2023 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI) | 979-8-3503-7336-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/EICEEAI60672.2023.10590147

Heart Disease Risk Assessment and Prediction: A Robust Ensemble Approach with Extra Tree Classifier

Md. Simul Hasan Talukder[1]
Bangladesh Atomic Energy Regulatory Authority
Dhaka, Bangladesh
simulhasantalukder@gmail.com

Sohag Kumar Mondal[2]
Electrical and Electronic Engineering
Khulna University of Engineering and Technology, Khulna, Bangladesh
Khulna, Bangladesh
ssohagkumar@gmail.com

Mohammad Aljaidi[3]
Department of computer Science
Faculty of Information Technology
Zarqa University
Zarqa, Jordan
mjaidi@zu.edu.jo

Rejwan Bin Sulaiman[4]
School of Computer Science & Technology
Northumbria University
Tayne, United Kingdom
rejwan.sulaiman@northumbria.ac.uk

Taminul Islam[5]
Department of Computer Science
Southern Illinois University Carbondale
IL, USA
islamtaminul@gmail.com

Abstract— **Heart disease remains a formidable global health challenge, affecting countless individuals and placing a substantial strain on healthcare systems worldwide. Various factors, including exercise, diabetes, age, gender, dietary habits, weight, height, and even emotional well-being, influence an individual's cardiac health. Detecting heart disease early and making precise predictions regarding its risk factors are pivotal in assessing, managing, and preventing its occurrence, ultimately leading to improved patient outcomes. In this research, a range of promising preprocessing techniques and a meticulously tuned ensemble of ten machine learning models, including Extra Trees (ET), Decision Trees (DT), Random Forest (RF), Logistic Regression (LR), Extremely Gradient Boosting (EGB), AdaBoost (AB), Support Vector Machine (SVM), Ridge Classifier (RC), Light Gradient Boosting (LGB), and Gradient Boosting (GB), were employed to analyze a publicly available heart assessment dataset. The models' robustness was assessed through a rigorous 10 k-fold validation process. Seven comprehensive evaluation metrics, encompassing accuracy, AUC, precision, recall, F1 Score, kappa, MCC, were used to gauge the models' performance and reliability. Impressively, ET model outshone the rest, achieving the highest accuracy (0.9559 ± 0.0006), AUC (0.9870 ± 0.0003), precision (0.9262 ± 0.0012), recall (0.9847 ± 0.0012), F1 Score (0.9546 ± 0.0006), kappa (0.9119 ± 0.0012), MCC (0.9135 ± 0.0012), and showcasing the lowest standard deviation.**

Keywords— heart diseases assessment; Extra Tree; hyperparameters tuning; Light Gradient Boosting;

I. Introduction

Cardiovascular disease (CVD) represents a wide spectrum of disorders affecting the heart and blood vessels, constituting a significant global health challenge [1]. These conditions encompass various ailments, including coronary artery disease, hypertension, heart failure, stroke, arrhythmias, and peripheral artery disease, among others [2]. As a collective group, CVD remains the leading cause of morbidity and mortality worldwide, accounting for a substantial proportion of premature deaths and disability-adjusted life years [3]. About 17.9 million fatalities worldwide are attributed to CVD each year, accounting for 31% of all deaths [4]. By 2030, it is anticipated to generate around 23.6 million [5]. CVD can affect individuals of all ages and demographics, presenting complex and multifaceted clinical scenarios [6]. Their impact extends beyond the affected individuals, encompassing families, communities, healthcare systems, and economies. Several modifiable and non-modifiable risk factors contribute to the development of CVD. Age, gender, genetics, high blood pressure, elevated cholesterol levels, tobacco use, diabetes, obesity, physical inactivity, and poor dietary habits are among the key factors that influence an individual's susceptibility to CVD [7,8]. Understanding these risk factors and their interplay is crucial in devising effective preventive strategies and tailored interventions to mitigate CVD incidence and progression.

However, diagnosis of cardiovascular diseases in an earlier stage can reduce the mortality rate hence an automated system for diagnosis of cardiovascular diseases is crucial especially for developing and underdeveloped countries where the facilities of modern healthcare are not sufficient.

In recent years, owing to significant advancements in computational power and Artificial Intelligence (AI), Computer-Aided Diagnosis (CAD) has garnered increased prominence as a research subject within the different organizations and research communities [9]. Artificial intelligence techniques, including data mining, machine learning, deep learning, and expert systems, are being actively employed within the healthcare industry for the purposes of diagnosis, detection, and prediction of various diseases such as diabetes, waterborne illnesses, COVID-19, malaria, and typhoid, among others [10]. Machine learning (ML) is an emerging technology utilized for the analysis of clinical data and the generation of predictive insights, particularly in the realm of early disease diagnosis, within the context of modern computer-aided detection approaches [11].

In this research work, using the 2021 BRFSS Dataset from CDC, we examine the effectiveness of ten machine learning algorithms (MLAs) in this work: AB, SVM, LGB, RC, ET, DT, RF, LR, EGB, and GB. We implemented a 10-fold cross-validation approach in our training and validation datasets to bolster the model's robustness. We incorporate a missing value handling procedure to ensure data completeness for the benefit of our machine learning models. We employ data scaling techniques to maintain data uniformity and apply the Synthetic Minority Over-sampling Technique (SMOTE) to address overfitting, promote a more balanced data distribution and increase diversity within the dataset.

In this article, Section I introduces the work, while Section II covers related works. Section III, titled "Materials and Methodology," includes three sub-sections: Dataset Description, Proposed Model, and Model Evaluation. Section IV presents Results and Discussion, and Section V concludes the paper.

II. LITERATURE REVIEW

In recent studies, the state-of-the-art research on machine learning increased the potential development with satisfactory results in prediction and detection tasks on cardiovascular disease. For example, a prediction of heart disease using a combination of machine learning and deep learning was proposed by R. Bharti et al. [12]. In this paper, researchers applied different machine learning algorithms to achieve better performance and analysis the UCI Machine Learning Heart Disease dataset. This dataset contains some irrelevant features which are controlled using Isolation Forest and data are also normalized for getting better results. An extensive study was carried out by Tiwaskar et al. [7] to evaluate the appropriateness of statistical, machine learning, and data mining techniques for heart failure risk prediction. Convolutional neural networks, decision trees, random forests, and statistical techniques were all evaluated in their investigation. The prediction accuracy of these approaches was found to be, respectively, 85%, 80.1%, 85.38%, and 93%.

R. Kasabe et al. have discussed cardiovascular ailments prediction and analysis based on deep learning techniques [13]. In this research, authors evaluate different classification techniques in heart diagnosis and for this purpose the heart numeric dataset is extracted and preprocessed. Then using machine learning techniques, they classified extracted features. Researchers also calculated data precision, performance criteria involving accuracy and they examined that machine learning provides better results and efficiency, compared to existing systems.

Authors in [14] have selected an electrocardiogram sensing mechanism using deep neural networks to classify cardiac disorder. In this paper, researchers use a 12-lead-based ECG Image system to detect cardiac disorders. This research study suggests a generalized methodology to process all formats of ECG. Single Shot Detection (SSD) Mobile-Net v2-based Deep Neural Network architecture was used to detect cardiovascular disease. This study focused on sensing the four major cardiac abnormalities (i.e., myocardial infarction, abnormal heartbeat, previous history of MI, and normal class) with 98% accuracy results calculated.

Furthermore, S. Dalal et al. has developed another robust model of machine learning for CVD detection and risk prediction in accordance with a dataset that contains 11 features which may be used to forecast the disease [15]. This dataset was collected from Kaggle on cardiovascular disease and includes approximately 70,000 patient records that were used to determine the outcome. Different ML models were designed using neural networks, random forests, Bayesian networks, C5.0, and QUEST were compared for this dataset. On training and testing data sets, the results achieved a high accuracy (99.1%), which is considerably higher than previous methods.

Another group of researchers Subramani et al. [16] has shown a collection of machine learning models that can be used to address the CVD problem. The data observation mechanisms and training procedures are calculated using different ML algorithms. The proposed method provides nearly 96% accuracy result than other existing methods and the complete analysis over several metrics has been analyzed and provided.

In their study, Mezzatesta et al. [17] rigorously evaluated both linear and non-linear algorithms. Their investigation revealed that the optimal performance was achieved using the non-linear Support Vector Classification (SVC) algorithm with a Radial Basis Function (RBF) kernel. This optimization was carried out using a Grid-Search approach. The authors also introduced an optimized set of hyperparameters for their model, resulting in a notable accuracy of 95.25% for the Italian dataset and 92.15% for the American dataset.

Conversely, Rubini et al. [5] conducted an in-depth feature importance analysis, identifying 10 significant features out of the 14 available in the UCI's Heart Diseases Dataset. Furthermore, they conducted a comprehensive comparative assessment of machine learning methodologies, including Random Forest, Logistic Regression, Support Vector Machine, and Naïve Bayes, for the purpose of cardiovascular disease classification. Their findings demonstrated that Random Forest emerged as the most accurate and reliable algorithm, achieving an accuracy rate of 84.81%. As a result, the authors recommended its adoption for this specific classification task.

In a study conducted by Boursalie et al. and detailed in reference [18], they introduced a Mobile Machine Learning Model for Cardiovascular Disease Monitoring, denoted as M4CVD. This innovative system employs wearable sensors to continuously monitor vital signs, integrating this real-time data with information from clinical databases. Rather than transmitting raw data directly to healthcare professionals, the system performs local data analysis. Utilizing a support vector machine (SVM) and features derived from the amalgamated dataset, it classifies individuals as either "no longer at risk" or "persistently at risk" for cardiovascular disease (CVD). In a preliminary assessment conducted with a synthetic clinical database encompassing 200 patients, the system achieved a remarkable classification accuracy of 90.5% in predicting CVD risk. This outcome underscores the potential of the M4CVD system, marking it as a promising proof-of-concept in the field of cardiovascular disease monitoring.

Al-Batah et al [9] employed different ML algorithms to predict heart disease on public available dataset and got better performance with accuracy and AUC values.

MSH Talukder et al [19] proposed a hybrid (ET + RF) model for heart failure survival prediction. The approach yields 98.33%.

Recently, Lupague et al. [20] employed the 2021 Behavioral Risk Factor Surveillance System (BRFSS) dataset from the World Health Organization (WHO) to conduct predictive modeling for cardiovascular diseases. To bolster the robustness of their models, they implemented the 10-Fold Cross Validation technique in both their training and validation datasets. Additionally, in order to rectify dataset imbalances, the authors incorporated sampling techniques. Their study encompassed the training of various machine learning models, including K-Nearest Neighbor, Naive Bayes, Decision Tree Classifier, and Random Forest. However, these models exhibited relatively low F1 scores. Notably, in their research, it was the Logistic Regression model that exhibited superior performance following hyper-parameter tuning and reported 74% precision, recall and f1 score.

The literature review reveals that only one study has been conducted using the BRFSS dataset to assess heart diseases. While previous research focused on predicting heart diseases using various datasets with diverse features, the significance of the BRFSS dataset lies in its real-time nature and inclusion of crucial features.

III. Materials and Methodology

A. Dataset

The dataset was obtained from the public Kaggle platform [21], using the 2021 annual BRFSS data provided by the Center for Disease Control (CDC, 2021). This data set consisted of 438,693 records with a total of 304 attributes, which were accessed on a local machine. However, not all these attributes were relevant to our specific research focus. Therefore, a meticulous selection process was conducted to choose a subset of 19 attributes that were deemed pertinent. This deliberate curation of attributes led to a reduction in the number of records, resulting in a total of 308,854 data instances that were used for our comprehensive analysis. The Dataset contains 19 variables. 12 are numerical and 7 are categorical variables. The attributes are like General Health, Checkup, Exercise, Heart Disease, Skin Cancer, Other Cancer, Depression, Diabetes, Arthritis, Sex, Age Category, Height_(cm), Weight_(kg), BMI, Smoking History, Alcohol Consumption, Fruit Consumption, Green Vegetables Consumption, Fried Potato Consumption.

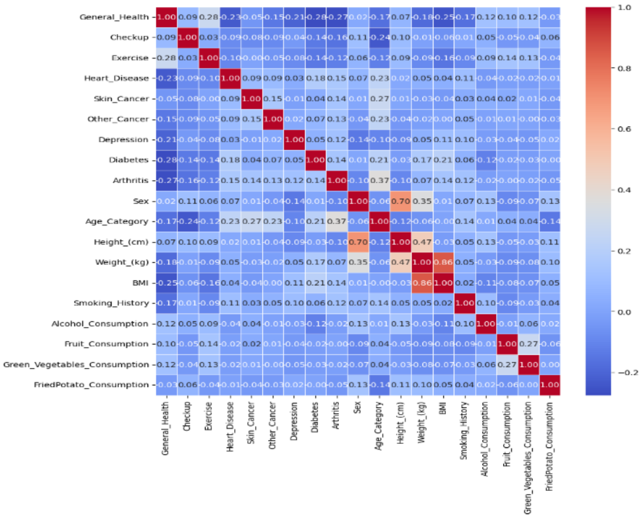

Fig. 1. Correlation among the features of BRFSS dataset.

B. Proposed Model

In our study, we conducted a series of essential preprocessing steps on the dataset to facilitate the understanding of machine learning models. These preprocessing procedures encompassed the handling of missing data, ordinal encoding, scaling, addressing range-type data, and implementing oversampling techniques. To begin, we meticulously checked the dataset for missing data using the isnull().sum() function in Pandas, and fortunately, no missing values were detected within our dataset. As a significant portion of the data was of object type, encoding became imperative. We effectively encoded the object type features using the OrdinalEncode() function within the Python environment. To ensure uniformity in data scale, we employed the MinMaxScaler technique, which was particularly useful for handling range-type data, focusing on the lower values. Subsequently, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to counter overfitting, create a more balanced distribution, and enhance diversity within the dataset.

In the next phase, we conducted exploratory data analysis, which encompassed the examination of a correlation matrix and a pie chart analysis that are shown in Fig. 1 and 2 respectively. This step allowed us to gain insights into the relationships between various variables and understand the distribution of data. Moving forward, the dataset was partitioned using the 10 K-fold cross-validation technique to ensure robust model assessment and validation. In the fourth step, we fine-tuned ten traditional machine learning models-namely, AB, SVM, LGB, RC, ET, DT, RF, LR, EGB, and GB. These models were optimized through a randomized search technique to identify the most suitable hyperparameters.

The best-performing model, as determined by the model evaluation, was found to be the Extra Trees (ET) classifier for this dataset. The whole procedures are shown in Fig. 3.

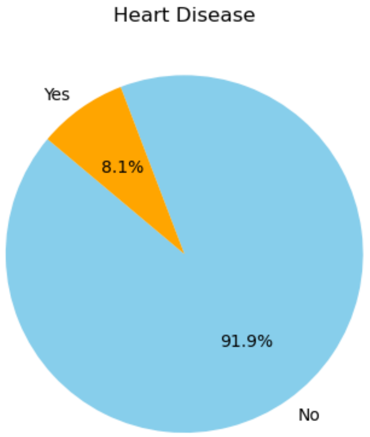

Fig.2. Pie chart of heart diseases counts.

C. Model Evaluation

Model evaluation is a critical step in machine learning [16], and in our study, we employed various metrics to comprehensively assess the performance of our models. These metrics provide a well-rounded view of the model's effectiveness. The formulae of the measuring parameters are given in equations 1 to 6.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}*100 \qquad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}*100 \qquad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}*100 \qquad (3)$$

$$\text{F1 Score} = 2 * \left(\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}\right) * 100 \qquad (4)$$

$$\text{Cohen's Kappa (Kappa)}, k = \frac{p_0 - p_e}{1 - p_e} \qquad (5)$$

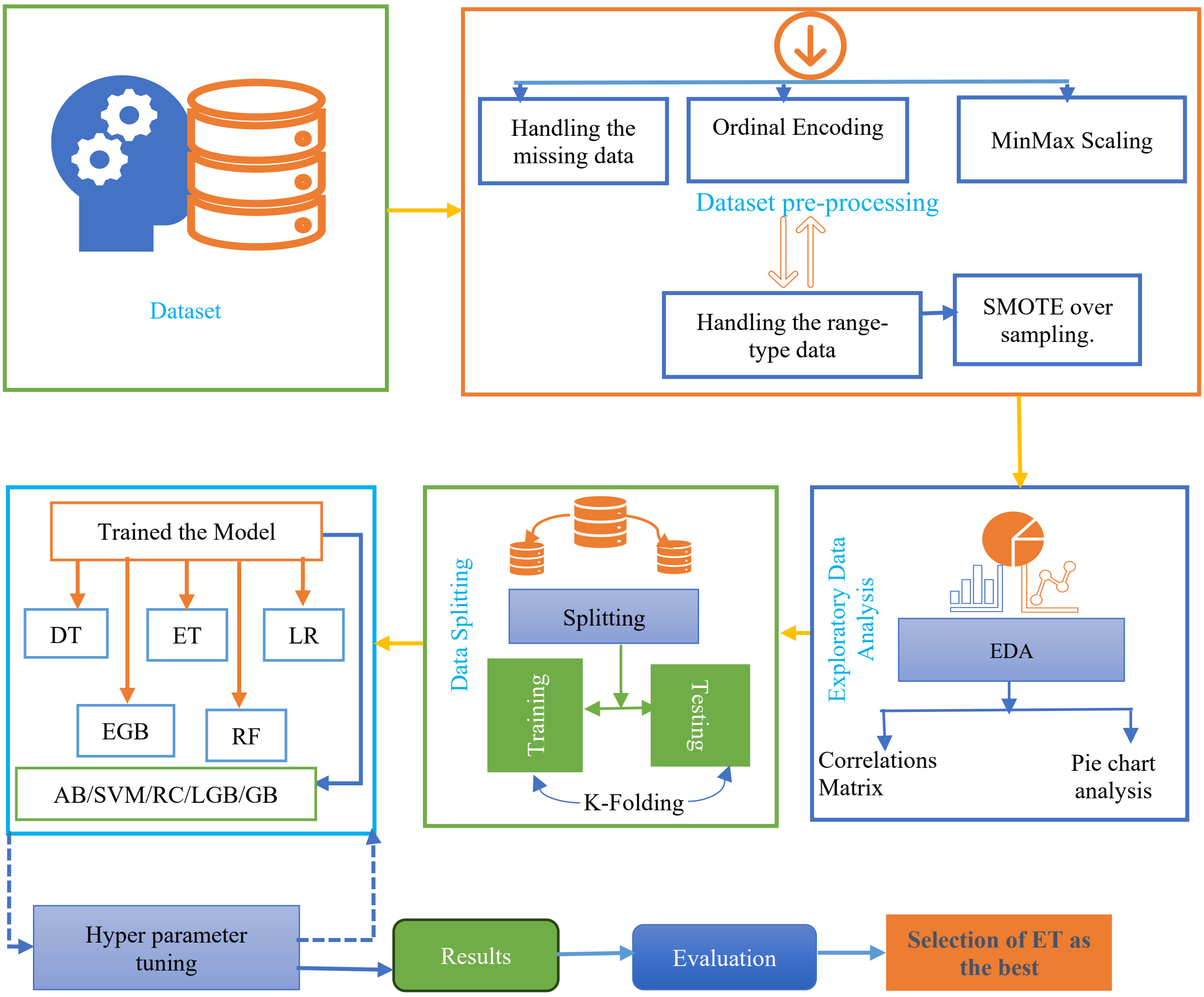

Fig 3. Proposed method for this research.

Table 1. Tuned hyper parameters of different ML models.

Models Name	Tuned parameters
ET	bootstrap=False, criterion='gini', max_features='sqrt', min_samples_leaf=1, min_samples_split=2, n_estimators=100, n_jobs=-1, random_state=6978
RF	bootstrap=True, criterion='gini', max_features='sqrt', min_samples_leaf=1, min_samples_split=2, n_estimators=100, n_jobs=-1, random_state=6978
DT	criterion='gini', min_samples_leaf=1, min_samples_split=2, random_state=6978, splitter='best'
EGB	booster='gbtree', enable_categorical=False, min_child_weight=None, missing=nan, n_estimators=None, n_jobs=-1, num_parallel_tree=None, objective='binary:logistic'
LGB	boosting_type='gbdt', colsample_bytree=1.0, importance_type='split', learning_rate=0.1, max_depth=-1, min_child_samples=20, min_child_weight=0.001, n_estimators=100, n_jobs=-1, num_leaves=31, random_state=6978, subsample=1.0, subsample_for_bin=200000
GB	criterion='friedman_mse', learning_rate=0.1, loss='log_loss', max_depth=3, min_samples_leaf=1, min_samples_split=2, n_estimators=100, random_state=6978, subsample=1.0, tol=0.0001, validation_fraction=0.1 warm_start=False
AD	algorithm='SAMME.R', base_estimator='deprecated', estimator=None, learning_rate=1.0, n_estimators=50, random_state=6978
RC	alpha=1.0, class_weight=None, copy_X=True, fit_intercept=True, max_iter=None, positive=False, random_state=6978, solver='auto',tol=0.0001
SVM	alpha=0.0001, average=False, early_stopping=False, epsilon=0.1, eta0=0.001, fit_intercept=True, l1_ratio=0.15, learning_rate='optimal', loss='hinge', max_iter=1000, n_iter_no_change=5, n_jobs=-1, penalty='l2', power_t=0.5, random_state=6978, shuffle=True, tol=0.001, validation_fraction=0.1
LR	C=1.0, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=1000, multi_class='auto', penalty='l2', random_state=6978, solver='lbfgs', tol=0.0001, warm_start=False

Table 2. Performance metrics of the parameters.

Models Name	Accuracy	AUC	Precision	Recall	F1 Score	Kappa	MCC
SVM	0.7731 ±0.0022	0.0000 ±0.0000	0.8257 ±0.0262	0.7477 ±0.0122	0.7843 ±0.0057	0.5462 ±0.0045	0.5500 ±0.0051
LR	0.7733 ±0.0025	0.8431 ±0.0020	0.8064 ±0.0029	0.7563 ±0.0025	0.7805 ±0.0024	0.5466 ±0.0049	0.5478 ±0.0049
RC	0.7742 ±0.0025	0.0000 ±0.0000	0.8233 ±0.0027	0.7497 ±0.0025	0.7848 ±0.0024	0.5484 ±0.0051	0.5511 ±0.0051
DT	0.9222 ±0.0010	0.9222 ±0.0010	0.9300 ±0.0013	0.9157 ±0.0013	0.9228 ±0.0009	0.8443 ±0.0019	0.8444 ±0.0019
AD	0.9333 ±0.0011	0.9787 ±0.0004	0.9049 ±0.0011	0.9593 ±0.0022	0.9313 ±0.0011	0.8665 ±0.0023	0.8679 ±0.0023
GB	0.9465 ±0.0009	0.9831 ±0.0003	0.9071 ±0.0013	0.9847 ±0.0013	0.9443 ±0.0010	0.8929 ±0.0018	0.8957 ±0.0018
LGB	0.9539 ±0.0008	0.9853 ±0.0003	0.9109 ±0.0014	0.9967 ±0.0005	0.9518 ±0.0008	0.9078 ±0.0015	0.9112 ±0.0014
EGB	0.9543 ±0.0007	0.9850 ±0.0004	0.9141 ±0.0012	0.9941 ±0.0007	0.9524 ±0.0007	0.9087 ±0.0013	0.9117 ±0.0013
RF	0.9555 ±0.0009	0.9856 ±0.0003	0.9188 ±0.0016	0.9916 ±0.0005	0.9538 ±0.0010	0.9110 ±0.0019	0.9134 ±0.0018
ET	**0.9559 ±0.0006**	**0.9870 ±0.0003**	**0.9262 ±0.0012**	**0.9847 ±0.0012**	**0.9546 ±0.0006**	**0.9119 ±0.0012**	**0.9135 ±0.0012**

where $p_o = obsereved\ agremeent, p_e = Expected\ agrement$

$$\text{Matthews Correlation Coefficient (MCC)} = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \qquad (6)$$

Where, TP= True positive; TN=True negative; FP= False positive; FN= False negative

IV. Results and discussion

Our proposed method involves the fine-tuning of hyperparameters and evaluating model performance. Table 1 presents the results of hyperparameter tuning. These tuned hyperparameters are then used to define and train the models on the cross-validation dataset, and the test results are summarized in Table 2, where we are dealing with a three-category classifier.

In the first category, RC, SVM, and LR achieve lower accuracy, precision, recall, F1 score, MCC, Kappa, and AUC, with their accuracy hovering around 77.32% to 77.42%. RC stands out with the highest F1 score at 78.48%, along with superior MCC and Kappa values, making it the best classifier among the three. Moving to the second category, DT, AD, and GB produce more promising results with accuracy ranging from 92.22% to 94.65%. Their precision, recall, and F1 score also surpass the first category. Among these, GB emerges as the top performer with an impressive F1 score of 94.43%, accompanied by superior Kappa and MCC values. In the third category, ET, RF, EGB, and LGB outshine the rest, achieving over 95% accuracy and AUC values exceeding 98%. Their precision, recall, and F1 score outperform the previous two categories. ET excels with an F1 score of 95.46%, making it the highest among this category and all the classifiers, along with the highest MCC and Kappa values exceeding 91%.

Notably, the standard deviations of each parameter for all models are quite small, indicating their robustness. Specifically, for the Extra Tree classifier (ET), the deviations for accuracy and F1 score are only 0.006, reaffirming its exceptional robustness. This suggests that ET is exceptionally stable in its performance. Analyzing the results logically, it becomes evident that ensemble models (LGB, EGB, RF, ET) benefit from their collective approach, enhancing their predictive power and resilience against overfitting. Considering accuracy, robustness, and other performance metrics, ET emerges as the top-performing model for this dataset. A cooperative study on this dataset has been conducted and concluded in table 3. It also shows the superiority of our proposed ET model.

V. Conclusion and future work

In conclusion, our study underscores the critical importance of accurately assessing and predicting heart disease risk factors, given its profound impact on public health. We employed an ensemble approach, specifically the Extra Tree Classifier, to achieve remarkable results, demonstrating the model's superiority in terms of accuracy, AUC, precision, recall, F1 Score, kappa, and MCC. The ET model's minimal standard deviation further attests to its reliability.

Looking ahead, there are several avenues for future research. First, we can explore the integration of additional health-related parameters and data sources to enhance the model's predictive power. Moreover, the use of deep learning techniques and the analysis of unstructured medical data, such as medical images, could provide valuable insights for more comprehensive heart disease risk assessment. Additionally, developing a user-friendly interface or mobile application for real-time risk prediction and disease management can be a promising direction. This would empower individuals to take proactive measures for their cardiac health. Furthermore, continuous model refinement and updates are crucial to adapt to evolving healthcare scenarios and ensure the most accurate predictions. Collaboration with medical experts and institutions for real-world validation is another vital step in the translation of our research into practical clinical applications.

However, while our study has achieved significant progress in heart disease risk assessment, there is a promising future ahead, marked by innovation, collaboration, and a continued commitment to improving public health outcomes.

Table 3. Comparison with the recent work on the same dataset.

Parameters	**Reference paper [14] (%)**	**Our proposed ET model (%)**
Accuracy	74	95.59
Precision	59	92.62
Recall	76	98.47
F1 score	74	95.46

REFERENCES

[1] Labarthe, Darwin R. “Epidemiology and Prevention of Cardiovascular Diseases: A Global Challenge: A Global Challenge”, 2010.

[2] Arunachalam, S. “Cardiovascular disease prediction model using machine learning algorithms”, Int. J. Res. Appl. Sci. Eng. Technol, vol. 8, pp.1006-1019, 2020.

[3] Alalawi, H.H. and Alsuwat, M.S., “Detection of cardiovascular disease using machine learning classification models”, International Journal of Engineering Research & Technology, vol. 10(7), pp.151-7, 2021.

[4] Princy, R.J.P., Parthasarathy, S., Jose, P.S.H., Lakshminarayanan, A.R. and Jeganathan, S. “Prediction of cardiac disease using supervised machine learning algorithms” In 2020 4th international conference on intelligent computing and control systems (ICICCS), pp. 570-575, 2020.

[5] Rubini, P.E., Subasini, C.A., Katharine, A.V., Kumaresan, V., Kumar, S.G. and Nithya, T.M., “A cardiovascular disease prediction using machine learning algorithms”, Annals of the Romanian Society for Cell Biology, pp.904-912, 2021.

[6] “World Health Organization, Cardiovascular Diseases, WHO, Geneva, Switzerland”, Online Link: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed on 4 Nov, 2023)

[7] R. Mythili, Dr. A.S. Aneetha, “A COMPARATIVE STUDY OF DEEP LEARNING ALGORITHMS USED IN DETECTING CARDIOVASCULAR DISEASES”, ISSN: 2096-3246, Volume 55, 2023.

[8] Louridi, N., Amar, M. and El Ouahidi, B. “Identification of cardiovascular diseases using machine learning”, In 2019 7th mediterranean congress of telecommunications (CMT), pp. 1-6, 2019.

[9] Al-Batah, M.S., Alzboon, M.S. and Alazaidah, R., “Intelligent Heart Disease Prediction System with Applications in Jordanian Hospitals”, International Journal of Advanced Computer Science and Applications, vol. 14, pp. 9, 2023.

[10] Muhammad, L.J., Al-Shourbaji, I., Haruna, A.A., Mohammed, I.A., Ahmad, A. and Jibrin, M.B., “Machine learning predictive models for coronary artery disease”, SN Computer Science, vol. 2(5), p.350, 2021.

[11] Guarneros-Nolasco, L.R., Cruz-Ramos, N.A., Alor-Hernández, G., Rodríguez-Mazahua, L. and Sánchez-Cervantes, J.L., “Identifying the main risk factors for cardiovascular diseases prediction using machine learning algorithms”, Mathematics, vol. 9(20), p.2537, 2021.

[12] R. Bharti, A. Khamparia ,Md. Shabaz ,G. Dhiman , S. Pande and Parneet Singh,” Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning”, Computational Intelligence and Neuroscience , 2021.

[13] R. Kasabe, Dr. Prof. G. Narang,” Cardiovascular Ailments Prediction and Analysis Based On Deep Learning Techniques”, IJEAP, vol. 1, No. 2, pp. 174~178, 2021.

[14] A. H. Khan, M. Hussain and Md. K. Malik, “Cardiac Disorder Classification by Electrocardiogram Sensing Using Deep Neural Network”, Hindawi, Complexity, 2021.

[15] S. Dalal, P. Goel , E. M. Onyema ,A. Alharbi, A. Mahmoud, M. A. Algarni and Halifa Awal, “Application of Machine Learning for Cardiovascular Disease Risk Prediction”, Hindawi, Computational Intelligence and Neuroscience , 2023.

[16] S. Subramani, N. Varshney, M. V. Anand, M. Soudagar, L. A. Al-keridis, T. K. Upadhyay, N. Alshammari, M. Saeed, K. Subramanian, K. Anbarasu and K. Rohini, “cardiovascular diseases prediction by machine learning incorporation with deep learning”, frontiers, 2023.

[17] Mezzatesta, S., Torino, C., De Meo, P., Fiumara, G. and Vilasi, A., “A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis”, Computer methods and programs in biomedicine, 177, pp.9-15, 2019.

[18] Boursalie, O., Samavi, R. and Doyle, T.E., “M4CVD: Mobile machine learning model for monitoring cardiovascular disease” Procedia Computer Science, vol. 63, pp.384-391, 2015.

[19] Talukder, Md Simul Hasan, Rejwan Bin Sulaiman, and Mouli Bardhan Paul Angon. “Unleashing the Power of Extra-Tree Feature Selection and Random Forest Classifier for Improved Survival Prediction in Heart Failure Patients” arXiv preprint arXiv:2308.05765, 2023.

[20] Lupague, R.M.J.M., Mabborang, R.C., Bansil, A.G. and Lupague, M.M., 2023. Integrated Machine Learning Model for Comprehensive Heart Disease Risk Assessment Based on Multi-Dimensional Health Factors. European Journal of Computer Science and Information Technology, 11(3), pp.44-58.

[21] “Cardiovascular Diseases Risk Prediction Dataset” online Link: https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?fbclid=IwAR0uSUu9SqBlc0f5C98DIId9InjJ4CQgu5N5PIc3N43V9rofgThzA4lkNo (accessed on 4 Nov, 2023)

[22] Talukder, M. S. H., & Akter, S, “An improved ensemble model of hyper parameter tuned ML algorithms for fetal health prediction” International Journal of Information Technology, pp.1-10, 2023.