due to its combined strengths of SVM's accuracy and DT's interpretability. Conversely, kNN demonstrates the poorest performance, particularly at higher epochs (300 and 400), with accuracies dropping to 0.63, due to its sensitivity and high-dimensional data. SVM and DT show improvement over epochs but do not match the robustness of the SVM-DT hybrid.

Figure 3 compares the performance of kNN, SVM, DT, and a hybrid SVM-DT model based on precision, recall, F1-score, and accuracy in case of Statlog dataset. The SVM-DT hybrid model performs the best across all metrics, achieving a precision of 0.98, recall of 0.97, F1-score of 0.96, and accuracy of 0.98, indicating its superior capability in accurately predicting heart disease and managing both false positives and false negatives effectively. In contrast, kNN shows the worst performance with the lowest accuracy (0.78), due to its sensitivity and less effective handling of complex data structures. SVM and DT perform well but are outperformed by the hybrid model.

Figure 4 illustrates the accuracy of kNN, SVM, DT, and SVM-DT models across different epochs in case of Cleveland dataset. The SVM-DT hybrid consistently demonstrates superior performance, reaching the highest accuracy of 0.98 at 500 epochs. This reflects its ability to effectively combine SVM's precision with DT's interpretability. On the other hand, kNN exhibits the poorest performance, particularly at 400 epochs with an accuracy of 0.71, due to its sensitivity and high-dimensional data. While SVM and DT perform well, showing improvements with more epochs, they are outperformed by the robust hybrid SVM-DT model.

Figure 5 compares kNN, SVM, DT, and SVM-DT models based on precision, recall, F1-score, and accuracy in case of Cleveland dataset. The SVM-DT hybrid model outperforms the others with the highest scores across all metrics: precision (0.98), recall (0.97), F1-score (0.97), and accuracy (0.98). This highlights its robust predictive capability and effective handling of both false positives and false negatives. In contrast, kNN performs the worst, with a notably lower F1-score (0.81) and accuracy (0.86),

due to its sensitivity to data variability. SVM and DT show good performance but are surpassed by the hybrid model's superior results.

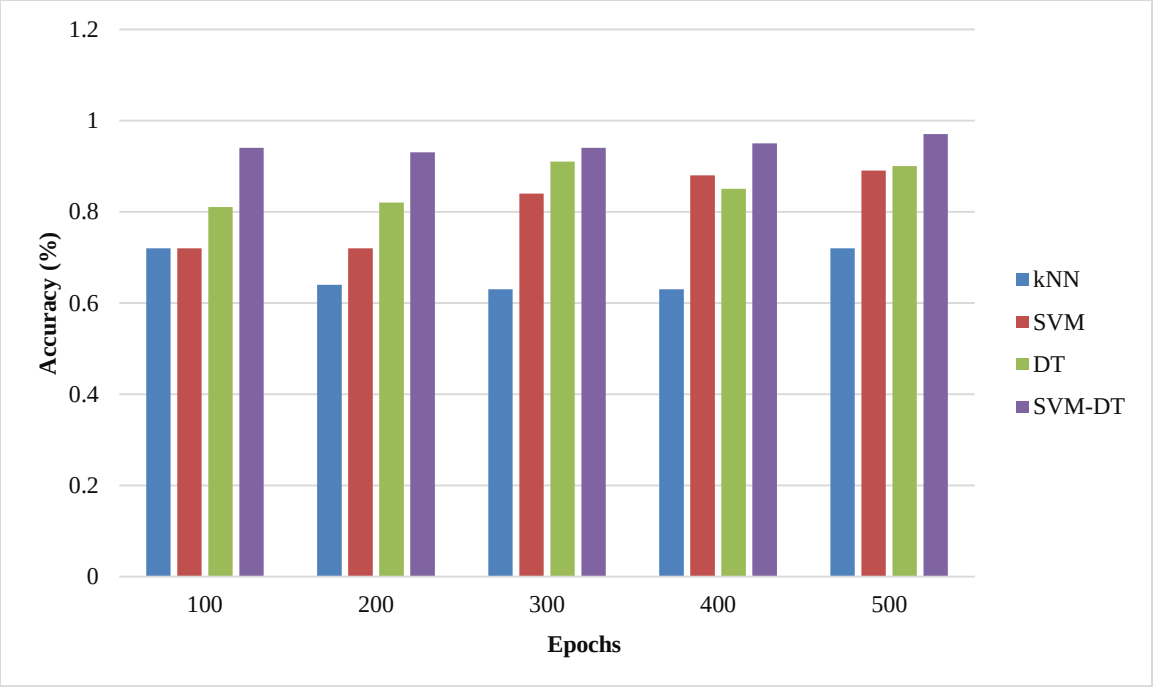

Figure 2 Performance of various machine learning models—kNN, SVM, DT, and a hybrid SVM-DT—across different epochs in case of Statlog dataset

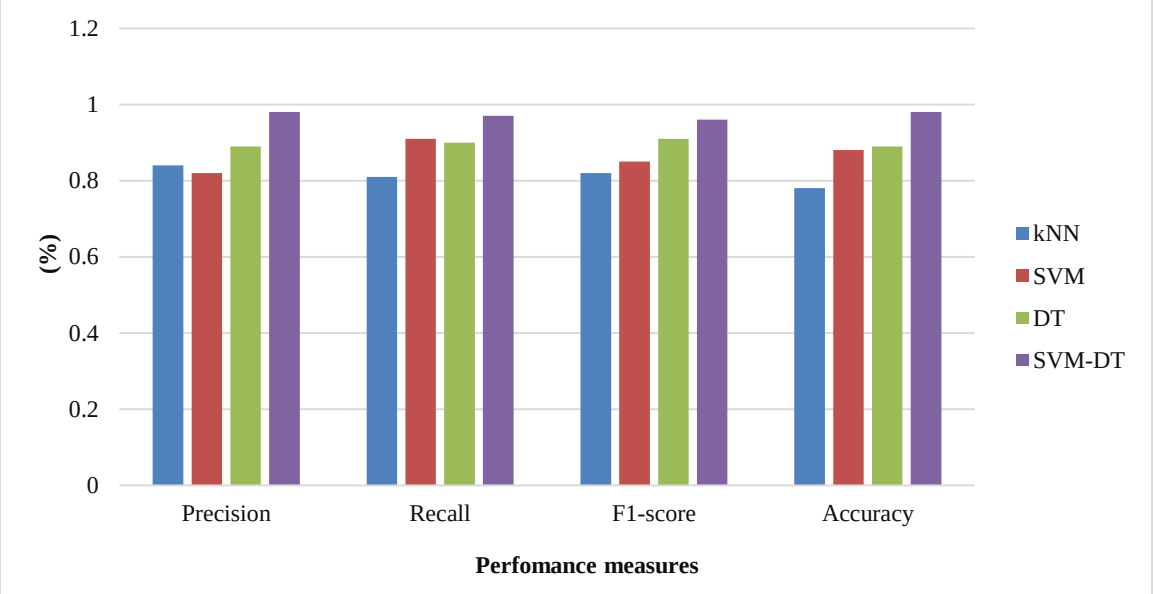

Figure 3 Performance of kNN, SVM, DT, and a hybrid SVM-DT model based on precision, recall, F1-score, and accuracy in case of Statlog dataset

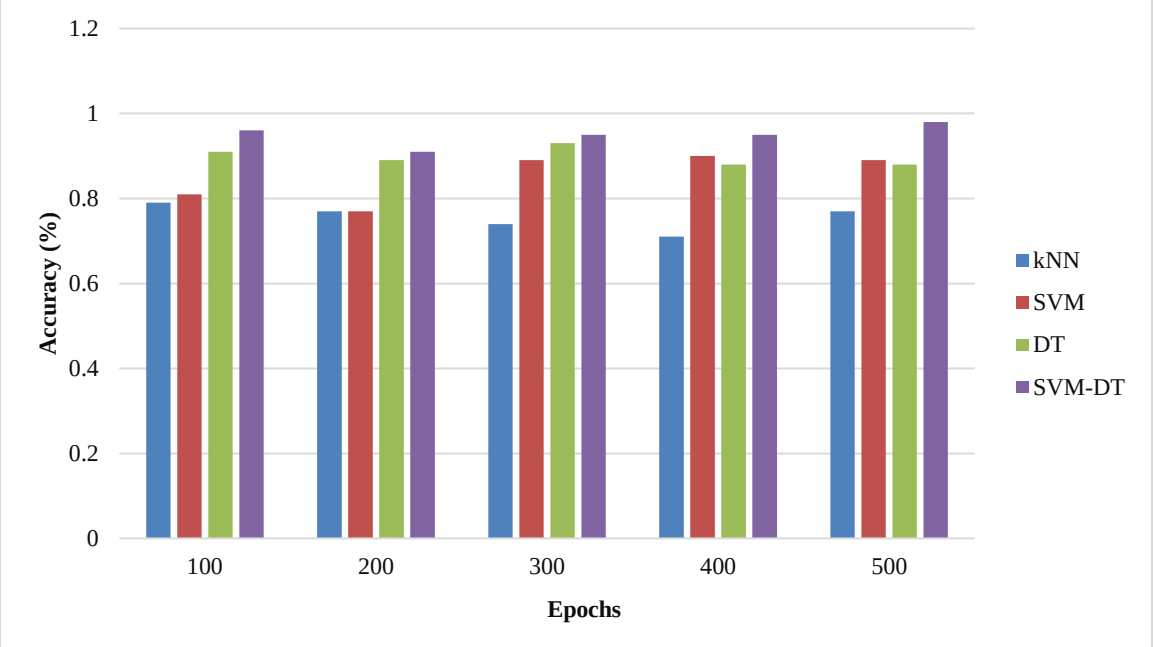

Figure 4 Accuracy of kNN, SVM, DT, and SVM-DT models across different epochs in case of Cleveland dataset

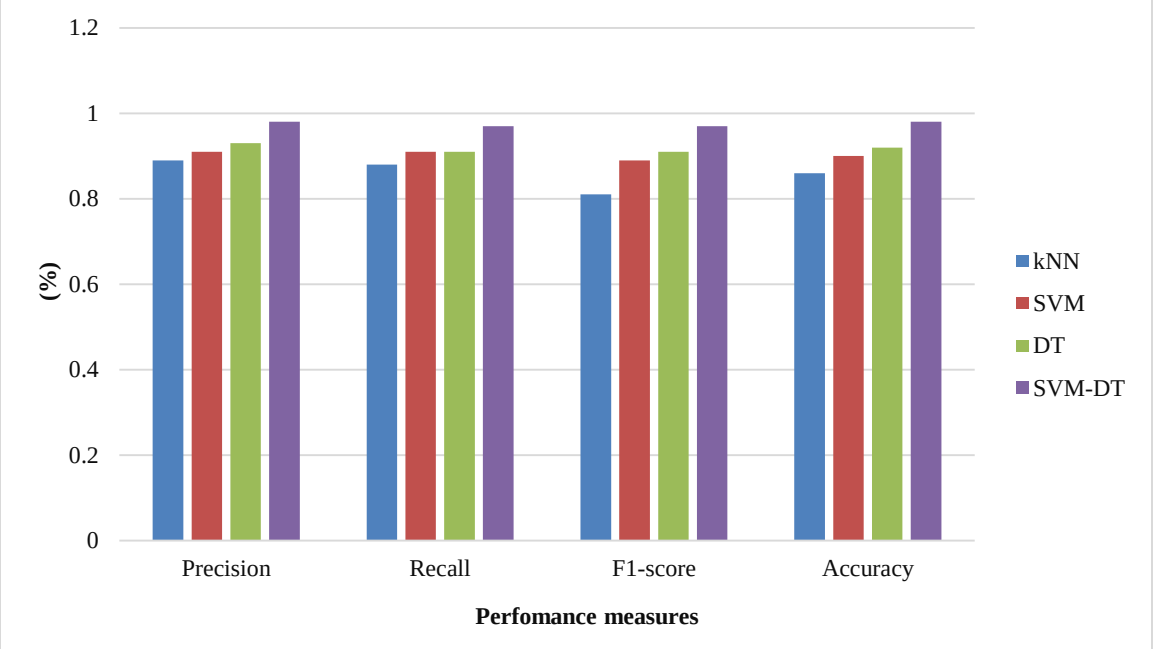

Figure 5 Comparison of kNN, SVM, DT, and SVM-DT models based on precision, recall, F1-score, and accuracy in case of Cleveland dataset

5.Conclusion

The study effectively demonstrates the superiority of the hybrid DT-SVM model in accurately classifying heart disease. By integrating the interpretability of DTs with SVMs, the hybrid approach provides a robust tool for medical diagnostics. The extensive analysis of various models highlights the DT-SVM hybrid's ability to handle complex and voluminous datasets, achieving an accuracy of 98%. This method not only enhances diagnostic accuracy but also offers insights into the decision-making process. The results suggest that the DT-SVM hybrid model

outperformed other algorithms. Future research should focus on further refining the model, exploring additional features, and testing its applicability in real-world clinical settings to fully harness its potential.

Acknowledgment
None.

Conflicts of interest
The authors have no conflicts of interest to declare.

References
[1] Ma CY, Luo YM, Zhang TY, Hao YD, Xie XQ, Liu XW, et al. Predicting coronary heart disease in Chinese diabetics using machine learning. Computers in Biology and Medicine. 2024; 169:107952.
[2] Rani P, Kumar R, Jain A, Lamba R, Sachdeva RK, Kumar K, et al. An extensive review of machine learning and deep learning techniques on heart disease classification and prediction. Archives of Computational Methods in Engineering. 2024:1-9.
[3] Griffeth EM, Stephens EH, Dearani JA, Shreve JT, O'Sullivan D, Egbe AC, et al. Impact of heart failure on reoperation in adult congenital heart disease: an innovative machine learning model. The Journal of Thoracic and Cardiovascular Surgery. 2024; 167(6):2215-25.
[4] Islam MA, Majumder MZ, Miah MS, Jannaty S. Precision healthcare: a deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction. Computers in Biology and Medicine. 2024; 176:108432.
[5] Hoque R, Billah M, Debnath A, Hossain SS, Sharif NB. Heart disease prediction using SVM. International Journal of Science and Research Archive. 2024; 11(2):412-20.
[6] Marelli AJ, Li C, Liu A, Nguyen H, Moroz H, Brophy JM, et al. Machine learning informed diagnosis for congenital heart disease in large claims data source. JACC: Advances. 2024; 3(2):100801.
[7] Dubey AK, Dubey AK, Agarwal V, Khandagre Y. Knowledge discovery with a subset-superset approach for mining heterogeneous data with dynamic support. In CSI sixth international conference on software engineering (CONSEG) 2012 (pp. 1-6). IEEE.
[8] Türk F. Investigation of machine learning algorithms on heart disease through dominant feature detection and feature selection. Signal, Image and Video Processing. 2024; 18(4):3943-55.
[9] Thangavel S, Selvaraj S, Keerthika K. Analyzing machine learning classifiers for the diagnosis of heart disease. EAI Endorsed Transactions on Pervasive Health and Technology. 2024; 10.
[10] Pachiyannan P, Alsulami M, Alsadie D, Saudagar AK, AlKhathami M, Poonia RC. A novel machine learning-based prediction method for early detection and diagnosis of congenital heart disease using ECG signal processing. Technologies. 2024; 12(1):4.
[11] Jandy K, Weichbroth P. A machine learning approach to classifying New York Heart Association (NYHA) heart failure. Scientific Reports. 2024; 14(1):11496.
[12] Singh P, Kourav PS, Mohapatra S, Kumar V, Panda SK. Human heart health prediction using GAIT parameters and machine learning model. Biomedical Signal Processing and Control. 2024; 88:105696.
[13] Chiriac A, Ngufor C, van Houten HK, Mwangi R, Madhavan M, Noseworthy PA, et al. Beyond atrial fibrillation: machine learning algorithm predicts stroke in adult patients with congenital heart disease. Mayo Clinic Proceedings: Digital Health. 2024; 2(1):92-103.
[14] Lakshmi A, Devi R. Heart disease prediction using ensemble feature selection method and machine learning classification algorithms. Conversational Artificial Intelligence. 2024:237-47.
[15] Ahmad N. Elevating E-healthcare: machine learning insights into heart disease identification. Journal Environmental Sciences and Technology. 2024; 3(1):195-206.
[16] Ogunpola A, Saeed F, Basurra S, Albarrak AM, Qasem SN. Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics. 2024; 14(2):144.
[17] Singh G, Guleria K, Sharma S. Machine learning and deep learning models for early detection of heart disease. In international conference on computing, communication, and intelligent systems (ICCCIS) 2023 (pp. 419-24). IEEE.
[18] Lakshmi A, Devi R. Heart disease prediction using enhanced whale optimization algorithm based feature selection with machine learning techniques. In 12th international conference on system modeling & advancement in research trends (SMART) 2023 (pp. 644-8). IEEE.
[19] Hemalatha S, Kavitha T, Chitra K, Babitha B, Kaya N, Keshika R. Extensive review on predicting heart disease using machine learning and deep learning techniques. In international conference on computer communication and informatics (ICCCI) 2023 (pp. 1-7). IEEE.
[20] El Hamdaoui H, Boujraf S, Chaoui NE, Maaroufi M. A clinical support system for prediction of heart disease using machine learning techniques. In 5th international conference on advanced technologies for signal and image processing (ATSIP) 2020 (pp. 1-5). IEEE.
[21] Rustagi T, Vijarania M. Extensive analysis of machine learning techniques in the field of heart disease. In 3rd international conference on technological advancements in computational sciences (ICTACS) 2023 (pp. 913-917). IEEE.
[22] Srivastava D, Pandey H, Agarwal AK, Sharma R. Opening the black box: explainable machine learning for heart disease patients. In international conference on advanced computing technologies and applications (ICACTA) 2023 (pp. 1-5). IEEE.
[23] Jadhav SR, Kulkarni R, Yendralwar A, Pujari P, Patwari S. Monitoring and predicting of heart diseases using machine learning techniques. In 8th

international conference for convergence in technology (I2CT) 2023 (pp. 1-4). IEEE.
[24] Tyagi N, Jain P. A review of machine learning algorithms for predicting heart disease. In 2nd international conference on disruptive technologies (ICDT) 2024 (pp. 961-5). IEEE.
[25] Selvakumar V, Achanta A, Sreeram N. Machine learning based chronic disease (Heart Attack) prediction. In international conference on innovative data communication technologies and application (ICIDCA) 2023 (pp. 1-6). IEEE.
[26] Aburayya RA, Alomar RA, Alnajjar DK, Athamnah S, Alquran H, Mustafa WA, et al. Automated heart diseases detection using machine learning approach. In 6th international conference on engineering technology and its applications (IICETA) 2023 (pp. 108-14). IEEE.
[27] Dibaji A, Sulaimany S. Community detection to improve machine learning based heart disease prediction. In 20th CSI international symposium on artificial intelligence and signal processing (AISP) 2024 (pp. 1-6). IEEE.
[28] https://archive.ics.uci.edu/dataset/145/statlog+heart. Accessed 11 March 2024.
[29] https://archive.ics.uci.edu/dataset/45/heart+disease. Accessed 17 March 2024.

Mukesh Kumar is currently pursuing an M.Tech in Software System at Patel College of Science and Technology, RGPV in Bhopal, Madhya Pradesh. He has completed his MCA. at Sikkim Manipal University. His areas of interest include Machine Learning.

Email: mukesh.patna05@gmail.com

Mohan Kumar Patel is working as Assistant professor with the department of Computer Science and Engineering at Patel College of Science & Technology, Bhopal, Madhya Pradesh, India. He has completed his Bachelor of Engineering and Master of Technology in Computer Science Engineering from RGPV Technical University Bhopal (M.P). He has 1 publication and conferences. His reserch area in network and network security, cryptography etc.
Email: patel.mohan67@gmail.com