Heart Health Status Detection Using Ensemble Learning with Hyperparameter Optimization

Abstract

Heart diseases are a group of medical conditions that disrupt the heart's normal functioning. Cardiovascular disease (CVD) is a prevalent condition caused by the narrowing or blockage of blood vessels, leading to chest pain or heart attacks. Another type of heart disease is congestive heart failure, which occurs when the heart cannot pump enough blood to the body, resulting in fluid accumulation in the lungs. Timely diagnosis of heart disease can help to reduce mortality rates.

To improve the performance of machine learning (ML) models for CVD detection and severity classification, optimization techniques were utilized. The application of ML to medical records has shown promise in predicting patient survival rates and identifying heart conditions. To build and compare ML models, such as Naive Bayes, Logistic Regression, Gradient Booster, Support Vector Machine, Decision Tree, Random Forest, and Ensemble models, the Cleveland Dataset publicly available in the UCI Repository was used. The combination of AI with medical records and imaging has demonstrated predictive abilities, making it a useful tool for clinical prediction and feature ranking. Our final model achieved an accuracy of about 90, which was significantly better than the accuracy range of 78-82 when we began comparing models. Simple ML techniques utilized in related work were not able to achieve such high accuracies.

I. INTRODUCTION

Heart diseases are a life-threatening condition that has a significant impact on individuals. Due to the accelerated pace of life, increased portion sizes, and inactivity, the majority of individuals consistently disregard their health. Moreover, as a result of environmental deterioration, these variables can contribute to the increasing prevalence of heart failure in the future. If people ignored the issue of heart health, it would eventually result in death. In order to forecast heart health, different researchers have employed a variety of data collection and analysis techniques in recent years [4]. These data consist of electronic health record (EHR) data of patients with heart failure from various hospitals in various countries, the Cleveland heart disease dataset, biomedical science datasets from UCI, etc.

Given the importance of a crucial organ such as the heart, predicting heart health status has become a top goal for doctors and physicians. However, until now, it has been difficult to predict heart health with high accuracy in clinical practice. Particularly when applied to medical records, machine learning can be a useful technique for predicting the survival of patients with heart health detection systems symptoms and identifying the most relevant characteristics (or risk factors) that may lead to heart failure. Not only may scientists utilize machine learning for clinical prediction, but also for feature rating. When used with medical information or combined with imaging, artificial intelligence demonstrates its prediction ability in particular [5].

We intend to accomplish this by employing many data mining techniques to estimate the survival of patients and then to prioritise the most relevant characteristics contained in their medical records. As a main conclusion, we demonstrate that the best prediction performance can be achieved by performing hyperparameter tuning to the dataset for getting the features importance and ensemble learning so as to combine the best four machine learning methods that work best with this type of dataset.

II. LITERATURE SURVEY

In this section, we give a short review of recent related works.

In this research article Abdallah Abdellatif, Hamdan Abdellatef, et Al. have proposed employing optimization methods to improve the ML model performance for CVD detection and severity level classification. And to achieve this they are using hyperparameter optimization, SMOTE and ExtraTreesClassifier on two datasets [25].

Ganjar Alfian, Jongtae Rhee et al, have proposed a heart disease prediction model (HDPM) for a CDSS which consists of DBSCAN to detect and eliminate the outliers, a SMOTE-ENN to balance the training data distribution and XGBoost to predict heart disease and they compared the results they achieved with other models. [26]

In this research article the author has proposed to investigate the effect of sample age and prediction resolution on the performance of Myocardial infarction risk prediction models. Their experiments indicate that both sample age and prediction resolution do not have a significant impact on prediction models developed using subjects aged 65 and above. Their study suggests the use of different prediction resolution to provide a more detailed health screening of elderly subjects.[23]

In this research article the author has proposed to use a switching vector autoregressive framework to systematically learn and identify a collection of vital sign time series dynamics, which are possibly recurrent within the same patient. They use HR and BP dynamics of an ICU from a dataset [18].

The authors [2] use electronic health record (EHR) data in conjunction with Machine Learning and Deep Learning models. The results indicate that it is possible for novel machine learning models to improve the accuracy of a model's predictions.

In [3], the dataset derived from EHR data and included 26,575 individuals with heart failure in 2018 is utilised. In conclusion, the results indicate that age, creatinine, body mass index, and blood pressure levels were significant predictors of mortality within one year among heart failure patients.

Minh Tuan [9], utilised these data from the Faisalabad Institute of Cardiology and the Faisalabad United Hospital regarding patients with heart failure. Analyze using a multilayer perceptron neural network. With an accuracy of 88 percent, the heart failure dataset beat prior studies in predicting heart failure.

III. PROPOSED APPROACH

Our study is comprised of two key parts: the pre-processing phase, in which we first clean our dataset, remove any outliers if present then identify the most relevant features through feature ranking, also employing Exploratory Data Analysis to better understand the data we are working with. And the Model Building phase, in which we compare the different classification algorithms based on precision and recall metrics and the top four algorithms are automatically selected by model for the calibration of final model.

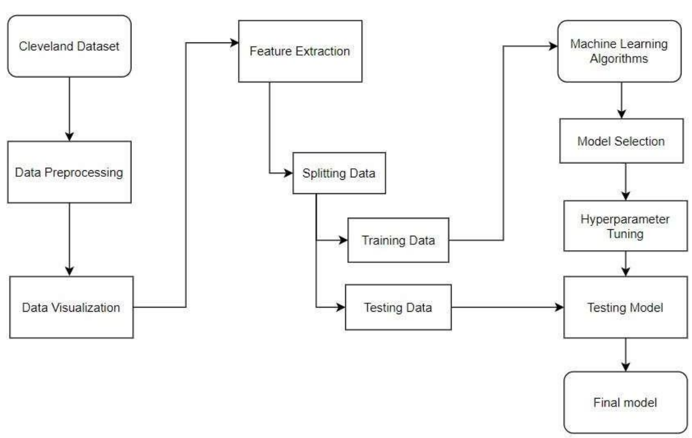

Fig 1. Architecture Diagram

A. DATASET COLLECTION

We utilise a dataset of individuals who have undergone heart disease detection analyses and testing. The data set is a matrix, with rows representing patients and columns representing testable characteristics as shown in Table1.

Table 1. Dataset

	0	1	2	3	4
Age	40	49	37	48	54
Sex	M	F	M	F	M
ChestPainType	ATA	NAP	ATA	ASY	NAP
RestingBP	140	160	130	138	150
Cholesterol	289	180	283	214	195
FastingBS	0	0	0	0	0
RestingECG	Normal	Normal	ST	Normal	Normal
MaxHR	172	156	98	108	122
ExerciseAngina	N	N	N	Y	N
Oldpeak	0.000000	1.000000	0.000000	1.500000	0.000000
ST_Slope	Up	Flat	Up	Flat	Up
HeartDisease	0	1	0	1	0

B. DATA PRE-PROCESSING

Data pre-processing is a crucial stage in Machine Learning since the quality of the data and the relevant information that can be extracted from it directly influences our model's capacity to learn; consequently, it is crucial that we pre-process our data prior to feeding it to the model.

C. FEATURES SELECTION

Features available in the Dataset:

- Age: age of the patient [years]
- Sex: sex of the patient [M: Male, F: Female]
- ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- ExerciseAngina: exercise-induced angina [Y: Yes, N: No]

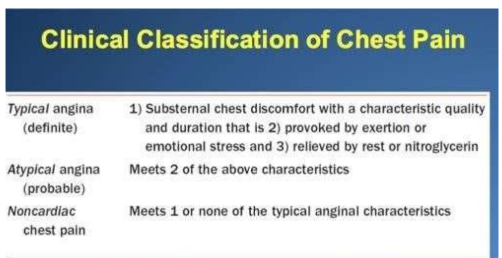

Fig 2. Classification of Chest Pain

- RestingBP: resting blood pressure [mm Hg]
- MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
- Cholesterol: serum cholesterol [mm/dl]
- FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- Oldpeak: oldpeak = ST [Numeric value measured in depression]
- ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- HeartDisease: output class [1: heart disease, 0: Normal]

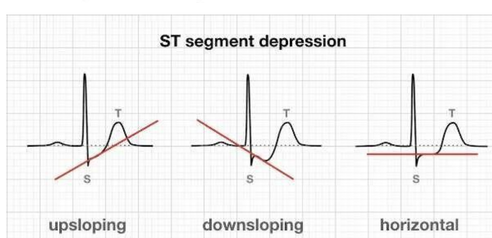

Fig 3. ST Slope

Using the Correlation matrix, EDA, Feature Permeation and partial dependence we select the most pertinent characteristics from a huge range of data features. To determine the relationships between attributes, we exclusively select highly dependent attributes in order to apply Machine Learning methods with greater precision.

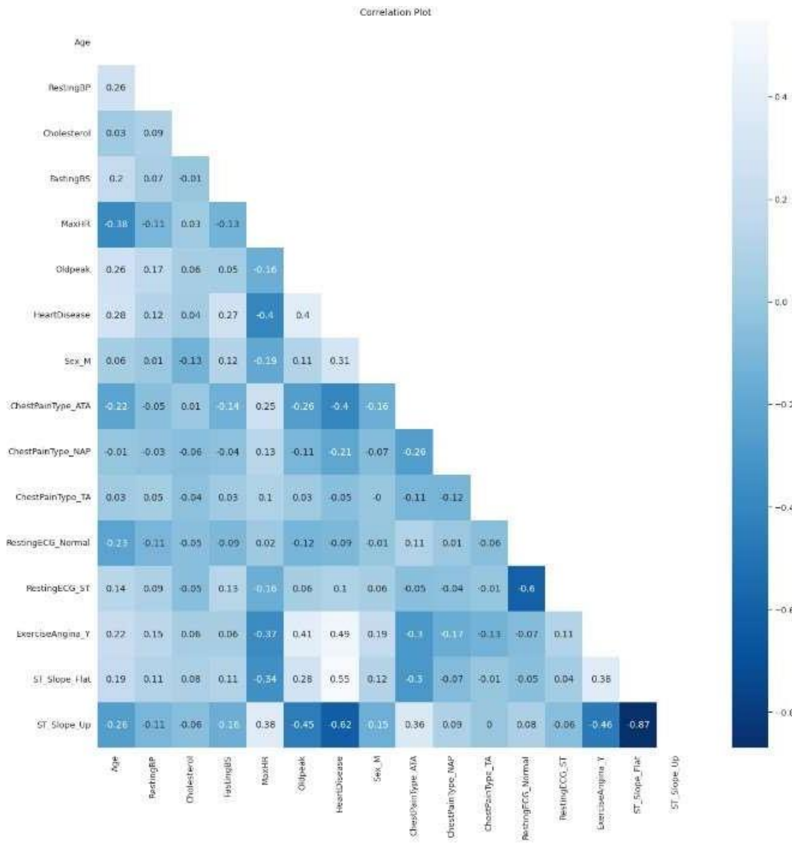

Fig 4: Feature Importance Heatmap

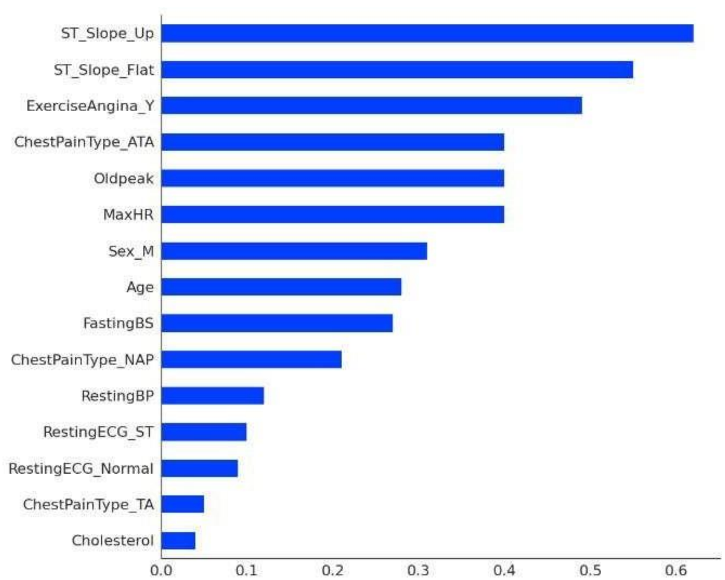

Fig 5. Feature Importance Graph

D. SPLITTING DATASET

To train the Machine Learning algorithm, we identify the target column in the dataset and then partition it into two smaller data sets. Training-set is the Test-set for testing the algorithm.

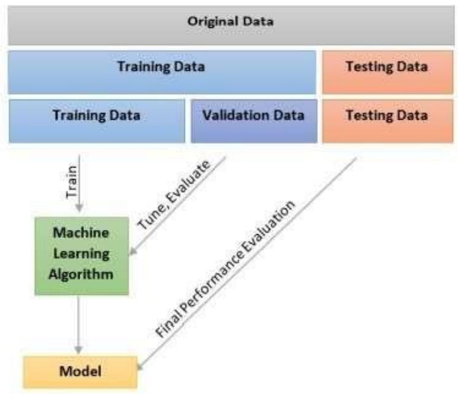

Fig 6. Model Approach

E. MODELLING

This is the phase in which we apply and evaluate the selected algorithms. Our first step is comparing the top classifying algorithms (SVM, Random Forest, Gradient Booster, Extra Trees, CatBoost, Logistic Regression, lightgbm, Naive Bayes, Decision Tree and various Ensemble Learning Methods).

Table 2. Model Comparison

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.8674	0.9093	0.9077	0.8659	0.8851	0.7283	0.7325	0.2710
gbc	Gradient Boosting Classifier	0.8643	0.9206	0.9025	0.8700	0.8639	0.7209	0.7260	0.0520
rf	Random Forest Classifier	0.8534	0.9125	0.8967	0.8525	0.8730	0.6997	0.7033	0.3060
catboost	CatBoost Classifier	0.8565	0.9187	0.8941	0.8605	0.8755	0.7062	0.7101	1.2030
lr	Logistic Regression	0.8534	0.9120	0.8912	0.8578	0.8730	0.6997	0.7030	0.5110
lda	Linear Discriminant Analysis	0.8472	0.9117	0.8829	0.8539	0.8671	0.6874	0.6901	0.0090
lightgbm	Light Gradient Boosting Machine	0.8503	0.9093	0.8716	0.8697	0.8689	0.6942	0.6962	0.1050
nb	Naive Bayes	0.8488	0.9120	0.8663	0.8666	0.8658	0.6925	0.6937	0.0090
dt	Decision Tree Classifier	0.7413	0.7377	0.7687	0.7709	0.7688	0.4749	0.4766	0.0110

F. ALGORITHMS APPLICATION

In this step, the model determines that which algorithms are best for further model building based on the input features. Our strategy is based on the application of ensemble learning algorithms to data sets of varying sizes in order to assess the accuracy of each algorithm and, most importantly, to identify the one that is more stable throughout training and provides high accuracy with certainty.

a) Extra Trees Classifier

Extra Trees Classifier is an ensemble method that uses multiple decision trees to make predictions. It is efficient, less prone to overfitting, and can handle high-dimensional datasets. However, it can be sensitive to noisy data and difficult to interpret. It came on top in terms of the metrics used by compare models in built function of PyCaret[12].

b) Gradient Boosting Classifier

Gradient Boosting Classifier is an ensemble method that combines multiple weak decision trees into a strong predictive model. It can handle both numerical and categorical data, and can handle missing data. However, it can be computationally expensive and sensitive to overfitting [15].

c) Random Forest Classifier

Random Forest Classifier is an ensemble method that combines multiple decision trees into a strong predictive model. It is less prone to overfitting, can handle high-dimensional datasets, and can handle missing data. However, it can be computationally expensive and difficult to interpret [24].

d) CatBoost Classifier

CatBoost Classifier is a gradient boosting algorithm designed to handle categorical variables in machine learning. It uses a unique algorithm to process categorical features, which reduces the need for pre-processing and improves model accuracy. It is computationally efficient and has become popular in various domains, including finance, e-commerce, and image recognition [13].

e) Logistic Regression

Logistic Regression is a statistical model used for binary classification problems. It estimates the probability of an event occurring based on a set of input features. It uses a logistic function to transform the input features into probabilities, which are used to predict the binary outcome. It is a simple and efficient algorithm used in various domains, including healthcare, marketing, and finance [20].

G. TESTING ALGORITHMS

We test the algorithms on the test set using the confusion matrix and the accuracy ratio as opposed to manual exploration. A confusion matrix is a table frequently used to describe the performance of a classifier on a collection of known-true test data.

H. TUNING HYPERPARAMETERS

We used tune_model function from pycaret for hyperparameter tuning. This tune_model function tunes the hyperparameters of a given estimator. The output of this function is a score grid with CV scores by fold of the best selected model based on optimized parameters.

I. BUILDING THE ENSEMBLE MODEL

We employed to use Stacking, Soft Voting and Hard Voting Ensemble Models. Then we calibrated our model and finally built our Final model. We used in-build functions from PyCaret for blending and calibrating our models.

a) Stacking

Stacking is an ensemble machine learning technique used for improving model accuracy. It involves training multiple models, then using a meta-model to combine their predictions. The meta-model takes the predictions of the individual models as input features and learns how to best combine them [24].

Stacking can improve the accuracy and robustness of the model by reducing the impact of individual model errors and leveraging the strengths of different models. It is used in various domains, including finance, healthcare, and natural language processing.

b) Soft Voting Classifier

Soft Voting Classifier is an ensemble machine learning algorithm used for classification problems. It combines multiple individual classifiers to make a final prediction by taking the average of the predicted probabilities of the individual models. This approach can improve the accuracy and robustness of the model by reducing the impact of individual model errors. It is widely used in various domains, including finance, healthcare, and image recognition [25].

c) Hard Voting Classifier

Hard Voting Classifier is an ensemble machine learning algorithm used for classification problems. It combines multiple individual classifiers to make a final prediction by taking the majority vote of the individual models. This approach can improve the accuracy and robustness of the model by leveraging the strengths of different models. It is widely used in various domains, including finance, healthcare, and image recognition [24].

After comparison of the performance of the three ensemble models by recall, the soft voting model had the best performance. Therefore, the soft voting model is selected as the final model and calibration is performed.

IV. RESULTS

After employing to various methodologies of improving our dataset like data cleaning, feature ranking, data visualization along with hyperparameter tuning in our models. We were successfully able to improve our precision metrics, especially accuracy by a reasonable extent. We were able to improve our initial accuracy which was in the range of 78-82 to the range of 86-88 after using various ensemble learning techniques. But we wanted to further improve our metrics which we achieved by the boosting technique. With the help of this technique our accuracy was boosted above 90.

The severity of a CVD can be classified into three outputs through our model namely; Low Risk – which means currently the heart is in an optimal condition, Moderate Risk – There are some abnormalities in the input data and there is moderate risk of suffering from a CVD, High Risk – the entered data is highly deviated from the safe limits suggesting high chances of suffering from CVDs and poses high risk of getting a heart failure if not treated immediately. The proposed model is evaluated and benchmarked against earlier works.

V. CONCLUSION

The research paper introduces a decision support system that utilizes the Optimized Ensemble model to predict the occurrence and severity of CVDs. The results indicate that optimizing hyperparameters and feature ranking enhanced the model's performance. The research team employed data balancing, classification, and hyperparameter optimization techniques to achieve this. Additionally, the Streamlit library enabled straightforward visualization of the model’s outcomes. The study aimed to create an optimal machine learning framework, including outlier detection and removal, feature selection, hyperparameter tuning and ensemble learning to improve the detection and severity level detection of various CVDs using real-time clinical data.

REFERENCES

[1] Chicco, Davide, and Giuseppe Jurman. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." BMC medical informatics and decision making 20.1 (2020):

[2] Guo, Aixia, et al. "Heart Failure Diagnosis, Readmission, and Mortality Prediction Using Machine Learning and Artificial Intelligence Models." Current Epidemiology Reports (2020)

[3] Guo, A., et al. "The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation. Front. Digit." Front. Digit. Health 2 (2020): 576945.

[4] Olsen, Cameron R., et al. "Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure." American Heart Journal (2020).

[5] Ali, L., and S. A. C. Bukhari. "An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction." IRBM (2020).

[6] Fang, Hao, Cheng Shi, and Chi-Hua Chen. "BioExpDNN: Bioinformatic Explainable Deep Neural Network." 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2020

[7] Eletter, Shorouq, et al. "Building an Intelligent Telemonitoring System for Heart Failure: The Use of the Internet of Things, Big Data, and Machine Learning." 2020 21st International Arab Conference on Information Technology (ACIT). IEEE, 2020.

[8] Kim, Young-Tak, et al. "A Comparison of Oversampling Methods for Constructing a Prognostic Model in the Patient with Heart Failure." 2020 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2020.

[9] Le, Minh Tuan, et al. "Predicting heart failure using deep neural network." 2020 International Conference on Advanced Technologies for Communications (ATC). IEEE, 2020.

[10] Venkatalakshmi, B., Shivsankar, M.: Heart disease diagnosis using predictive data mining. International Journal of Innovative Research in Science, Engineering and Technology 3(3), 1873–7 (2014)

[11] Peter, T.J., Somasundaram, K.: An empirical study on prediction of heart disease using classification data mining techniques. In: IEEE-International conference on advances in engineering, science and management (ICAESM-2012). pp. 514–518. IEEE (2012)

[12] Gavhane, A., Kokkula, G., Pandya, I., Devadkar, K.: Prediction of heart disease using machine learning. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). pp. 1275–1278. IEEE (2018)

[13] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, ‘‘MIFH: A machine intelligence framework for heart disease diagnosis,’’ IEEE Access, vol. 8, pp. 14659–14674, 2020.

[14] B. A. Tama and S. Lim, ‘‘A comparative performance evaluation of classification algorithms for clinical decision support systems,’’ Mathematics, vol. 8, no. 10, p. 1814, Oct. 2020.

[15] B. A. Tama, S. Im, and S. Lee, ‘‘Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble,’’ BioMed Res. Int., vol. 2020, pp. 1–10, Apr. 2020.

[16] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, ‘‘HDPM: An effective heart disease prediction model for a clinical decision support system,’’ IEEE Access, vol. 8, pp. 133034–133050, 2020.

[17] M. Waqar, H. Dawood, H. Dawood, N. Majeed, A. Banjar, and R. Alharbey, ‘‘An efficient SMOTE-based deep learning model for heart attack prediction,’’ Sci. Program., vol. 2021, pp. 1–12, Mar. 2021.

[18] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, ‘‘Improving the prediction of heart failure patients’ survival using SMOTE and effective data mining techniques,’’ IEEE Access, vol. 9, pp. 39707–39716, 2021.

[19] N. Salari, S. Shohaimi, F. Najafi, M. Nallappan, and I. Karishnarajah, ‘‘A novel hybrid classification model of genetic algorithms, modified knearest neighbor and developed backpropagation neural network,’’ PLoS ONE, vol. 9, no. 11, Nov. 2014, Art. no. e112987.

[20] W. Wiharto, H. Kusnanto, and H. Herianto, ‘‘Performance analysis of multiclass support vector machine classification for diagnosis of coronary heart diseases,’’ 2015, arXiv:1511.02352.

[21] N. Khateeb and M. Usman, ‘‘Efficient heart disease prediction system using K-nearest neighbor classification technique,’’ in Proc. Int. Conf. Big Data Internet Thing (BDIOT), 2017, pp. 21–26.

[22] G. Magesh and P. Swarnalatha, ‘‘Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction,’’ Evol. Intell., vol. 14, no. 2, pp. 583–593, Jun. 2021.

[23] H. B. Kibria and A. Matin, ‘‘The severity prediction of the binary and multi-class cardiovascular disease—A machine learning-based fusion approach,’’ Comput. Biol. Chem., vol. 98, Jun. 2022, Art. no. 107672.

[24] S. Shin, B. Ko, and H. So, ‘‘Noncontact thermal mapping method based on local temperature data using deep neural network regression,’’ Int. J. Heat Mass Transf., vol. 183, Feb. 2022, Art. no. 122236.

[25] Abdallah Abdellatif, Hamdan Abdellatef, Jeevan Kanesan, Chee-Onn Chow, Joon Huang Chuah, And Hassan Muwafaq Gheni,” An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods”, IEEE Access, vol. 9, 2022.

[26] Norma Latif Fitriyani, Muhammad Syafrudin, Ganjar Alfian, And Jongtae Rhee, “HDPM: An Effective Heart Disease Prediction Model For A Clinical Decision Support System”, IEEE Access, vol. 8, 2020.