Journal of Robotics and Control (JRC)
Volume 5, Issue 3, 2024
ISSN: 2715-5072, DOI: 10.18196/jrc.v5i3.21606

Heart Disease Prediction Using Hybrid Machine Learning: A Brief Review

Mohammed Ahmed [1*], Idress Husien [2]
[1, 2] College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq
Email: [1] Stcm22011@uokirkuk.edu.iq, [2] idress@uokirkuk.edu.iq
*Corresponding Author

***Abstract*—Cardiovascular disease is a widespread and potentially fatal condition that requires proactive preventive measures and efficient screening approaches on a global scale. To tackle this issue, recent studies have investigated novel machine-learning frameworks that propose to diagnose and forecast cardiovascular disease by capitalizing on enormous datasets and predictive patterns linked to this condition. The research contribution is a thorough examination and implementation of ensemble learning and other hybrid machine-learning techniques for heart disease prediction. By employing ensemble learning on datasets including The Cleveland heart disease dataset and The IEEE Dataport heart diseases dataset such as age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate, and four types of chestpain. To predict heart disease, our methodology integrates numerous machine learning models. By capitalising on the merits of specific algorithms while addressing their drawbacks, this approach yields a predictive model that is more resilient. The findings of our research exhibit encouraging outcomes in the realm of heart disease prediction, attaining enhanced precision and dependability in contrast to discrete algorithms. Through the utilisation of ensemble learning, we successfully discerned predictive patterns that are linked to heart disease, thereby augmenting the capabilities of diagnostics. In summary, the findings of our study emphasise the considerable potential of ensemble techniques within the realm of machine learning for the advancement of cardiac disease prediction. By providing a more dependable method for rapid diagnosis and prognosis of cardiac disease, this strategy has substantial ramifications for healthcare practices.**

Keywords—Heart Disease; Hybrid Machine; Machine Learning; Supervised; Unsupervised.

I. INTRODUCTION

Health data must be transformed to account for behavior and can be monitored globally. A considerable number of individuals encounter difficulties when searching for health information about maladies online, and diagnosing various medications, which can be time-consuming, and wasting money [1].

It is the vital function of the heart to circulate oxygen-rich blood throughout the body via the capillaries and arteries. Anything that impacts the heart is referred to as cardiac disease [2]. A World Health Organization report estimates that cardiac disease is responsible for the demise of around 17 million individuals annually on a global scale [3]. A multitude of heart diseases exist, including but not limited to coronary artery disease, congenital heart disease, and arrhythmia. Manifestations of heart disease in patients include chest discomfort, vertigo, and excessive perspiration.

Diabetes, obesity, smoking, and hypertension, among others, are the leading causes of cardiovascular disease [4]. Early detection of cardiac disease is not adequately addressed by current diagnostic methods for many reasons, including execution time and accuracy. Consequently, scientists are striving to develop a methodology that effectively identifies cardiovascular diseases at an early stage. In the modern era, diagnosing and treating cardiac disease is extraordinarily challenging—the absence of technology and medical specialists [5]. Heart disease is a specific pathological state characterized by irregularities affecting the circulatory system and originating in the heart. It can range from complications with the heart rhythm to blood vessel disease [6]. Medical specialists ascertain heart disease via a comprehensive evaluation of the patient's medical history and physical symptoms, which may involve blood pressure, glucose, and cholesterol readings. Furthermore, to ascertain the presence of coronary angiography, advanced medical examinations including the utilization of X-rays, echocardiograms, electrocardiograms (ECGs), exercise stress tests, radionuclide assessments, MRI scans, and CT scans may be employed [7].

The field of machine learning is crucial for early disease detection, aiming to conclude new data by revealing hidden patterns in observations [8]. Early detection of cardiac disease has the potential to reduce mortality rates, leading researchers from diverse backgrounds to employ machine learning in examining this matter. Recent advancements in machine learning have sparked renewed interest in predicting heart disease diagnosis for individual patients [9]. Decision trees are commonly used in the early detection of heart disease, offering advantages in clinical contexts for classifying unseen data and ease of comprehension by decision-makers [10][11]. The increasing volume of newly generated data leads to a rise in effective implementations of decision tree variants in this field [12][13]. Although numerous studies in various application domains primarily concentrate on prediction through the utilization of decision trees, The sub-domain of cardiac disease has seen a paucity of graphical representations for modeling and extracting rule sets; this could aid in the making of decisions. To compensate for the aforementioned deficiencies, the purpose of this research is to model and forecast cardiac disease. A comprehensive data set comprising five data sets, comprising 1190 observations and eleven features, was utilized for this investigation. To enhance the precision and efficiency of the model's execution, the data underwent pre-processing. Furthermore, an analysis was conducted on the Classification

and Regression Tree (CART) algorithm, which was suggested and evaluated for its ability to accurately detect cardiac disease and generate a graphical representation of the model's rule inferences. In contrast to the majority of studies, the implications and outcomes of this investigation can be readily integrated into future clinical decision support systems by utilizing the provided rule sets and modeling outcomes. Ensemble methods, including gradient boosting and random forests, enhance the robustness and dependability of cancer classification models [14]. By combining diverse models, each with its advantages, ensemble methods address these shortcomings by leveraging the collective intelligence of multiple algorithms. This collaborative approach reduces the risk of overfitting and enhances the generalization ability, resulting in more resilient and robust predictions. Ensemble learning proves effective in cancer classification systems, improving overall precision and consistency. Machine learning algorithms play a crucial role in rapid analysis and decision support for cancer diagnosis and treatment when integrated into real-time systems. Capable of processing large amounts of data in real-time, these algorithms extract relevant information efficiently, providing clinicians with evidence-based insights. The timely and accurate dissemination of information empowers medical professionals to utilize this technology to enhance patient outcomes, customize treatment strategies, and make informed decisions [15].

Classification is employed to train a model, known as a classifier, by utilizing a substantial quantity of labeled data and instances. Additionally, it is utilized to classify a test example into one of the classes by employing an informed ideal termed analysis—classification-based techniques for variance discovery function in a similar fashion across two phases. Instructing a classifier to utilize the accessible characterized training statistics [16] will comprise this course segment. By employing classification—variance detection techniques for classification functions, the test segment categorizes a test instance as normal or aberrant into one or more classes. One-class-specific variance recognition performances are predicated on the premise that every training instance includes a solitary class identifier. Both approaches utilize a one-gender classification algorithm to establish a biased boundary between standard cases. A test instance is considered "anomalous" if it deviates from the examined limits during a procedure that identifies anomalies across multiple classes. The training data demonstrates the classification of instances into numerous common classes [17]. The categorization is intended to differ between each typical course and the remaining courses, including anomaly detection methods. An instance of a test is deemed abnormal if any of the classifiers fail to identify it as natural. Certain techniques within this subdivision establish a connection between trust and classification prediction. An anomaly exists when neither of the classifiers is certain that the test instance has been labeled as natural.

The clustering algorithm is a process of unsupervised learning that has numerous industrial and academic applications. Partitioned clustering is a prevalent technique among the many available [18] clustering methods due to its minimal computational demands. The process of partition clustering may be conceptualized as an optimization problem in which locating the optimal clustering center is the objective. A solitary objective function is employed by McDowell to assess the grade of clustering. The majority of these clustering methods evaluate the clustering results using a singular objective function [19]. By reading previous studies, the researchers' goals and my goal are to identify the main risk factors associated with heart disease.

Developing predictive models for early detection of heart disease. Evaluating the effectiveness of different machine learning algorithms in predicting heart disease. Evaluate the influence of lifestyle factors on the development of heart disease. Investigate potential interventions or treatments to reduce the risk of heart disease.

The remainder of the paper is arranged as follows. The most popular dataset for heart disease utilized by researchers in this area is presented in Section 2, along with background information on machine learning and classification methods. Section 3 offers a tabular comparison and a study of the literature on the projected research effort currently being done in this field.

II. BACKGROUND

This section provides concise explanations of the subjects that are pertinent to the paper, including machine learning and hybrid machine learning and its supervised and unsupervised techniques.

Machine learning (ML), which falls under the umbrella of artificial intelligence, utilizes mathematical models of varying complexity to facilitate uninstructional learning in computers. ML has gained much traction in healthcare research, primarily due to its adaptability and capacity to assist in making sense of the ever-growing quantity of electronic data. To date, however, clinically applicable applications appear to be restricted to fields that utilize imaging [20][21][22] and high-frequency time-series data [23][24]. ML models, despite their considerable strength, possess certain limitations that may impede their applicability in pharmacovigilance. One such vulnerability is their black-box structure [25][26], which currently precludes them from being compatible with certain branches of causal inference [27]. However, to be effective, they require sizable (or information-rich) datasets when applicable [28]. Machine learning utilizes algorithms to extract valuable insights from data, a rapidly expanding discipline. Various fields, including medicine, benefit from informed predictions and decisions drawn from this data. Machine learning encompasses surveillance-based, unsupervised, and other algorithms, including learning ensembles [29] used to assess dataset precision and categorization. Numerous methodologies, such as Naïve Bayes classification, decision trees, and artificial neural networks, are applied in cardiac disease diagnosis. These algorithms facilitate the creation of predictive models for effectively identifying cardiac disease prevalence from patient data, particularly valuable in medical diagnosis. Moreover, machine learning extends to forecasting other conditions like liver disease, diabetes, and tumors [30] .The term "hybrid" generally refers to a composite of two or more elements possessing characteristics that are either dissimilar or similar. Various constituents possess while they possess

different qualities; however, upon joining, the final element may possess both qualities. A hybrid strategy integrates two or more algorithms, each possessing its own unique set of advantages. When algorithms are combined, novel outcomes are generated that may be more precise and accurate than when the techniques are used in isolation [31]. It is used in many fields, including diagnosing hepatitis [32]. heart disease, and others. Fig. 1 shows the percentage of use of machine learning during the previous four years.

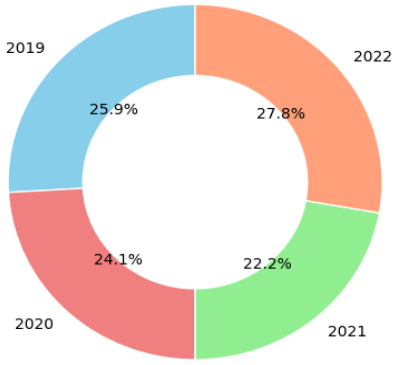

Fig. 1. Shows the percentage of use of machine learning during the previous four years

The principal ML techniques fall into the following categories:

A. Supervised Machine Learning

This method utilizes a dataset containing instances and their corresponding responses (the output). The algorithm is capable of acquiring knowledge from the dataset via a training procedure, and subsequently applying that knowledge to any novel input. The supervised learning technique is illustrated by classification and regression [32]. This section provides concise definitions of the classification techniques that are most frequently employed in the prediction of cardiac disease.

1) Support Vector Machine

Supervised learning models are employed in classification and regression analysis to examine data and identify patterns. SVM is specifically engineered to identify hyperplanes that partition data in N-dimensional regions (N-features). There are two possible interpretations regarding the linear distinction of data within a given dataset. However, if the data are not linearly separable, the linear kernel is ineffective.

When data are linearly distinct, they are challenging to separate. For classification, samples from one class are positioned on one side of a hyperplane, and samples from another class are positioned on the opposite side. The hyperplane is optimized to ensure the maximum possible separation between the two classes. [34] Support vectors consist of class data elements that are in closest proximity to the hyperplane. Fig. 2.

Example-SVM was used to predict cardiac illness in a research that was published in the Journal of Medical Internet Research. Researchers gathered information on a range of patient characteristics, including age, gender, blood pressure, cholesterol, and heart disease in the family history. Using this dataset, they trained an SVM model to categorize individuals into heart disease risk groups: high-risk and low-risk. The SVM model demonstrated a high degree of accuracy in predicting the risk of heart disease, enabling medical professionals to identify patients who could benefit from medication or lifestyle modifications as preventative measures.

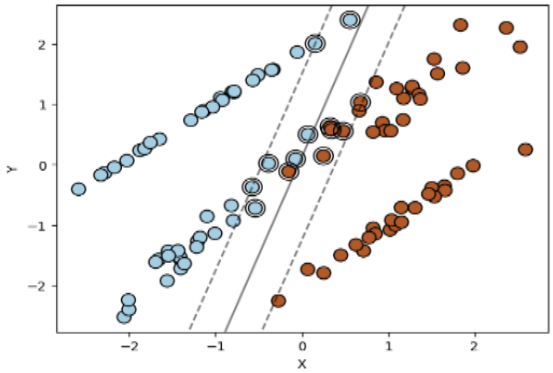

Fig. 2. Support Vector Machine

2) K-Nearest Neighbour

One of the most extensively used machine learning procedures for classification problems, pattern recognition, and regression, KNN is a straightforward model. KNN obtains neighbors from data. by utilizing the Euclidean distance between data locations [35]. This algorithm is implemented to solve classification and regression issues. When k is a user-defined constant, it will locate all existing feature cases that are similar to the new case and encircle all cases to locate the new case for a similar category. Consequently, the K value is crucial and must be selected with care, as an extremely small K may result in overfitting of the system. A few limitations exist, including subpar performance when confronted with a large training dataset.

Example-KNN was used by researchers to predict cardiac illness in a study that was published in the International Journal of Medical Informatics. They obtained information from individuals having heart evaluations, such as age, BMI, blood pressure, and ECG readings. They created a prediction model that could categorize individuals into various heart disease risk groups by using the KNN algorithm on this dataset. Promising outcomes were shown by the KNN model in correctly identifying those who are susceptible to heart-related problems, allowing for early intervention and individualized treatment plans.

3) Decision Tree

At present, decision trees (DTs) are among the most potent and extensively implemented predictive and classification techniques in machine learning. Numerous scholars have employed it as a classifier within the healthcare domain to evaluate data and arrive at conclusions. By learning, DT generates a model that anticipates the value of a target variable. By generating fundamental decision rules from data properties and segmenting data into branch-like units. Input values can be classified as either continuous or discontinuous. Probability scores or class designations are returned by the leaf nodes. The tree can be transformed into

a collection of decision principles [36]. Fig. 3 The figure illustrates the idea of a decision tree.

Example-Decision trees were used in a study by a group of cardiologists and data scientists to estimate the risk of heart disease. From electronic health records, they collected information on the patient's demographics, medical history, and results of diagnostic tests. They were able to determine important risk variables and create a prediction tool for determining each patient's chance of developing heart disease by using a decision tree model. Through focused treatments and optimized patient care pathways, the decision tree model gave doctors practical insights.

4) Gradient Boosting (GB)

An ML algorithm utilized for regression and classification is GB. It is an ensemble method in which a robust predictor is generated by combining numerous poor learners. To minimize the loss function, the algorithm operates by applying decision trees to the residuals of the current model iteratively. With each iteration, the by directing attention toward the blunders that occurred during the preceding iteration. By summing the predictions generated by each tree in the model, the final forecast is ascertained [37]. Hyperparameters, namely the number of trees, learning rate, and complexity of the trees, are modifiable to optimize the algorithm's efficacy [38].

Example-Using patient data gathered from wearable technology and health monitoring applications, a healthcare organization used gradient boosting in the real world to forecast heart disease. Features including sleep patterns, physical activity levels, and heart rate variability were included in the dataset. They developed a prediction model that might identify early indicators of heart illness and notify users to seek medical assistance by using the gradient boosting technique. By enabling people to actively track their cardiovascular health and schedule timely physician visits, the GB model improved preventive healthcare.

5) Random Forest

Random Forest is a strategy for classifying data that makes use of an extensive collection of decision trees. By combining bagging and feature randomization, an uncorrelated forest of trees is generated, where the accuracy of the committee prediction surpasses that of any individual tree frequently employed in classification and regression tasks. To attain the most favorable result, this classification methodology generates and merges numerous decision trees. For tree learning, it primarily employs bootstrap aggregation or bagging [39]. Fig. 3, The figure illustrates the idea of a Random Forest.

Example-The random forest algorithm was used in retrospective research at a cardiology clinic to predict the course of heart disease in patients with coronary artery disease. Clinical characteristics including medication history, imaging results, and cardiac biomarkers were included in the dataset. Researchers created a prognostic tool for evaluating the likelihood of adverse cardiac events, such as heart attack or stroke, by training a random forest model on this dataset. When it came to risk stratification, the random forest model performed better than other models, which helped doctors better customize treatment programs and improve patient outcomes.

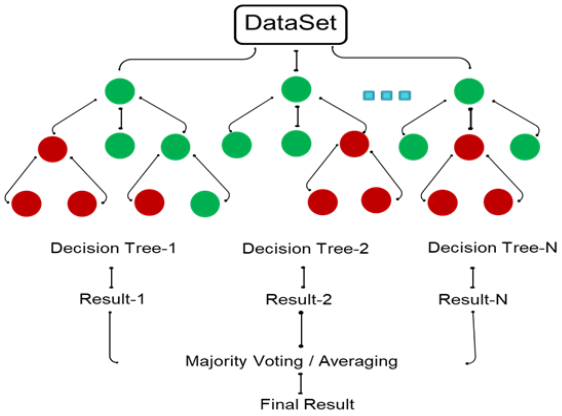

Fig. 3. Random Forest

B. Unsupervised Machine Learning

Unsupervised learning systems can identify distinctive features like retractable talons, whiskers, and long tails, revealing concealed patterns within data for forecasting. Employing principal component analysis (PCA) and cluster analysis, unsupervised ML removes associated features using covariance matrices, eigenvalues, and eigenvectors. Recognized algorithms include the Self-Organized Model (SOM), K-Mean, PCA, and LDA. The K-means clustering algorithm, incorporating preprocessing and outlier detection, initially identifies anomalies in the diabetes dataset [40]. Subsequently, the SVM classification algorithm is applied. Using K-means, patients are categorized as healthy or diabetic in a healthcare dataset for expectant women. The study achieves 78% accuracy with the K-Means algorithm [41]. Given the extensive scale of the diabetes dataset, understanding the fundamental components of diabetes detection and prognosis is crucial.

This section briefly defines unsupervised techniques used in heart disease prediction, spanning various fields such as botnet detection [42].

1) K-means

The k-means algorithm is frequently applied in data science across multiple iterations. Based on their similarity, the k-means clustering algorithm divides components into groups. Fig. 4 illustrates a graphical representation of the K-mean clustering workflow. K represents the number of groups. Consequently, three groupings will result if k equals three: [43], [44], [45]. For each data point, this clustering algorithm divides the unlabeled dataset into distinct clusters that share similar characteristics. The challenge lies in identifying cluster centroids, which are K-shaped centers. There will be a cluster centroid for each group. Once a new data point is introduced, the algorithm will employ metrics such as Euclidean distance to determine to which cluster the data point is associated. Utilizing the K-mean clustering method, which is iterated until the optimal centroid is obtained, the centroids are computed. Most likely, the number of groupings is known. It is also referred to as the flat clustering algorithm.

"K" represents the quantity of groups identified by the algorithm in K-means [46]. Implementations in Market segmentation, Grouping of documents and images, Algorithm for dynamic data trend analysis, Customer segmentation, and Image compression.

Example-Using k-means clustering, researchers were able to distinguish between different patient groupings according to their heart disease risk factors in a study that was published in the Journal of Healthcare Informatics Research. They gathered information on the demographics of the patients, their lifestyle choices, and clinical parameters including blood pressure, cholesterol, and glucose levels. They discovered common patterns of heart disease risk variables inside each cluster by utilising the k-means algorithm to cluster individuals into groups with comparable risk profiles. In the end, this data improved cardiovascular health outcomes by enabling tailored risk management plans and targeted therapies for various patient groups.

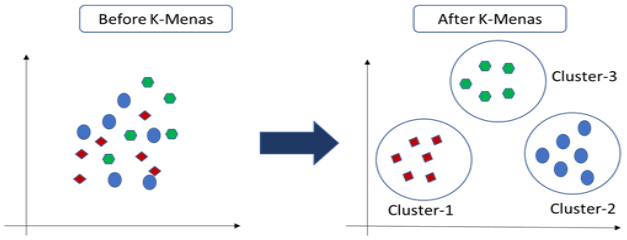

Fig. 4. Workflow of K-mean Clustering

2) Principal Components Analysis (PCA)

Principle Component Analysis (PCA) is a method for reducing the dimensionality of large data sets. It accomplishes this by converting numerous variables into a smaller group while preserving the majority of the data information present in the "large set." The depletion of accuracy Although it is inevitable to reduce the number of variables in a dataset, doing so in exchange for simplification compromises some accuracy when attempting to reduce dimensionality. machine learning algorithms find data analysis simpler and more efficient due to the absence of superfluous variables to examine and the ease of studying and visualizing smaller datasets [47]. PCA, in summary, aims to reduce the number of variables in a given dataset while preserving the maximum amount of information possible.

Example-PCA was used in a research study at a cardiac rehabilitation center to analyze multidimensional data from echocardiograms, stress tests, and blood assays, among other heart assessment tests. Numerous associated variables about heart function and health markers were included in the dataset. Latent variables, or principal components, that describe underlying patterns in heart health were found by researchers using principle component analysis (PCA) to minimize the dimensionality of the data while maintaining its essential information. These key elements were useful markers for determining patients' heart disease risk and tracking their progress via rehabilitation treatments.

3) Frequent Pattern Growth Algorithm

The Frequent Pattern (FP) Growth algorithm has been implemented as an enhancement to the Apriori algorithm. The database is represented by this algorithm as a pattern or frequent occurrence. tree structure (FT). By utilizing this regular tree, the most prevalent patterns are identified. The Apriori method requires n + 1 database scans (where n is the length of the most extended model), whereas the FP growth method only requires two scans [48]. Uses Application A Frequent pattern (FP) growth algorithm can resolve a variety of issues, including clustering, classification, software issue identification, and recommendations.

C. Reinforcement Learning

This method strikes a balance between supervised and unsupervised learning in that the performance of the model is enhanced through its interactions with the environment. Understand how to rectify its errors as a result. It should achieve the desired outcome by conducting analysis and exploring various potential solutions [36]. Example-Using the Frequent Pattern Growth (FP-Growth) algorithm, researchers analysed electronic health data in a clinical trial at a cardiology clinic to find recurrent patterns in the symptoms, test findings, and treatment outcomes of patients with heart disease. Through the use of FP-Growth to mine common patterns from the dataset, they were able to identify correlations between certain clinical variables and the course of the illness or the effectiveness of therapy. By forecasting patients' risk of suffering from unfavorable cardiac events and creating customized care plans, these insights assisted healthcare practitioners in improving patient outcomes and treatment tactics.

III. PERFORMANCE EVALUATION

The performance metrics employed in this study were pivotal in assessing the classifiers' efficacy and precision. Several metrics were utilized in the analysis, such as AUC-ROC, F1-score, Cohen's kappa (κ), recall, precision and mean squared error (MSE). The aforementioned metrics contributed significant insights regarding various facets of the classifiers' efficacy. The assessment was conducted utilizing the confusion matrix presented in Table I, which illustrates the outcomes of the classification process concerning false positives (FN), true negatives (TP), true positives (TN), and false positives (FP). TN denotes instances that were accurately forecasted as belonging to the negative class, whereas TP signifies instances that were accurately predicted as belonging to the positive class. Positive-class inaccurate predictions (FP) and negative-class incorrect predictions (FN) are denoted as FP and FN, respectively. The MSE and the errors table comprise the additional performance evaluators in the predictive models. Frequently, these relationships are employed to quantify the

TABLE I. THE CONFUSION MATRIX

	Predicted Positive	**Predict Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)
Accuracy = (TN + TP)/(TN + FP + TP + FN) ∗ 100%		
Precision =(TP/) / ((TP + FP) × 100%)		
Recall = (TP) / ((TP + FN) × 100%)		
$\kappa = 2 \times (TP \cdot TN - FP \cdot FN)/TP + FP) \cdot (FP + TN) + (TP + FN) \cdot (FN + TN)$		
F1-score = 2∗ Precision × Recall/Precision + Recall		
$MSE = \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i)^2$		

IV. PREVIOUS SURVEYS

Heart disease datasets have undergone extensive experimentation and investigation. The compilation of past research includes datasets that scientists have worked on, amalgamated based on shared attributes. The present study focuses on the analysis of this combined dataset.

In terms of global mortality, fatalities related to cardiac disease remain the most prevalent. Throughout the years, a multitude of traditional approaches have been utilized to detect cardiac disease. Conversely, traditional methodologies for evaluating risk are often implemented, such as the Hazard Ratio, Framingham Risk Score, Thrombolysis in Myocardial Infarction, Systematic Coronary Risk Evaluation, Global Registry of Acute Coronary Events, and QResearch Risk. However, these techniques possess specific limitations in comparison to approaches based on machine learning [49] [50].

These limitations include the need for additional assumptions, dealing with low-dimensional data, and increased time consumption, especially with larger datasets . Moreover, it is possible that the statistical methods mentioned above may not be adequate when it comes to analyzing unstructured or semi-structured data. Therefore, it can be argued that the emerging risk factors incorporated into machine learning methodologies offer a more comprehensive assessment than those utilized in conventional risk-scoring techniques. Notwithstanding this, it is crucial to note that machine learning is employed to uncover latent patterns and correlations among parameters in a dataset, thereby forecasting the probability of future occurrences, as opposed to statistical methods being utilized to conclude from samples [51].

A substantial body of literature exists that delineates the risk factors associated with cardiovascular disease. Emerging risk factors associated with heart-related mortality were investigated [52]. They proposed associations between heart-related fatalities and various factors such as sex, age, race, socioeconomic status, diabetes diagnosis, obesity, and diabetes.

Moreover, machine-learning approaches have suggested biomarkers like electrocardiograms (ECGs), serum markers, and imaging parameters as reliable predictors across diverse and integrated cases. In the prediction of heart disease using machine learning, a review of papers from the last three decades identified age, sex, blood pressure, hypertension, ST depression, chest pain type, and cholesterol as the primary determinants [53]. Another investigation delved into the social determinants of heart disease predicted by machine learning algorithms [54]. Their findings imply that social determinants, including gender and ethnicity, could be considered alongside medical inputs when predicting the risk of cardiac disease in patients.

This study predicts cardiac disease through the use of machine learning, utilizing six algorithms and the Cleveland and IEEE Dataport datasets. To optimize, GridsearchCV and five-fold cross-validation are utilized. AdaBoost outperformed logistic regression in IEEE Dataport, whereas the former dominated the Cleveland dataset. A flexible voting ensemble outperformed logistic regression and AdaBoost, attaining 93.44 percent (Cleveland) and 95 percent (IEEE Dataport). Innovatively, the study optimizes hyperparameters utilizing GridSearchCV in conjunction with five-fold cross-validation, evaluating the accuracy and negative log loss metrics. The ensemble method significantly outperforms previous studies on the prediction of cardiac disease. Based on ML algorithms, [58] developed an intelligent computational model for the early and accurate detection and diagnosis of cardiac disease. They applied a variety of ML algorithms to the UCI repository's Cleveland and Hungarian heart disease datasets.

Upon comparing the algorithms concerning performance evaluation metrics, it was determined that Gradient Boosting and Extra-Tree Classifier attained the highest values of overall accuracy. To develop a more effective solution, several ensembles and hybrid representations were suggested by researchers to predict cardiovascular disease. The accuracy of CVD obtained from the Mendeley Centre, Cleveland datasets, and IEEE Port was attained by the method proposed in [59]. The accuracy of the hybridized LR and RF models for predicting heart disease was demonstrated in reference [60]. These studies intend to examine the relationship between coronary artery calcium and carotid plaque in asymptomatic individuals, as well as its potential correlation with the risk of developing cardiovascular disease [61]. Presently, IoT-integrated machine-learning techniques are widely employed in the prediction and detection of diseases.

The author of [62] illustrates how ML could be utilized to resolve the issue, this was the aim. ML is employed to examine instances linked to health conditions and maladies through the analysis of numerous healthcare datasets [63]. The objective of one investigation [64] was to construct a predictive model for cardiovascular disease by employing machine learning methodologies. The information utilized for this objective was acquired from the Cleveland heart disease dataset, which was obtained from the UCI machine learning repository. The dataset comprised 303 instances and 17 attributes. The research utilized various supervised classification techniques, such as k-nearest neighbor (KKN), decision tree, random forest, and naïve Bayes. The findings revealed that the KKN model demonstrated the greatest degree of precision, measuring at 90.8%. The study highlights the potential usefulness of machine learning methods in forecasting cardiovascular disease and emphasizes the criticality of choosing suitable models and techniques to attain the most favorable outcomes. A proposed enhancement to the k-means clustering algorithm for large data was introduced in 2020 [65]. The proposed solution sought to enhance the clustering of data in terms of both efficiency and precision. In contrast, a private artificial dataset containing up to 50,000 records was utilized in the experiment to evaluate the efficacy of clustering in comparison to basic k-means. The findings demonstrated a potential 50% increase in the efficiency of clustering. In reference [66], an innovative method was proposed for improving the efficiency of k-means clustering on large data sets through the utilization of the Hadoop parallel framework.

The clustering procedure was enhanced through the division and merging of clusters based on their distance from one another. Non-clustered data points were partitioned into clusters according to their proximity. This approach is expected to increase the effectiveness of k-means clustering. Utilizing the KDD99 dataset, the efficacy of the solution utilizing a parallel Hadoop platform and shared memory space was assessed. The simulation outcomes demonstrated a 10% increase in precision. An improved iteration of the k-means clustering algorithm specifically designed for large datasets is presented in Reference [67]. By implementing this enhanced iteration, processing demands are mitigated without compromising precision levels. In the final two iterations, the method makes use of the distance between points and their variations, in addition to the cluster radius of the research index. The findings demonstrate a substantial enhancement in the efficacy of clustering, with gains of as much as 41.8% observed in the baseline scenario. A variety of real-world datasets were utilized in the experiment, such as Concrete, Abalone, Facebook, and CASP. Additionally, the solution necessitated significant computational power, which was accomplished by utilizing a 12-GB of RAM and a multi-core Core i7 processor.

To optimize the clustering of a dataset in 2021, Reference [68] evaluated the performance of the simple k-means algorithm and enhanced it using the parallel k-means technique. The solution was specifically engineered to optimize performance on multi-core devices—clustering effectiveness. The investigation employed a dataset from the education sector that underwent evaluation a maximum of ten times to determine the elapsed time and the number of clustering iterations. Up to three times speedier overall performance improvement was achieved compared to the straightforward K-means algorithm and utilized the K-means clustering method to identify anomalies in their data set. Following the elimination of anomalies through clustering, they evaluated and discussed the performance of KNN, Random Forest, and Support Vector. Various clusters are analyzed using Logistic Regression, Naïve Bayes, and Machines [69]. Table II briefly reviews Cardiovascular disease prediction studies conducted on large data sets, using supervised, unsupervised, and hybrid machine-learning technique.

TABLE II. COMPARISON OF VARIOUS EXISTING ML METHODS

Papers/year	Methods/Classifiers	Datasets	Key Findings	Challenges and limitation	Technique
[55] 2020	LR, K-NN, ANN, SVM, NB, DT, FCMIM (Feature Selection)	Cleveland Heart Disease Dataset	Best result: NB classifier, LASSO FS features.	Enhancing accuracy, reducing processing time.	Classification Techniques
[56] 2021	KNN, DT, RF	Kaggle Heart Disease Dataset	Accurate heart disease predictions using simple supervised algorithms.	Cardiovascular expertise shortage, misdiagnosis prevalence.	Classification Techniques
[57] 2021	RF, KNN, LR, NB, GB, AB, SVE classifier	UCI Kaggle Cleveland.	Ensemble models improve accuracy	Data Prep, Real-time Integration, The feature selection strategy does not give priority to selecting the most relevant features of the data set.	Classification, clustering
[58] 2020	LR, RF, XGB, SVM, DT, KNN	Cleveland and Hungarian heart disease datasets (UCI repository)	Algorithmic Performance: ETC and GB Outperform	Future research does not prioritize exploring additional improvement techniques,	Classification Techniques
[59] 2022	NB, RF, SVM	Cleveland Dataset (UCI repository, Mendeley Data	Achieved 96.75% accuracy	Improve the accuracy of existing algorithms for predicting heart disease by validating additional data sets, optimising parameters, and employing advanced optimisation techniques.	Strong Voting Classifier
[64] 2022	LR, MLP, RF	UCI repository heart disease dataset	Achieved more accuracy compared to existing work	CNN underperformed; RF, GBT, and MLP excelled among classifiers	Classification Techniques
[65] 2020	Fast k-means clustering algorithm	Using image dataset only	Accuracy and performance.	Exploration of algorithm acceleration conditions and validation	Clustering Techniques
[66] 2020	k-means clustering improvement with Hadoop	KDD99 dataset	Efficiency and execution time.	Artificial and actual datasets (AT&T, Yale, COIL-20, CMD).	Clustering Techniques
[67] 2021	k-means improvement	Multiple datasets (Facebook, CASP, etc.)	Execution time.	The proposed method's scalability and efficacy when applied to larger datasets require high-performance processing devices.	Clustering Techniques
[68] 2021	Parallel k-means clustering	Education sector dataset	Many iterations and time elapsed.	In order to address constraints and expand the functionality of K-Means clustering to encompass a wide range of data types and duties.	Clustering Techniques
[69] 2021	KNN, RF, SVM, NB, LR	UCI machine learning repository	Heart disease prediction using ML algorithms after K-means clustering	Handling anomalies through clustering	Clustering & Classification

V. CONCLUSION

In summary, this review analysis emphasizes the significance and potential of hybrid machine learning in the domain of heart disease prediction, specifically in the development of personalized risk prediction models. Incorporating diverse machine learning approaches enables a more sophisticated and precise evaluation of the risk factors linked to coronary heart disease. The ongoing development of machine learning is anticipated to bring about a significant paradigm shift in the healthcare industry regarding disease prediction and prevention through the implementation of these techniques.

In the future, healthcare professionals and researchers should work together to implement and validate the hybrid machine learning model proposed in actual clinical settings. This may entail the implementation of prospective studies to evaluate the performance of the model on a wide range of patient populations, as well as the integration of ongoing updates to augment its predictive capabilities. In essence, the effective execution of these models has the potential to render a substantial contribution to proactive healthcare administration, thereby mitigating the strain that cardiovascular disease places on both individuals and healthcare systems.

REFERENCES

[1] K. D. K. Venkatesh, M. Prathyusha, and C. H. Naveen Teja, "Identification of Disease Prediction Based on Symptoms Using Machine Learning," *JAC: A Journal Of Composition Theory*, vol. 14, no. 6, 2021.

[2] R. Kumar and P. Rani, "Comparative analysis of decision support system for heart disease," *Adv. Math. Sci. J*., vol. 9, no. 6, pp. 3349–3356, 2020.

[3] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Inf. Med. Unlocked*, vol. 16, p. 100203, 2019.

[4] K. Subhadra and B. Vikas, "Neural network based intelligent system for predicting heart disease," *Int. J. Innov. Technol. Explore. Eng*., vol. 8, no. 5, pp. 484–487, 2019.

[5] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/access.2020.3001149.

[6] R. R. Razonable *et al*., "A framework for outpatient infusion of antispike monoclonal antibodies to high-risk patients with mild-to-moderate coronavirus disease-19: the Mayo Clinic model," in *Mayo Clinic Proceedings*, vol. 96, no. 5, pp. 1250-1261, 2021.

[7] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 128, p. 102289, 2022.

[8] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Diagnosing of disease using machine learning," in *Machine learning and the internet of medical things in healthcare*, pp. 89-111, 2021.

[9] V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthc. Anal*., vol. 2, 2022.

[10] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," *Artif. Intell. Med*., vol. 128, p. 102289, 2022.

[11] A. Alanazi, "Using machine learning for healthcare challenges and opportunities," *Inform. Med. Unlocked*, vol. 30, p. 100924, 2022.

[12] L. Yahaya, N. David Oye, and E. Joshua Garba, "A comprehensive review on heart disease prediction using data mining and machine learning techniques," *Am. J. Artif. Intell*., vol. 4, no. 1, p. 20, 2020.

[13] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, and N. Kumar, "Latest trends on heart disease prediction using machine learning and image fusion," *Mater. Today Proc*., vol. 37, no. 2, pp. 3213–3218, 2020.

[14] R. M. Aziz, "Application of nature-inspired soft computing techniques for gene selection: a novel framework for cancer classification," *Soft Comput*., vol. 26, no. 22, pp. 12179–96, 2022.

[15] A. Yaqoob, R. M. Aziz, N. K. Verma, P. Lalwani, A. Makrariya, and P. Kumar, "A review on nature-inspired algorithms for cancer disease prediction and classification," *Mathematics*, vol. 11, no. 5, p. 1081, 2023.

[16] W. Zhang, Z. Zhang, H. C. Chao, and M. Guizani, "Toward Intelligent Network Optimization in Wireless Networking: An Au-toLearning Framework," *IEEE Wirel. Commun*., vol. 26, pp. 76–82, 2019.

[17] S. Rajagopal, P. P. Kundapur, and K. S. Hareesha, "A Stacking Ensemble for Network Intrusion Detection Using Heterogeneous Datasets," *Secure. Commun. Netw*., vol. 2020, p. 4586875, 2020.

[18] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," *Atmos. Pollut. Res*., vol. 11, pp. 40–56, 2020.

[19] I. C. McDowell, D. Manandhar, C. M. Vockley, A. K. Schmid, T. E. Reddy, and B. E. Engelhardt, "Clustering gene expression time series data using an infinite Gaussian process mixture model," *PLoS Comput. Biol.*, vol. 14, p. e1005896, 2018.

[20] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI," *J. Magn. Reson. Imaging*, vol. 49, no. 4, pp. 939-954, 2019.

[21] P. Prahs *et al*., "OCT-based deep learning algorithm for the evaluation of treatment indication with antivascular endothelial growth factor medications," *Graefes Arch. Clin. Exp. Ophthalmol*., vol. 256, no. 1, pp. 91-98, 2018.

[22] A. BenTaieb and G. Hamarneh, "Deep learning models for digital pathology," *arXiv preprint arXiv:1910.12329*, 2019.

[23] H. Thorsen-Meyer *et al*., "Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records," *Lancet Digital Health*, vol. 2, no. 4, pp. e179-e191, 2020.

[24] H. Thorsen-Meyer *et al*., "Discrete-time survival analysis in the critically ill: A deep learning approach using heterogeneous data," *Npj Digital Med*., vol. 5, no. 1, p. 142, 2022.

[25] F. Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

[26] S. Russell. *Human compatible: AI and the problem of control*. Penguin Uk, 2019.

[27] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Commun. ACM*, vol. 62, no. 3, pp. 54-60, 2019.

[28] R. Eshleman and R. Singh, "Leveraging graph topology and semantic context for pharmacovigilance through Twitter-streams," *BMC Bioinform*., vol. 17, no. 13, p. 335, 2016.

[29] K. Shailaja and M. T. Scholar, "Machine Learning in Healthcare : A Review," *2018 Second Int. Conf. Electron. Commun. Aerosp. Technol*., pp. 910–914, 2018.

[30] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid Machine Learning Model," *Proc. 6th Int. Conf. Inven. Comput. Technol. ICICT 2021*, pp. 1329–1333, 2021.

[31] A. Ahmed, M. Bibi, and S. Syed, "Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative Study of SVM and KNN Algorithms," *International Journal of Computations, Information and Manufacturing (IJCIM)*, vol. 3, no. 1, pp. 49-54, 2023.

[32] I. I. Ahmed, D. Y. Mohammed, and K. A. Zidan, "Diagnosis of hepatitis disease using machine learning techniques," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 3, pp. 1564−1572, 2022.

[33] R. Waigi, S. Choudhary, P. Fulzele, and G. Mishra, "Predicting the risk of heart disease using advanced machine learning approach," *Eur. J. Mol. Clin. Med*., vol. 7, pp. 1638–1645, 2020.

[34] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine

learning," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263-275, 2021.

[35] A. R. Lubis and M. Lubis, "Optimization of distance formula in K-Nearest Neighbor method," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 1, pp. 326-338, 2020.

[36] M. A. Kadhim and A. M. Radhi, "Heart disease classification using optimized Machine learning algorithms," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 2, pp. 31-42, 2023.

[37] L. V. Fulton, D. Dolezel, J. Harrop, Y. Yan, and C. P. Fulton, "Classification of Alzheimer's disease with and without imagery using gradient boosted machines and resnet-50," *Brain Sci*., vol. 9, no. 9, 2019.

[38] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)," *Diagnostics*, vol. 11, no. 9, p. 1714, 2021.

[39] H. S. Abdul Kareem and M. S. Mahdi Altaei, "Detection of Deep Fake in Face Images Based Machine Learning," *Al-Salam Journal for Engineering and Technology*, vol. 2, no. 2, pp. 1–12, 2023.

[40] N. I. Alghurair, "A Survey Study Support Vector Machines and K-MEAN Algorithms for Diabetes Dataset," *Academic Journal of Research and Scientific Publishing*, vol. 2, pp. 14-25, 2020.

[41] D. C. Sujatha, D. M. Kumar, and M. C. Peter, "Building a predictive model for diabetics data using K Means Algorithm," *International Journal of Management, IT and Engineering*, vol. 8, pp. 58-65, 2018.

[42] D. H. Mustafa and I. M. Husien, "Adaptive DBSCAN with Grey Wolf Optimizer for Botnet Detection," *International Journal of Intelligent Engineering & Systems*, vol. 16, no. 4, 2023.

[43] S. Park, Y. Hwang, and B.-J. Yang, "Unsupervised learning of topological phase diagram using topological data analysis," *Physical Review B*, vol. 105, no. 19, p. 195115, 2022.

[44] P. An, Z. Wang, and C. Zhang, "Ensemble unsupervised autoencoders and Gaussian mixture model for cyberattack detection," *Information Processing & Management*, vol. 59, no. 2, p. 102844, 2022.

[45] M. Rashid, H. Singh, and V. Goyal, "Cloud storage privacy in health care systems based on IP and geo-location validation using mean clustering technique," *International Journal of E-Health and Medical Communications (IJEHMC)*, vol. 10, no. 4, pp. 54–65, 2019.

[46] M. Mateen, J. Wen, S. Song, and Z. Huang, "Fundus image classification using vgg-19 architecture with PCA and SVD," *Symmetry*, vol. 11, no. 1, 2018.

[47] C.-H. Chee, J. Jaafar, I. A. Aziz, M. H. Hasan, and W. Yeoh, "Algorithms for frequent itemset mining: a literature review," *Artificial Intelligence Review*, vol. 52, no. 4, pp. 2603–2621, 2019.

[48] L. Gubu *et al*., "Robust mean–variance portfolio selection using cluster analysis: A comparison between Kamila and weighted k-mean clustering," *Asian Economic and Financial Review*, vol. 10, no. 10, pp. 1169–1186, 2020.

[49] A. S. Mohd Faizal, T. M. Thevarajah, S. M. Khor, and S. W. Chang, "A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach," *Comput. Methods Programs Biomed.*, vol. 207, 2021.

[50] N. Biswas, K. Mohammad, M. Uddin, and S. Tasmin, "Healthcare analytics a comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Healthc. Anal*., vol. 2, 2022.

[51] D. Mpanya, T. Celik, E. Klug, and H. Ntsinjana, "Machine learning and statistical methods for predicting mortality in heart failure," *Heart Fail. Rev*., vol. 26, no. 3, pp. 545–552, 2021.

[52] A. C. T. Ha, B. S. Doumouras, C. Wang, J. Tranmer, and D. S. Lee, "Prediction of sudden cardiac arrest in the general population: Review of traditional and emerging risk factors," *Can. J. Cardiol*., vol. 38, no. 4, pp. 465–478, 2022.

[53] R. Alizadehsani *et al*., "Coronary artery disease detection using artificial intelligence techniques: A survey of trends, geographical differences and diagnostic features 1991–2020," *Comput. Biol. Med*., vol. 128, 2021.

[54] Y. Zhao, E. P. Wood, N. Mirin, S. H. Cook, and R. Chunara, "Social determinants in machine learning cardiovascular disease prediction models: A systematic review," *Am. J. Prev. Med*., vol. 61, no. 4, pp. 596–605, 2021.

[55] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in E-healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020.

[56] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput. Biol. Med*., vol. 136, 2021.

[57] G. N. Ahamad *et al*., "Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease," *Processes*, vol. 11, no. 3, p. 734, 2023.

[58] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Sci. Rep*., vol. 10, no. 1, 2020.

[59] B. P. Doppala, D. Bhattacharyya, M. Janarthanan, and N. Baik, "A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using Ensemble Techniques," *J. Healthc. Eng*., vol. 2022, 2022.

[60] A. Kondababu, V. Siddhartha, B. B. Kumar, and B. Penumutchi, "WITHDRAWN: A comparative study on machine learning based heart disease prediction," *Mater. Today Proc*., 2021.

[61] E. F. Gudmundsson *et al*., "Carotid plaque is strongly associated with coronary artery calcium and predicts incident coronary heart disease in a population-based cohort," *Atherosclerosis*, vol. 346, pp. 117–123, 2022.

[62] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Comput. Intell. Neurosci*., vol. 2021, 2021.

[63] G. Gupta, U. Adarsh, N. S. Reddy, and B. A. Rao, "Comparison of various machine learning approaches used in heart ailments prediction," in *Journal of Physics: Conference Series*, vol. 2161, p. 012010, 2022.

[64] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci*., vol. 1, p. 345, 2020.

[65] T. Xie, R. Liu, and Z. Wei, "Improvement of the Fast Clustering Algorithm Improved by-Means in the Big Data," *Appl. Math. Nonlinear Sci.*, vol. 5, pp. 1–10, 2020.

[66] W. Lu, "Improved K-means clustering algorithm for big data mining under Hadoop parallel framework," *J. Grid Comput*., vol. 18, pp. 239–250, 2020.

[67] F. Moodi and H. Saadatfar, "An improved K-means algorithm for big data," *IET Softw*., vol. 16, pp. 48–59, 2021.

[68] R. Shang, B. Ara, I. Zada, S. Nazir, Z. Ullah, and S. U. Khan, "Analysis of simple K-mean and parallel K-mean clustering for software products and organizational performance using education sector dataset," *Sci. Program*., vol. 2021, 2021.

[69] R. C. Ripan *et al*., "A data-driven heart disease prediction model through Kmeans clustering-based anomaly detection," *SN Comput. Sci.*, vol. 2, no. 2, pp. 1–12, 2021.