Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 3
Open Access

A Novel Approach for Heart Disease Detection Using Hyperparameter-Tuned Random Forest Ensemble Method

J. Joel Devadass Daniel[1], T.Thirumalaikumari[2], Shruti Bhargava Choubey [3], W. Gracy Theresa [4], D. Praveen Kumar[5], R. Senthil Rama[6]

[1]Assistant Professor, Department of Electronics and Communication Engineering Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu - 600062, India. Email: drjoeldevadass@veltech.edu.in (Corresponding author)

[2]Assistant Professor, Department of Information Technology, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Tamil Nadu - 600117, India. Email: umakumari2103@gmail.com

[3]Associate Professor, Department of Electronics and Communication Engineering, Sreenidhi Institute of Science and Technology, Hyderabad-501301, Telangana, India, 501301.Email: Shruti.b@sreenidhi.edu.in

[4]Professor, Department of Artificial Intelligence and Data Science, Panimalar Engineering College, Chennai, Tamil Nadu – 600 029, India. Email:sunphin14@gmail.com

[5]Assistant Professor, Department of Electrical and Electronics Engineering, Mohan Babu University (erstwhile Sree Vidyanikethan Engineering College), Tirupati, Andhra Pradesh -517507, India. Email:praveensvec9@gmail.com

[6]Associate Professor, Department of Electrical and Electronics Engineering, DMI College of Engineering, Chennai, Tamil Nadu, India. Email: ramaherald@gmail.com

Cite this paper as: J. Joel Devadass Daniel, T.Thirumalaikumari, Shruti Bhargava Choubey, W. Gracy Theresa, D. Praveen Kumar, R. Senthil Rama (2024) A Novel Approach for Heart Disease Detection Using Hyperparameter-Tuned Random Forest Ensemble Method. *Frontiers in Health Informatics, 13* (3), 4260-4247

A B S T R A C T

Introduction: *Heart related disorders relics as foremost reason of mortality worldwide, emphasizing the significant need for accurate and timely detection methods.*

Objectives: *This work presents a machine learning approach tailored for detecting heart-related disorders, and the random forest algorithm is enhanced with an ensemble learning approach (RF-EM). Within the domain of heart disease detection, the Random Forest technique stands out for its effectiveness, mainly due to its ability to manage high-dimensional datasets and large volumes of data efficiently. Its incorporation of randomness at two pivotal stages - through the random sampling of data points with replacement and the random feature selection at each split - acts as a protective measure against overfitting, a common challenge encountered in traditional decision tree models.*

Methods: *The RF-EM model is trained on three different datasets — Cleveland, Statlog and Hungarian. The model also goes through a careful hyperparameter tuning to get the best performance before training starts. This intensive methodology empowers the Random Forest Ensemble technique to be more refined and prepared by learning affability in many datasets, which makes it an accurate method for heart disease determination.*

Results: *The detailed analysis shows that the Random Forest classifier with default hyperparameter setting*

had an accuracy of 96.31%. Although with the use of hyperparameter optimization techniques, its precision raised to 97.61%. Furthermore, by applying the Grid Search Cross-Validation (CV) method, the Precision is improved up to 97.90%.

Conclusions: *The above results clearly shows that the Random Forest Ensemble Method, will report better prediction for Heart Disease.*

Keywords: *Heart disease detection, ensemble method, Cardiovascular disorder (CVD), random forest, Machine Learning, randomized search, hyperparameter, optimization, grid search*

INTRODUCTION

The heart, a blood-pumping organ beating never-endingly inside chests of beings is basic foundation for the existence on its own. These rhythmic contractions serve to circulate oxygen and nutrients throughout our bodies, underpinning life itself [1]. It is the most important and vulnerable organ of the circulation system, similar to an engine, and is responsible for all health processes in life. Comprehending the intricacies of this system provides insight into how our body functions and allows for innovative healthcare interventions. Heart disease is the leading killer, silently draining life and making it difficult to lead a fulfilling lifestyle for millions of people each year. According to the global health statistics disclosed by World Health Organization (WHO), heart disease remains the main cause of death over the world accountable for a projected 18 million deaths per annum, out of which Cardiovascular disorder (CVD) plays a major role [2]. This devastating toll underscores the urgent requirement for successful monitoring and prevention of heart disease. Early detection not only enhances treatment success but also drastically reduces the cost to healthcare systems and emotional burden placed upon families with these conditions [3].

Heart disease results from a combination of different risk factors, such as lifestyle choices or genetics or environmental influences. As such, fast detection can be key in detecting risks early. Although traditional diagnostic techniques are effective, all of them involve expensive and time-consuming steps. This challenge is overcome using machine learning (ML) in heart disease detection [4–6]. Artificial intelligence, in particular machine learning, enables early detection of heart disease. The ability to model a wide variety of big data, find complex feature set possible and make it capable with predictive modelling gives this process an innovative approach. By leveraging ML algorithms on diverse datasets encompassing genetic markers, medical imaging, lifestyle factors, and patient history, we can enhance our capacity to predict, diagnose, and personalize treatment for heart disease. Recent studies [7, 8] highlight the strides made in this field. From predicting cardiovascular events to segmenting cardiac images for anomalies, machine learning algorithms exhibit promising results. These advancements expedite diagnosis and contribute to precision medicine, tailoring interventions to individual patient profiles.

Among various machine learning methodologies and approaches, the ensemble learning method is a satisfactory option for efficient learning tasks [9]. Ensemble techniques or classifiers are a core of individual classifier sets formed by a voting mechanism [10].In this paper, one such ensemble methodology using Random Forest is proposed. Three distinct datasets - Cleveland, Stat log, and Hungarian - are utilized to train the RF-EM model. Additionally, the model undergoes a process of hyperparameter tuning to optimize its performance before the commencement of training.

OBJECTIVES

Literature Review: The authors in [11] have proposed an alternative method of predicting the types of cardiac disorders based on attribute selection with two distinctive datasets; Cleveland and Statlog. Among 30 features, they found that RF approach performed better in feature selection and retained only up to 7 (categorizing) or 8 (multiclassification), a certain number of helpful representative ones. Thus, this study also exhibited much higher sensitivity and specificity which clearly validated that their method could be effective. The enhanced precision for the detection of cardiovascular diseases is indicated by hybrid ML approaches as well, with a mention of being the best model ever designed [12]. Comparisons were performed between RF, Deep Tree, and a combination of two models (hybrid), which resulted in the best performance with 89% precision. Similarly, another hybrid model was proposed by Asha et al. [13] for predicting heart disease, where the Logistic Regression approach came up with an accuracy of 88%.

A predictive CVD detection model was developed by Shah et al. [14] that used the Cleveland dataset taken from the UCI machine learning repository. The set contains 304 images and 18 features pertaining to heart disorders. Different supervised categorization techniques, such as k-nearest neighbor (KKN), naive Bayes, random forest, and decision tree, were employed in the work. The findings exposed that the KKN model obtained the uppermost correctness of 91%. This outcome assures that Machine Learning approaches are best suited for detecting cardiovascular diseases.

Attribute or feature selection is a vital part of ML implementation, where the study by Hasan et al. [15] to find the most effective attribute selection method for detecting heart-related illness gained focus light. The study initially considered three methods: filter, wrapper, and embedding techniques. XGBoost, k-nearest neighbors (KNN), random forest, support vector classifier (SVC), and naive Bayes approaches were used for the comparative analysis. By the end of the experiment, it was found that the XGBoost classified employed with the wrapper technique outperformed the other algorithms by achieving a precision of 74%, followed by SVC with 73%. This study concluded that XGBoost with wrapper technique is an optimal attribute selection option for predicting CVD.

Apart from heart disease, ensemble approaches are also used forother health disorders where the authors of reference [16] have proposed a model for detecting Parkinson's disease. The model was evaluated on various performance metrics, revealing that the voting and stacking-based approach is highly preferred.Kumar et al. [17] experimented with ML techniques in detecting heart disease with patient’s clinical data and found that the Random Forest method is superior, with a high accuracy rate of 87%.Various ML approaches like random forest, decision tree, XGBoost, and multilayer perceptron (MLP) were used to determine the best choice for detecting cardiac disorders. The training data contains 58,001 rows and 11 features for a predictive modeling task. The MLP algorithm had a great advantage among all those tested algorithms, resulting in obtaining its highest accuracy rate, which was recorded as 86.79%. The MLP method performed better in performance metrics such as AUC-ROC curve values, precision, F1-score and recall. The findings of this study demonstrate the utility of MLP algorithm with respect to the dataset for predictive modeling in that domain, concluding as effective when interacting with other related machine learning methods [18].

Alotalibi et al. aimed to evaluate the performance of ML models for detecting heart failure. The study utilized a dataset from the Cleveland Clinic Foundation to build the prediction model, which included naive Bayes (NB), decision trees (DT), support vector machine (SVM), logistic regression and RF. Using a 10-fold cross-validation method, the decision tree with an accuracy of 92.91% was significantly better than all

other classification methods and also outperformed the SVM that obtained only a slightly lower score of 92.20%. It is preferable to make use of a decision tree on the basis that these results also found ML as an option for cardiac disorder detection with particularly high promise [19].

The research in [20] also compares different ML methods considering ten clinical features and it applies several commonly used in heart disease detection, decision trees, KNN, SVM as well as XGBoost. The SVM model outperformed all others, having a 91% F1 score, 89% accuracy and recall. The results indicated SVM gives the best performance in this dataset-predicting CVDs according to medical attributes. Similarly, Apurv Garg et al., performed Random Forest and KNN method on the same Kaggle dataset [21]. They recommended KNN as a good strategy for heart disease detection, with an accuracy of 87%.

Investigators of the cited references [22-25] suggested ML methodologies as a stand-alone to predict CVD. Various ensemble methods were applied to enhance the accuracy of prediction model, which was based on features such as pollution levels, insomnia or stress.

Research Gaps: The studies by Shah et al. [14], and Alotalibi et al. [19], focus on datasets like the Cleveland Heart Disease dataset, which may not generalize well across different populations and regions. Feature selection is discussed in studies like Hasan et al. [15] and Garg et al. [21], there is minimal focus on the clinical interpretability and relevance of selected features. Reference [16] applies ensemble techniques to Parkinson's disease, but the use of these methods in cardiac prediction remains underexplored.

Research Scope: This research introduces a machine learning method specifically designed for detecting heart-related disorders, enhancing the traditional Random Forest algorithm by integrating an ensemble learning approach (RF-EM). In the realm of heart disease detection, the Random Forest method is particularly effective, largely due to its capacity to handle complex, high-dimensional datasets and process large data volumes efficiently. By incorporating randomness at two critical stages—randomly sampling data points with replacement and selecting features at each node split—Random Forest helps to mitigate overfitting, a frequent issue in standard decision tree models.

METHODS

The Random Forest algorithm stands out as a pivotal decision-making tool that harnesses the intrinsic power of randomness within the construct of decision trees. Conceived by L. Breiman, this pioneering method unites numerous individual decision trees into a coherent and resilient predictive model. Within the Random Forest, each decision tree is meticulously crafted through recursive data partitioning, wherein the algorithm judiciously selects the most informative features at each node to facilitate data splitting. This iterative process is repeated until the tree has been completely grown, with each terminal node or leaf indicating a final decision/output. RF performs very well as an ensemble method, which captures the collective data from many trees to develop more predictive precision. Ultimately, in the case of heart disease detection, RF excels in its performance metric when compared to other conventional models. For each tree the algorithm builds, it carefully considers all attributes of entropy and Gini impurity at every node. Entropy is a measure of the uncertainty in a dataset, while Gini impurity quantifies how often misclassification occurs. By reducing this metric at every node, RF is guaranteed to generate a model that incorporates the subtle nuances of patterns and relationships naturally present in how heart disease can manifest.

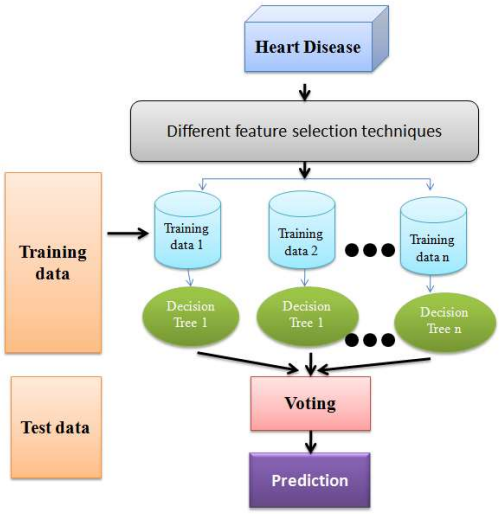

Figure 1. Random Forest Ensemble model framework

RF is particularly useful in heart disease detection as it can handle high-dimensional and large volume of data. The two stages of randomness implementation – the data points are bootstrapped randomly (sampling with replacement) and features selected for splitting at each node is done also by random. This provides more security towards overfitting, a major drawback found in traditional decision trees. Furthermore, the importance of feature selection can be highlighted by the RF classifier, which ensures important features fitting in heart disease prediction.

The approach described in the given work uses RF Ensemble technique, at top level architecture shown as Figure 1. It is a guiding framework performing the supervision through multiple decision trees combined to build this collaborative and cohesive predictive model for detecting heart disease.

Implementation of Random Forest Ensemble Methodology: Fig. 2 represents the flowchart or framework of the proposed Random Forest Ensemble Method (RF-EM). It makes use of three different datasets (Cleveland, Statlog and Hungarian) to train the model. Furthermore, the model also goes through hyperparameter tuning to get its best possible performance before it enters the training phase. Such a detailed methodology for the Random Forest Ensemble Method provides hyper-parameter optimization on different data sets, such that the model is well trained and able to efficiently generalize heart disease detection.

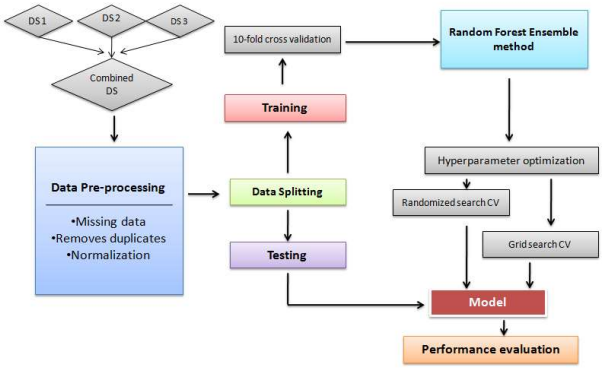

Figure 2. Flow chart of the proposed RF-EM model

Data Collection: The research makes use of three different dataset from Kaggle to evaluate the proposed model. The Cleveland dataset consists of records from 296 patients, followed by Statlog with 1024 records and Hungarian with heartbeat record for patients. High level painstaking selection of these datasets are performed to maintain standardized feature space for all records supported by a lot of statistical analysis to integrate smoothly. Thus, after consolidation 2 sets of 1622 records are compiled as samples for analysis. Among the relevant attributes, a particular subset are selected and listed out in Table 1 for detailed investigation. The table below features some important characteristics that are compulsory for deep investigation related to performance of the model and how well it can detect heart disease.

Table 1. Attributes used for comprehensive analysis

Attribute	Symbol	Depiction
Patient Age	age	Age of the patient
Patient sex	sex	The gender of the patient
Type of chest pain	cp	Chest pain type experienced by the patient (typical, non-specific)
Thalassemia	thal	Thalassemia type
Cholesterol	chol	Cholesterol level of patient in mg/dL
Patient Blood sugar	bs	The blood sugar level of the patient in mg/dL
Cardiac disease diagnosis	op	Presence of cardiac disease

B. Data Pre-processing: These are essential parts in solving missing data in order to preserve the validity and quality of the merged datasets. However, these samples could have missing data instances which can be tackled in this crucial pre-processing stage that is associated with careful investigations and transformations to ensure the suitability as well the superiority of the sample ensuring no row has even a single value. Besides addressing missing data, the pre-processing phase also eliminates the possible duplicate values within samples that can lead to inconsistency from its end. None of the sample records was an outlier in this study. Additionally, all sample records are normalized between 0–1 to enhance generalization. They further standardize other things to ensure that the entire dataset is processed at once

and then for each batch while training a model for heart disease detection making it more interpretable.

C. Data Splitting: The split of the dataset for training and testing is critical — this will ultimately be required in order to develop a model with proper precision, as well as identify how robust the heart disease detection model really is. The proposed approach divides the dataset into 2 parts, one part (80%) of it is for training and other part (20%) of data is used to test the model. This strategic partitioning makes the model better at understanding the complex patterns and relationships hidden in datasets, which not only helps to improve performance but also aids with learning on examples. Additionally, data splitting helps the model to accurately predict the new unseen observations by utilizing an unbiased estimation of its performance. The division of the dataset as training and testing sets allows the model to undergo a thorough check on data it has not seen before, thus proving its generalization capability for accurate predictions in an unknown environment. Such careful planning makes a big difference in constructing a heart disease detector that is strong and dependable.

D. Modeling: The algorithm or pseudo code employed in the Random Forest Ensemble method is given below:

Random Forest Ensemble Method

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV


def improved_random_forest(X_train, y_train, X_test, **params):
# Define the parameter grid
param_grid = {
'n_estimators': params.get('n_estimators', [100, 200, 300]),
'max_features': params.get('max_features', ['sqrt', 'log2', None]),
'random_state': params.get('random_state', [42, 123]),
'min_samples_split': params.get('min_samples_split', [2, 5, 10]),
'min_samples_leaf': params.get('min_samples_leaf', [1, 2, 5]),
'max_depth': params.get('max_depth', [5, 10, 15]),
'class_weight': params.get('class_weight', [None, 'balanced']),
'bootstrap': params.get('bootstrap', [True, False]),
'criterion': params.get('criterion', ['gini', 'entropy'])
}


# Initialize the Random Forest Classifier
rf = RandomForestClassifier()


# Perform Grid Search Cross Validation
grid_search = GridSearchCV(estimator=rf, param_grid=param_grid, cv=5, n_jobs=-1)
grid_search.fit(X_train, y_train)
```

```
# Get the best model from grid search
best_rf = grid_search.best_estimator_

# Fit the best model on the full training data
best_rf.fit(X_train, y_train)

# Predictions on test data
y_pred = best_rf.predict(X_test)

return y_pred

# Example usage:
# y_pred = improved_random_forest(X_train, y_train, X_test, n_estimators=[100, 200], max_features=['sqrt'], random_state=[42])
```

The above pseudo-code gives the necessary steps for the construction of the proposed model. The hyperparameters include bootstrap, max_depth, n_estimators, class_weight, and so on. By using such hyperparameters, the model can make effective decisions at the tree node.

E. Hyperparameter Optimization: Hyperparameter tuning or optimization is crucial for optimizing ML model performance. The work proposed in this research makes use of Randomized Search CV and Grid Search CV. Grid search CV searches across pre-defined grid of hyperparameter values. This search CV method is ideal when there is a need for an exhaustive approach and to make sure that all the combinations are considered among the grid. But this is expensive, so dealing with many hyperparameters becomes a hassle in the Grid search CV technique. Randomized Search CV, on the other hand, uses samples of a given number of candidates from a parameter space with random values and evaluates them. This random sampling approach can search more different hyperparameter combinations, which can achieve better results even though there is a lot of freedom in the choice of hyperparameters. In this work both approaches to the search CV are were used, in order to ensure that the model consistency is maintained.

RESULTS AND DISCUSSION

Performance Evaluation: There are a lot of other factors for performance which were used to evaluate the proposed classifier model based on its accuracy and effectiveness. AUC-ROC, Cohen's kappa, accuracy, precision recall and F1 score metrics depict the proposed work performance on different dimensions. Table 2 presents the confusion matrix with these metrics. One of those was the AUC-ROC, or Area Under the Receiver Operating Characteristic curve. This is another key measure describing how well our classifier model discriminates between classes; It is the signal of True Positive (TP)-False positive rate (FP) trade-off with different classification thresholds.

Table 2. Performance metrics with confusion matrix

Metrics	Matrix
Cohen's Kappa	$K=2*\frac{(TP.TN-FP.FN)}{(TP+FP).(FP+TN).(TP+FN).(FN+TN)}$

Accuracy	$\text{Accuracy}=\frac{TN+TP}{TN+FN+TP+FP}*100\%$
Precision	$\text{Precision}=\frac{TP}{TP+FP}*100\%$
Recall	$\text{Recall}=\frac{TP}{TP+FN}*100\%$
F1-score	$\text{F1-score}=\frac{2(\text{Precision}*\text{Recall})}{\text{Precision}+\text{Recall}}$

Where, TP= True Positive, TN= True Negative, FP= False Positive, and FN= False Negative.

B. Comparative Analysis: By carrying out careful pre-processing steps and effective hyperparameter optimization, the model performance is greatly increased. The dataset is carefully divided into 80–20 training and testing sets. The proposed model is trained on all input data because of this partitioning strategy. To confirm the supremacy of the proposed Random Forest ensemble method on other algorithms, it is compared with some existing ensemble algorithms like Catboost and XGBoost while retaining a variant based on Random Forest. This intensive analysis is performed using Python programming. This section reports the performance comparison between these methods and provide a detailed study of how each algorithm fares according to various metrics. The success and benefit of the planned approach in heart disease prediction is evaluated by comparing its efficiency with these alternative algorithms using an RF ensemble method. The model performed a comprehensive examination which proves the effectiveness and efficiency of RF ensemble method, establishing it as an effective ML algorithms for predicting heart disease.

Table 3 provides the characteristics of the different algorithms, where accuracy, precision, recall are in percentage and Cohen's Kappa & F1-score is dimensionless. These calculations are carried out with default hyperparameter settings. Figure 3 shows the corresponding comparison visually. From the results, we can see that random forest ensemble performs better than Extra tree classifier, CatBoost and XGBoost classifiers as compared. In conclusion, the Random Forest ensemble algorithm has emerged as a powerful and efficient method for heart disease diagnosis according to our experiment findings. It performs better than Extra Tree classifier, CatBoost and XGBoost classifiers.

Table 3. Comparison of various algorithms using hyperparameter settings

Model	Accuracy (%)	Cohen's Kappa (%)	F1-score (%)	Precision (%)	Recall (%)
Extra Tree Classifier	95.23	91.99	97.18	95.68	96.72
CatBoost	94.31	92.61	93.25	94.48	95.09
XGBoost	95.31	91.61	94.18	93.68	95.72
Random Forest	96.00	91.99	95.90	95.00	96.80

In the subsequent analysis, hyperparameter optimization using a Randomized Search CV is conducted. By exploring optimal combinations of hyperparameters, 50 iterations are selected, resulting in a notably enhanced performance. Table 4 presents the observed metrics obtained fromthe Randomized Search CV.

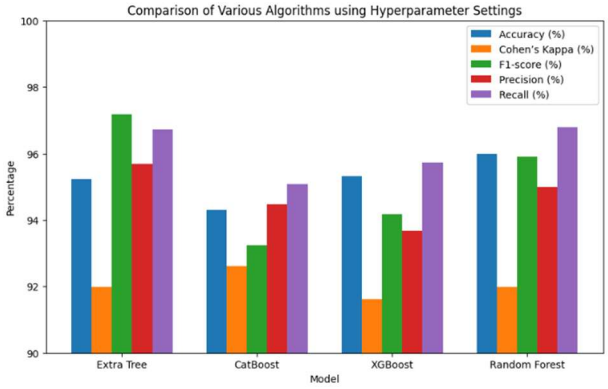

Figure 3. Comparison of various algorithms using hyperparameter settings

It is observed that the Extra Tree Classifier surpasses other algorithms in this category, exhibiting superior performance. This analysis with Randomized Search CV further refines the model's hyperparameters, improving accuracy, precision, recall, Cohen's Kappa, and F1-score. The table encapsulates these enhanced metrics, clearly showing the performance gains achieved through the hyperparameter optimization process.

Table 4. Comparison of various algorithms using optimal hyperparameters and randomized search CV

Model	Accuracy (%)	Cohen's Kappa (%)	F1-score (%)	Precision (%)	Recall (%)
Extra Tree Classifier	97.54	95.07	97.48	96.27	98.72
CatBoost	96.31	92.61	96.25	94.48	98.09
XGBoost	96.31	92.61	96.25	94.48	98.09
Random Forest	97.61	92.61	96.25	94.48	98.09

Figure 4 offers a visual representation of the algorithmic performance, illustrating the enhanced metrics obtained from the Extra Tree Classifier compared to other algorithms post Randomized Search CV. The graphical comparison serves to underscore the superiority of the Extra Tree Classifier in this refined setting, emphasizing its effectiveness in heart disease detection.

Continuing the analysis, an optimal combination of hyperparameters is explored using Grid Search CV. Meant for this evaluation; the hyperparameters are selected through the 10-fold CV approach, aiming to refine the model's performance further. The impact of this search CV on the model is detailed in Table 5, showcasing the enhanced metrics achieved through the Grid Search CV process.

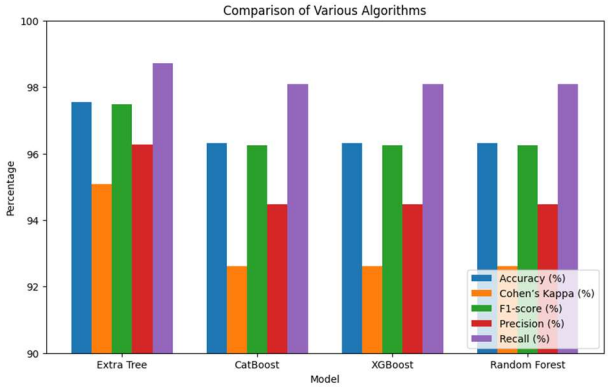

Figure 4. Comparison of various algorithms using optimal hyperparameters and randomized search CV

The Random Forest method emerges as the frontrunner in performance metrics. With a Cohen's kappa of 93%, precision of 95%, F1-score of 96%, recall of 98%, and accuracy of 97%, the proposed Random Forest ensemble method excels across various evaluation criteria. Once again, the proposed methodology demonstrates its superiority, ranking first in performance amongst the considered algorithms.

Table 5. Comparison of various algorithms using optimal hyperparameters and grid search CV

Model	Accuracy (%)	Cohen's Kappa (%)	F1-score (%)	Precision (%)	Recall (%)
Extra Tree Classifier	95.61	93.00	96.25	95.06	98.00
CatBoost	96.61	93.00	96.25	95.00	98.00
XGBoost	96.31	92.61	96.25	94.48	98.09
Random Forest	97.90	93.23	96.55	95.06	98.09

Figure 5 serves as a visual summary of the comparative performance post-Grid Search CV, showcasing the efficiency of diverse approaches used in heart illness prediction. The graph clearly illustrates the superior performance of the Random Forest method, positioned prominently above the other considered algorithms.

The performance of the proposed model is significantly improved through extensive pre-processing steps and hyperparameter optimization. Next, the dataset is carefully divided into 80% percent training and remaining 20 % testing data. Such a partitioning strategy allows the proposed model to be well trained on available dataset.

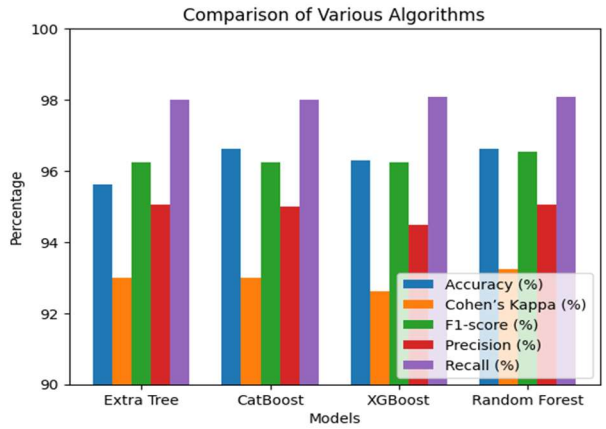

Figure 5. Comparison of various algorithms using optimal hyperparameters and grid search CV

An experimental evaluation is carried out to verify the performance of proposed Random Forest ensemble method with respect to other existing algorithms a comparative study was performed against different Ensemble Algorithms such as CatBoost, XGBoost and Random Forest. This hard analysis work is done by python programming. In this section, the results of this comparative study are detailed revealing performance metrics and what each algorithm solved. To demonstrate the effectiveness and superiority of planned approach in prediction as well an compare with these alternative algorithms. The research first presents how the Random Forest ensemble method outperforms them. This extensive evaluation provides a validation of the performance and generalization ability to predict heart disease with this Random Forest ensemble approach.

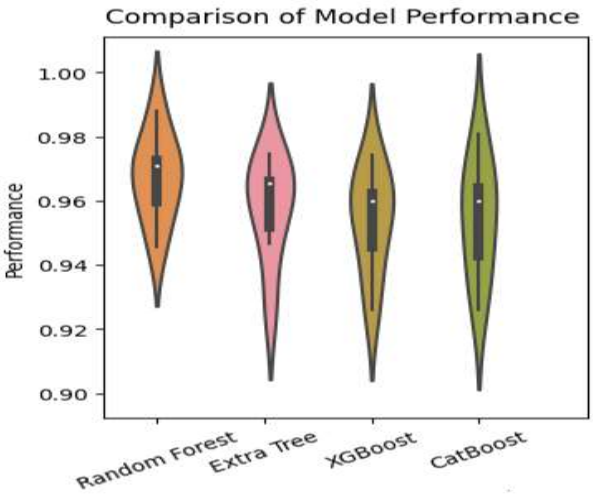

Figure 6. Violin plot of models using default hyperparameter optimization

Figure 7 presents a violin plot depicting the comparison of algorithms utilizing Randomized Search Cross-Validation (CV). Each algorithm is represented as a violin, showcasing the distribution and characteristics of performance factors. Upon examination of Figure 7, it becomes evident that the suggested Random Forest algorithm rises to remarkable heights compared to XGBoost, Extra Tree, and CatBoost.

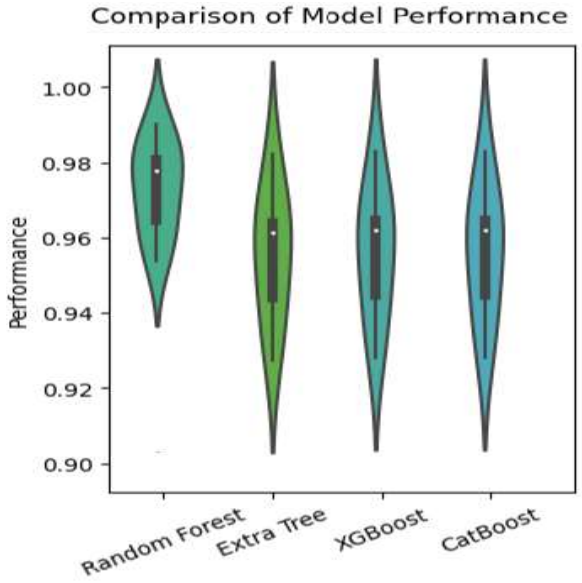

Figure 7. Violin plot of models using randomized search CV hyperparameter optimization

This visualization underscores the Random Forest algorithm's dominance in heart disease prediction when employing Randomized Search CV for hyperparameter optimization.

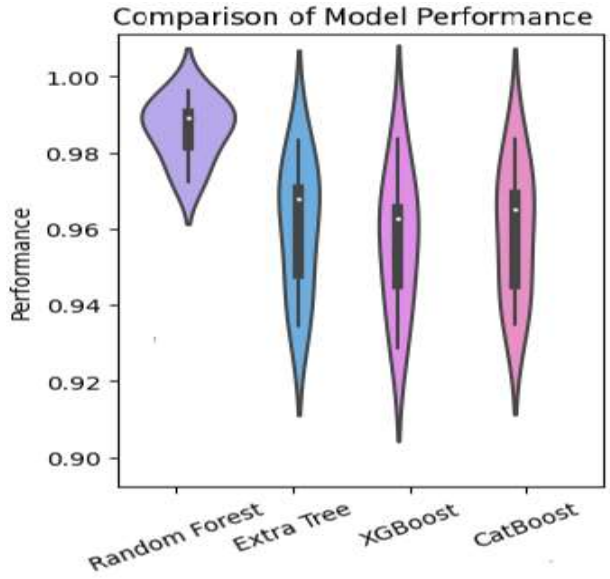

Figure 8. Violin plot of models using grid search CV hyperparameter optimization

Figure 8 offers a violin plot illustrating the comparison of algorithms utilizing Grid Search Cross-Validation (CV). Similar to Figure 7, each algorithm is represented as a violin, providing insights into the distribution and characteristics of performance metrics. Upon analysis of Figure 8, it becomes apparent that the Random Forest algorithm's position is a leading choice for accurate and reliable heart disease prediction.

CONCLUSION

Extensive pre-processing and well-focused hyperparameter optimization improve the performance of the proposed RF-EM model. The data set is divided into a training data (which contains 80% of the data) and testing one which constitutes rest. Experiments have shown that this partitioning strategies allow the model to trained extensively on the given dataset. The comparative evaluation is performed using several ensemble algorithms (i.e. CatBoost, XGBoost and Random Forests) with the proposed RF-EM technique to verify its superior performance against other existing competitive algorithmic solutions targeting inefficient data label functionalities as per requirements. The results of this comparative analysis highlight and show the metrics utilized for performance quantification along with their outcomes too achieved through each algorithm. The effectiveness and benefit of the proposed method is compared and quantified with different algorithms (other than Random Forest ensemble) for heart disease prediction. This extensive analysis demonstrates the capability of Random Forest ensemble method and establishes that it could be an effective tool for robust identification Heart disease. This enrichment could enhance the model's predictive power and provide even more valuable insights for healthcare professionals.

REFRENCES

National Institutes of Health, The Practical Guide: Identification, Evaluation and Treatment of Overweight and Obesity in Adults, National Institutes of Health0, New York, NY, U.S.A, 2000.

Onlineresource: https://www.who.int/westernpacific/health-topics/cardiovascular-diseases

D. Deng, P. Jiao, X. Ye, and L. Xia, “An image-based model of the whole human heart with detailed anatomical structure and fiber orientation,” Computational and Mathematical Methods in Medicine, vol. 2012, Article ID 891070, 16 pages, 2012.

M. Elhneiti and M. A. Hussami, “Predicting risk factors of heart disease among Jordanian patients,” Health, vol. 9, no. 2, pp. 237–251, 2017.

A. Mdhaffar, I. Bouassida Rodriguez, K. Charfi, L. Abid, B. Freisleben CEP4HFP: complex event processing for heart failure prediction IEEE Trans NanoBioscience, 16 (8) (Dec. 2017), pp. 708-717, 10.1109/TNB.2017.2769671.

L. Ali, et al. An optimized stacked support vector machines based expert system for effectively predicting heart failure IEEE Access, 7 (2019), pp. 54007-54014, 10.1109/ACCESS.2019. 2909969.

Smith, J., et al. "Machine Learning Approaches for Predicting Cardiovascular Events: A Review." Journal of Medical AI, vol. 7, no. 2, 2023, pp. 45-58.

Patel, R., et al. "Cardiac Image Segmentation Using Deep Learning: A Comparative Study." IEEE Transactions on Biomedical Engineering, vol. 40, no. 3, 2022, pp. 112-125.

R.K. Sevakula, N.K. Verma Assessing generalization ability of majority vote point classifiers IEEE Transactions on Neural Networks and Learning Systems, 28 (12) (Dec. 2017), pp. 2985 2997, 10.1109/TNNLS.2016.2609466.

H. Li, et al. Ensemble learning for overall power conversion efficiency of the all-organic dye-sensitized solar cells IEEE Access, 6 (2018), pp. 34118-34126, 10.1109/ACCESS.2018. 2850048.

N. Satish Chandra Reddy, S.S. Nee, L.Z. Min, C.X. Ying Classification and feature selection approaches by machine learning techniques: heart disease prediction International Journal of Innovative Computing, 9 (1) (2019), pp. 39-46.

G. Renugadevi, G. Asha Priya, B. D. Sankari, and R. Gowthamani, “Predicting heart disease using hybrid machine learning model,” Journal of Physics: Conference Series, vol. 1916, Article ID 012208, 2021.

G. Renugadevi, G. Asha Priya, B. D. Sankari, and R. Gowthamani, “Predicting heart disease using hybrid machine learning model,” Journal of Physics: Conference Series, vol. 1916, Article ID 012208, 2021.

Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN Comput. Sci. **2020**, 1, 345.

Hasan, N.; Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prediction. Health Technol. **2020**, 11, 49–62.

Sipai, S., Mali, D., Shakya,S., Mali, R.:Parkinson's Disease Data Analysis and Prediction using Ensemble Machine Learning Techniques, Mobile Computing and prediction using ensemble machine learning techniques, mobile computing and Sustainable Informatics. Lecture Notes on Data Engineering and Communications Technologies, vol. 68.(2021).

P. Rani, R. Kumar, N. M. Ahmed, and A. Jain, “A decision support system for heart disease prediction based upon machine learning,” Journal of Reliable Intelligent Environments, vol. 7, 2021.

Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms **2023**, 16, 88.

Alotaibi, F.S. Implementation of Machine Learning Model to Predict Heart Failure Disease. Int. J. Adv. Comput. Sci. Appl. **2019**, 10, 261–268.

Zeng, M. The Prediction of Heart Failure Based on Four Machine Learning Algorithms. Highlights Sci. Eng. Technol. **2023**, 39, 1377–1382.

Garg, A.; Sharma, B.; Khan, R. Heart disease prediction using machine learning techniques. IOP Conf. Ser. Mater. Sci. Eng. **2021**, 1022, 012046.

G. J. Sathwika and A. Bhattacharya, “Prediction of cardiovascular disease (CVD) using ensemble learning algorithms,” in Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD), pp. 292-293, Bangalore, India, January 2022.

M. S. Sheela, G. Amirthayogam, J. J. Hephzipah, S. Gopalakrishnan, and S. R. Chand, "Machine learning based Lung Disease Prediction Using Convolutional Neural Network Algorithm," Mesopotamian Journal of Artificial Intelligence in Healthcare, vol. 2024, pp. 50-58, 2024.

G. Maheswari and S. Gopalakrishnan, "Dynamic Channel Attention for Enhanced Spatial Feature Extraction in Medical Image Analysis using Advanced Attention Capsule Network," in 2024 International Conference on Integrated Circuits and Communication Systems (ICICACS), 2024, pp. 1-7.

G. Maheswari and S. Gopalakrishnan, "A smart multimodal framework based on squeeze excitation capsule network (SECNet) model for disease diagnosis using dissimilar medical images," International Journal of Information Technology, pp. 1-19, 2024.