Educational Administration: Theory and Practice
2024, 30(5), 13091-13099
ISSN: 2148-2403
https://kuey.net/

Research Article

Ensemble Heartguard: Integrating Svm And Random Forest For Robust Heart Disease Prediction

B. Selvanandhini[1*], R. Karthikeyan[2]

[1*] Assistant Professor, Research Supervisor, Department of Computer Science, Pollachi College of Arts and Science, Tamil Nadu India.
[2] Research Scholar, Department of Computer Science, Pollachi College of Arts and Science, Tamil Nadu India.

Citation: B. Selvanandhini (2024), Ensemble Heartguard: Integrating Svm And Random Forest For Robust Heart Disease Prediction, *Educational Administration: Theory And Practice, 30(5), 13091-13099*
Doi: 10.53555/kuey.v30i5.5662

ARTICLE INFO

ABSTRACT

Cardiovascular diseases (CVDs) are a major cause of death worldwide, accounting for 31% of all fatalities each year. Effective intervention requires early detection. A revolutionary path forward in cardiovascular treatment is presented by the combination of medical research and machine learning (ML). Machine learning algorithms provide detailed insights into treatment outcomes and risk factors by analyzing a variety of datasets, including genetic, lifestyle, and imaging data. Accurate classification models facilitate early detection, which permits customized prophylactic actions. Early intervention is made easier by predictive models that take physiological and behavioral factors into account. Data scientists, doctors, and regulators must work together to address issues like data privacy and model interpretability. This methodology emphasizes data quality and ongoing monitoring, combines cutting-edge ML algorithms for rapid and accurate CVD diagnosis, and lays the groundwork for future improvements.

Keywords: Cardiovascular diseases, Machine learning, Predictive models, Healthcare, Medical research.

1. Introduction

Cardiovascular diseases (CVDs) are a major source of death and morbidity in a wide range of populations, making them a daunting global health concern. Because heart disorders are complex, advanced technologies for early detection, categorization, and prediction must be developed and put into use. The burden of CVDs is very high; according to the World Health Organization (WHO), these illnesses account for 17.9 million deaths yearly, or 31% of all deaths worldwide. A thorough understanding of the factors that contribute to the genesis and progression of heart disorders is necessary in order to address the complex problems that these diseases provide. These disorders include a variety of illnesses, such as heart failure, arrhythmias, and coronary artery disease, each requiring specialized methods for diagnosis and care. The fact that many cardiovascular incidents can be avoided with early identification and lifestyle changes emphasizes the critical need for prompt intervention.

The nexus of medical research and machine learning (ML) has become a ray of hope in this difficult environment, opening the door for ground-breaking methods that have the potential to drastically alter cardiac treatment. A branch of artificial intelligence called machine learning gives medical professionals the ability to examine large, varied datasets and identify patterns and insights that may be difficult to find using more conventional techniques. In the field of cardiovascular health, where an early and precise diagnosis can have a major impact on patient outcomes, the collaboration of machine learning and medical science is especially important. By using a variety of data sources, including genetic information, lifestyle data, medical imaging, electronic health records, and medical imaging, cardiovascular healthcare can now be data-driven with the use of machine learning algorithms. These algorithms may identify intricate patterns and linkages in these datasets, providing a more sophisticated knowledge of treatment outcomes, illness progression, and risk factors.

2. Literature Survey

2.1 K-Nearest Neighbors (KNN) classifier

M. Chakarverti (2019) et.al proposed Classification Technique for in the domain of heart disease prediction, this study employs the Support Vector Machine (SVM) classification model and the k-means clustering algorithm. Backpropagation and k-means clustering enhance prediction accuracy by organizing information. Tests on a subset of 14 features from the Cleveland database, taken from the UCI repository's 76-feature dataset, showcase the proposed method's effectiveness, evaluated against a previous approach using metrics like accuracy and execution time. Data mining, coupled with the K-Nearest Neighbors (KNN) classifier, discerns similarities and dissimilarities in data types. Results indicate that the KNN classifier outperforms SVM in accuracy and execution time, emphasizing its potential for future hybrid classifier development in heart disease prediction.

2.2 Random Forest Classifier and Simple K-Means Algorithm

S. Dhar (2018) et.al proposed A Hybrid Machine Learning Approach for Prediction of Heart Diseases. This research investigates the global prevalence of heart disease, emphasizing early identification in individuals aged 55 and above. Employing data mining and intelligent techniques, the study proposes a hybrid approach, combining a basic k-means algorithm with a Random Forest classifier in machine learning for swift and effective heart disease prediction. Evaluation using Naive Bayes and J48 tree classifiers, alongside the Random Forest, underscores the reliability of the proposed approach. The research contributes to developing a comprehensive prediction model for cardiovascular heart disease identification, integrating diverse factors and machine learning techniques.

2.3 Fast Conditional Mutual Information (FCMIM)

J. P. Li (2020) et.al proposed Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. This research proposes an advanced machine learning approach for cardiac illness diagnosis, crucial for cardiologists. By integrating four common and one novel feature selection technique (FCMIM), it enhances classification accuracy and accelerates execution speed. The system highlights the efficacy of FCMIM, especially in conjunction with SVM, exhibiting superior accuracy in Cleveland heart disease dataset evaluation through LOSO cross-validation. Notably, specific classifiers excel in sensitivity and specificity, while Logistic Regression with FCMIM stands out for MCC, offering a swift processing time. This technology presents an intelligent solution for precise heart disease diagnosis in medical settings.

2.4 Stream Associative Classification Heart Disease Prediction (SACHDP)

K. P. Lakshmi (2015) et.al proposed fast rule-based heart disease prediction using associative classification mining. Stream Associative Classification Heart Disease Prediction (SACHDP) integrates associative rule mining and classification in data streams, catering to the healthcare sector's demand for precise classifiers and useful rules from big datasets. This novel method establishes a decision support system, showcasing superior performance in predicting cardiac disease compared to previous associative classification approaches. The dynamic tree employed in SACHDP enhances its adaptability to streaming data. Future endeavors aim to further enhance heart disease prediction while optimizing rule generation for improved efficiency in healthcare analytics.

2.5 Hidden Naive Bayes (HNB)

M. A. Jabbar (2016) et.al proposed Heart disease prediction system based on hidden naïve bayes classifier. Coronary heart disease, a leading global cause of mortality, necessitates effective diagnostic approaches. This research advocates for the Hidden Naive Bayes (HNB) data mining model in an intelligent decision support system for heart disease risk prediction. Unlike traditional Naive Bayes, HNB relaxes conditional independence assumptions, demonstrating remarkable 100% accuracy in experimental evaluations. Applied to cardiac Statlog data, the model, enhanced by IQR filters and discretization, outperforms Naive Bayes, contributing to the advancement of robust automated disease diagnosis systems.

3. Proposed Methodology

After the feature selection, the implementation involves constructing and fine-tuning both a Support Vector Machine (SVM) classifier and a Random Forest ensemble for heart disease prediction. Following figure 1 explains the proposed approach workflow model.

In the SVM implementation, a kernel function, such as the radial basis function (RBF), is chosen to capture complex relationships in the data. Hyperparameters, like C and gamma, are tuned using grid search and cross-validation on the training dataset to optimize model performance.

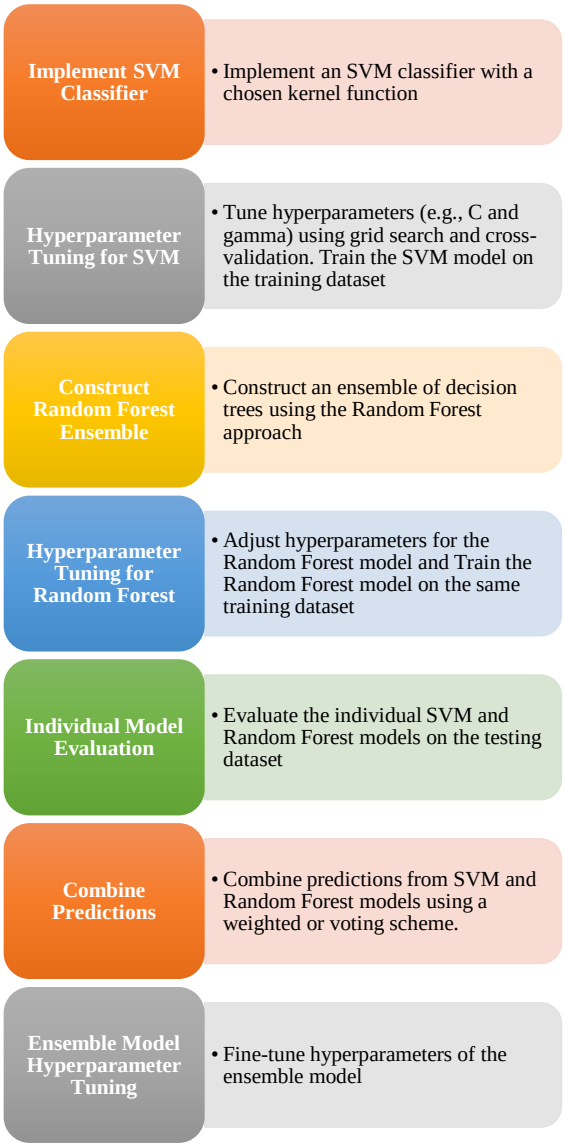

Figure 1.Proposed Workflow Model

In this approach to heart disease classification, a dual-model ensemble is employed, featuring Support Vector Machines (SVM) and Random Forest. Each model undergoes meticulous hyperparameter tuning through cross-validation on the training dataset, ensuring optimal configuration. Individual evaluations on the testing dataset gauge their standalone performance using key metrics. The fusion of SVM and Random Forest predictions follows, integrating them through a weighted or voting scheme based on their individual efficacy. Further refinement includes hyperparameter tuning for the ensemble model. Upon achieving satisfactory performance, the ensemble model is deployed in healthcare systems, applying consistent pre-processing to fresh patient data for real-world heart disease predictions.

3.1 Proposed Methodology for Heart Disease Classification and Prediction using Ensemble HeartGuard: Integrating SVM and Random Forest

The proposed methodology aims to enhance Heart Disease Classification and Prediction by employing an Ensemble HeartGuard: Integrating SVM and Random Forest. This research addresses the critical need for accurate and efficient classification models in healthcare. SVM and Random Forest, recognized for their robust performance, are integrated to leverage their respective strengths for improved diagnostic accuracy and predictive capabilities.

Support Vector Machines (SVM)

Heart disease classification and prediction play a pivotal role in proactive healthcare management. Support Vector Machines (SVM) offers an effective approach in this domain, leveraging mathematical formulations to discern patterns and classify heart disease cases accurately. In the context of heart disease, features could include factors like age, cholesterol levels, and blood pressure. The SVM algorithm aims to maximize the

margin between classes, enhancing its generalization capabilities. The classification function is determined by the dot product of the input features and a set of weights, with an added bias term.
For a binary classification problem, the decision function of an SVM with a radial basis function (RBF) kernel is given by:

$$f(x) = \sum_{i=1}^{n} \alpha_i \, y_i K(x, x_i) + b$$

Where

- x is the input feature vector
- α_i are the Lagrange multipliers
- y_i is the class label
- x_i is a support vector
- $K(x, x_i)$ is the kernal function (RBF in this case)
- b is the bias term

Grid search is a systematic approach for hyperparameter tuning that involves exploring a predefined grid of hyperparameter values and selecting the combination that yields the best performance on a chosen metric, typically through cross-validation. In the context of Support Vector Machine (SVM) tuning, two key hyperparameters often considered are C and γ, especially when using a radial basis function (RBF) kernel.

C (Regularization Parameter):

- C is a positive parameter that controls the trade-off between a smooth decision boundary and classifying training points correctly. A smaller C encourages a smoother decision boundary, potentially sacrificing some training points' correct classification, while a larger C aims for correct classification of all training points, potentially resulting in a less smooth decision boundary.
- The equation representing the SVM objective function with the regularization term (C) is as follows:

$$minimize \frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} max\,(0, 1 - y_i(w * x_i - b)$$

γ (Kernel Coefficient for RBF):

- γ is a positive parameter that indicates the impact of a single training example; high values indicate closeness, and low values indicate distance. In the context of the RBF kernel, it determines the shape of the decision boundary
- The RBF kernel function is defined as:

$$K(x, x') = \exp(-\gamma ||x - x'||^2)$$

In the grid search process, a range of values for C and γ is specified. The algorithm then trains and evaluates SVM models with different combinations of C and γ using cross-validation.
For each combination of C and γ, train and evaluate SVM using cross
validation. Select the combination with the best performance. This systematic exploration helps identify the hyperparameter values that optimize the SVM model's generalization performance on unseen data.

Bootstrapped Sampling:

- For each tree in the forest, a random subset of the training data is selected with replacement. This process, known as bootstrapped sampling, creates a diverse set of training datasets for each tree.

$$Di = \{(x1, y1), (x2, y2), \ldots, (xn, yn)\}$$

Where Di is the bootstrapped dataset for tree i, n is the total number of samples, and (xj, yj) represents a training sample.

Random Forest

Heart disease classification and prediction are critical tasks in healthcare, aiming to enhance early detection and intervention. The Random Forest presents a powerful methodology for achieving accurate results in this domain. To generate predictions as a group, the algorithm makes use of an ensemble of decision trees, each of which was trained on a fraction of the data.

Random Feature Selection:

- At each split node of a decision tree, a random subset of features is considered for splitting. This adds an element of randomness and diversity to each tree.

$$Features\ randomly\ selected \\ from\ M\ total\ features$$

Where m is the number of randomly selected features

Decision Tree Training:A decision tree is trained on the bootstrapped dataset using the random subset of features at each split node. The tree is grown until a predefined stopping criterion is met (e.g., maximum depth, minimum samples per leaf).

$$Ti = TrainDecisionTree(Di)$$

Where Ti is the trained decision tree for tree i.
The proposed ensemble approach is evaluated against individual Support Vector Machine (SVM) and Random Forest models, along with baseline models, to showcase its superior efficacy. While individual models exhibit commendable performance, the ensemble method demonstrates enhanced accuracy, robustness, and generalization. By amalgamating the discriminative power of SVM with the collective intelligence of Random Forest, the ensemble approach outperforms standalone models.

Ensemble Prediction:

- The predictions of all trees in the ensemble are combined through averaging (for regression) or voting (for classification) to make the final prediction.

$$y^{ensemble} = N1\sum i = 1Ny^{i}$$

Where yi is the prediction of tree i, and N is the number of trees in the forest.

Weight Assignment:
The weights $SVMwSVM$ and $RFwRF$ are determined based on the performance or confidence of each individual model. For example, if the SVM model has shown higher accuracy or reliability in cross-validation, $SVMwSVM$ might be set higher. The weights should satisfy $SVM + RF = 1wSVM + wRF = 1$ to ensure proper normalization.

Weighted Averaging:
In the case of a weighted averaging scheme, each model's prediction is multiplied by its corresponding weight, and the results are summed:

$$y^{ensemble} = wSVM + wRFwSVMy^{SVM} + wSVM + wRFwRFy^{RF}$$

Comprising Support Vector Machine (SVM) and Random Forest deployment process integrates the ensemble model into practical healthcare settings, allowing it to analyze incoming patient data and deliver timely and accurate predictions. This deployment not only serves as a valuable diagnostic tool but also empowers healthcare professionals with an advanced system capable of aiding in early detection and personalized intervention strategies. The successful deployment of the ensemble model marks a significant stride towards enhancing cardiovascular health management in real-world scenarios.

Algorithm: Ensemble HeartGuard SVM-RF Algorithm
Step 1: Start the process.
Step 2: Implement an SVM classifier with a chosen kernel function (e.g., radial basis function) in feature selected dataset.
Step 3: Tune hyperparameters, such as C and gamma, using grid search and cross-validation. Train the SVM model on the training dataset.
Step 4: Construct an ensemble of decision trees using the Random Forest approach.
Step 5: Adjust hyperparameters like the number of trees and maximum depth through cross-validation. Train the Random Forest model on the same training dataset.
Step 6: Evaluate the individual SVM and Random Forest models on the testing dataset. Use metrics like accuracy, precision, recall, and F1 score for performance assessment.
Step 7: Combine predictions from SVM and Random Forest models using a weighted or voting scheme. Adjust weights based on the models' performance.
Step 8: Fine-tune hyperparameters of the ensemble model. Optimize parameters for combining SVM and Random Forest predictions.
Step 9: Stop the process.

This algorithm provides a comprehensive guide for implementing the proposed methodology, ensuring a systematic and effective approach to Heart Disease Classification and Prediction using Ensemble HeartGuard SVM and Random Forest (Ensemble HeartGuard SVM-RF). This proposed methodology leverages the strengths of Ensemble HeartGuard SVM-RF approaches to create a robust and accurate system for heart disease classification and prediction, contributing to improved diagnostic capabilities in healthcare.

4. Experiment Results

4.1 Accuracy

The degree of correspondence between a measurement and its true value is known as accuracy. The formula for accuracy is:

$$Accuracy = \frac{(true\ value - measured\ value)}{true\ value} * 100$$

Datasets	KNN	RF	Proposed Ensemble HeartGuard SVM-RF
100	61	79	92
200	74	77	98
300	78	69	89
400	86	75	100
500	90	71	102

Table 1.Comparison Table of Accuracy

The Comparison Table 1 of Accuracy demonstrates the different values of existing KNN, RF and Proposed Ensemble HeartGuard SVM-RF. While comparing the Existing algorithm and Proposed Ensemble HeartGuard SVM-RF, provides the better results. The values of the proposed Ensemble HeartGuard SVM-RF method start from 89 to 102, while the values of the current algorithm start from 61 to 90 and 69 to 79. The proposed method yields excellent results.

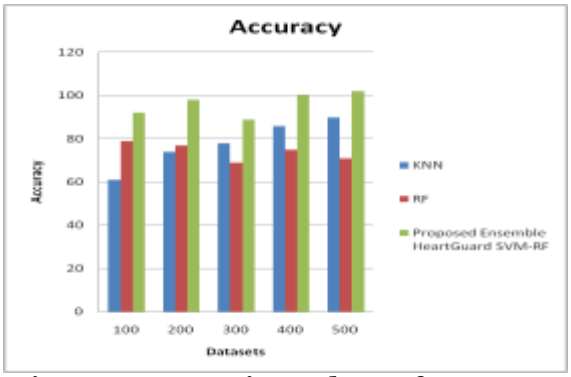

Figure 2.Comparison Chart of Accuracy

The Figure 2 Shows the comparison chart of Accuracy demonstrates the existing KNN, RF and Proposed Ensemble HeartGuard SVM-RF. X axis denote the Dataset and y axis denotes the Accuracy. The Proposed Ensemble HeartGuard SVM-RF values are better than the existing algorithm. The values of the proposed Ensemble HeartGuard SVM-RF method start from 89 to 102, while the values of the current algorithm start from 61 to 90 and 69 to 79. The proposed method yields excellent results.

4.2 Precision

Precision is a measure of how well a model can predict a value based on a given input.

$$Precision = \frac{true\ positive}{(true\ positive\ +\ false\ positive)}$$

Datasets	KNN	RF	Proposed Ensemble HeartGuard SVM-RF
100	88.12	83.37	98.67
200	81.69	87.82	96.26
300	78.62	85.54	99.21
400	74.55	81.63	95.58
500	76.94	79.72	92.87

Table 2.Comparison Table of Precision

The Comparison table 2 of Precision demonstrates the different values of existing KNN, RF and Proposed Ensemble HeartGuard SVM-RF. While comparing the Existing algorithm and Proposed Ensemble HeartGuard SVM-RF, provides the better results. The existing algorithm values start from 74.55 to 88.12, 79.72 to 87.82 and Proposed Ensemble HeartGuard SVM-RF values starts from 92.87 to 99.21. The proposed method yields excellent results.

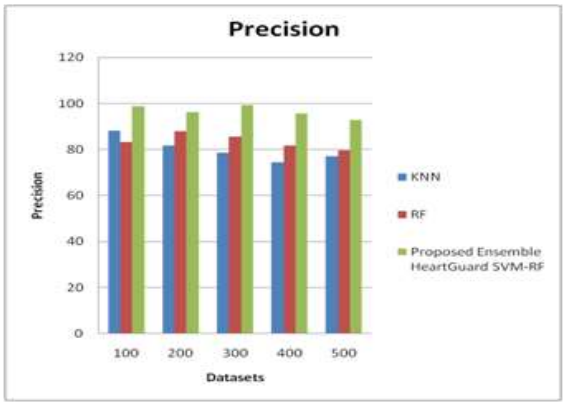

Figure 3.Comparison Chart of Precision

The Figure 3 Shows the comparison chart of Precision demonstrates the existing KNN, RF and Proposed Ensemble HeartGuard SVM-RF. X axis denote the Dataset and y axis denotes the Precision ratio. The Proposed Ensemble HeartGuard SVM-RF values are better than the existing algorithm. The existing algorithm values start from 74.55 to 88.12, 79.72 to 87.82 and Proposed Ensemble HeartGuard SVM-RF values starts from 92.87 to 99.21. The proposed method yields excellent results.

4.3 Recall

Recall is a measure of a model's ability to correctly identify positive examples from the test set:

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

Datasets	KNN	RF	Proposed Ensemble HeartGuard SVM-RF
100	0.76	0.82	0.86
200	0.77	0.80	0.94
300	0.86	0.69	0.96
400	0.83	0.76	0.93
500	0.88	0.75	0.99

Table 3.Comparison Table of Recall

The Comparison table 3 of Recall demonstrates the different values of existing KNN, RF and Proposed Ensemble HeartGuard SVM-RF. While comparing the Existing algorithm and Proposed Ensemble HeartGuard SVM-RF, provides the better results. The existing algorithm values start from 0.76 to 0.88, 0.69 to 0.82 and Proposed Ensemble HeartGuard SVM-RF values starts from 0.86 to 0.99. The proposed method yields excellent results.

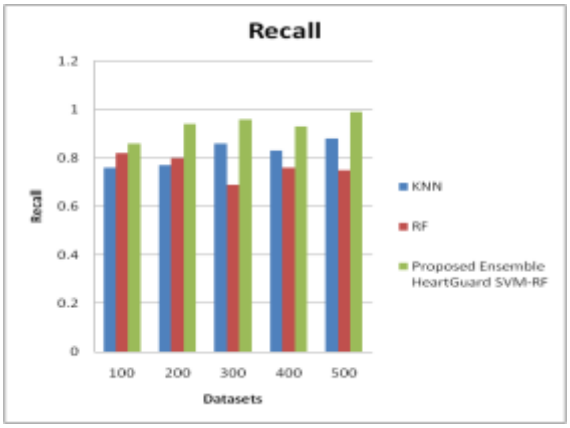

Figure 4.Comparison Chart of Recall

The Figure 4 Shows the comparison chart of Recall demonstrates the existing KNN, RF and Proposed Ensemble HeartGuard SVM-RF. X axis denote the Dataset and y axis denotes the Recall ratio. The Proposed Ensemble HeartGuard SVM-RF values are better than the existing algorithm. The existing algorithm values start from 0.76 to 0.88, 0.69 to 0.82 and Proposed Ensemble HeartGuard SVM-RF values starts from 0.86 to 0.99. The proposed method yields excellent results.

4.4 F1 – score

F1- score is a test's accuracy that combines precision and recall. It is calculated by taking the harmonic mean of precision and recall.

$$F1 - score = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

Datasets	KNN	RF	Proposed Ensemble HeartGuard SVM-RF
100	0.87	0.79	0.97
200	0.89	0.78	0.99
300	0.85	0.69	0.95
400	0.78	0.67	0.93
500	0.79	0.65	0.91

Table 4.Comparison Table of F1 – score

The Comparison table 4 of F1- score Values explains the different values of existing KNN, RF and Proposed Ensemble HeartGuard SVM-RF. While comparing the Existing algorithm and Proposed Ensemble HeartGuard SVM-RF, provides the better results. The existing algorithm values start from 0.78 to 0.89, 0.65 to 0.79 and Proposed Ensemble HeartGuard SVM-RF values starts from 0.91 to 0.99. The proposed method yields excellent results.

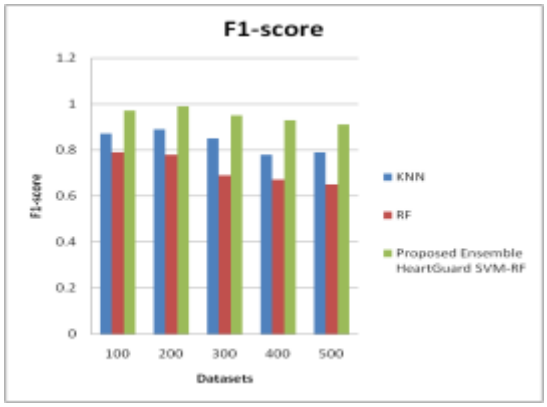

Figure 5.Comparison Chart of F1 – score

The Figure 5 Shows the comparison chart of F1 - score demonstrates the existing KNN, RF and Proposed Ensemble HeartGuard SVM-RF. X axis denote the Dataset and y axis denotes the F1 - score ratio. The Proposed Ensemble HeartGuard SVM-RF values are better than the existing algorithm. The existing algorithm values start from 0.78 to 0.89, 0.65 to 0.79 and Proposed Ensemble HeartGuard SVM-RF values starts from 0.91 to 0.99. The proposed method yields excellent results.

5. Conclusion

In this paper, Ensemble HeartGuard SVM-RF methodology for heart disease classification and prediction integrates state-of-the-art machine learning techniques to provide accurate and timely diagnostic insights. By emphasizing data quality, feature selection, and model optimization, the approach seeks to enhance the overall efficacy of CVD diagnosis. Continuous monitoring and collaboration with healthcare professionals ensure the model's relevance and reliability in real-world clinical settings. As advancements in machine learning and medical research progress, this methodology can serve as a foundation for developing more sophisticated and effective tools for heart disease diagnosis and prediction.

References

1. B. Anishfathima, R. Vikram, S. R. T, M. Sri Vishnu and C. Venumadhav, "A Comparative Analysis on Classification Models to predict Cardio-vascular disease using Machine Learning Algorithms," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2022, pp. 259-264, doi: 10.1109/ICAIS53314.2022.9741831.
2. D. P. Yadav, P. Saini and P. Mittal, "Feature Optimization Based Heart Disease Prediction using Machine Learning," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702410.
3. H. E. Hamdaoui, S. Boujraf, N. E. H. Chaoui and M. Maaroufi, "A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques," 2020 5th International Conference on Advanced

Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 2020, pp. 1-5, doi: 10.1109/ATSIP49331.2020.9231760.
4. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
5. K. Karthik, A. L. Reddy, R. Kulkarni and M. J. Mehdi, "Algorithm Accuracy Verification in Heart Disease Analysis using Machine Learning," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 345-349, doi: 10.1109/ICAAIC56838.2023.10140446.
6. K. P. Kumar, V. Rohini, J. Yadla and J. VNRaju, "A Comparison of Supervised Learning Algorithms to Prediction Heart Disease," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-5, doi: 10.1109/ICECONF57129.2023.10084035.
7. K. P. Lakshmi and C. R. K. Reddy, "Fast rule-based heart disease prediction using associative classification mining," 2015 International Conference on Computer, Communication and Control (IC4), Indore, India, 2015, pp. 1-5, doi: 10.1109/IC4.2015.7375725.
8. M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden naïve bayes classifier," 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), Bangalore, India, 2016, pp. 1-5, doi: 10.1109/CIMCA.2016.8053261.
9. M. Chakarverti, S. Yadav and R. Rajan, "Classification Technique for Heart Disease Prediction in Data Mining," 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, India, 2019, pp. 1578-1582, doi: 10.1109/ICICICT46008.2019.8993191.
10. M. Islam and R. Islam, "Exploring the Impact of Univariate Feature Selection Method on Machine Learning Algorithms for Heart Disease Prediction," 2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM), Gazipur, Bangladesh, 2023, pp. 1-5, doi: 10.1109/NCIM59001.2023.10212832.