IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 May 2023, accepted 27 May 2023, date of publication 30 May 2023, date of current version 9 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3281484

RESEARCH ARTICLE

Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning

AZAM MEHMOOD QADRI[1], **ALI RAZA**[1], **KASHIF MUNIR**[2], **AND MUBARAK S. ALMUTAIRI**[3]

[1]Institute of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan
[2]Institute of Information Technology, Khawaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan
[3]College of Computer Science and Engineering, University of Hafr Al Batin, Hafr Al Batin 31991, Saudi Arabia

Corresponding authors: Ali Raza (ali.raza.scholarly@gmail.com) and Kashif Munir (kashif.munir@kfueit.edu.pk)

This work was supported by the University of Hafr Al Batin, Saudi Arabia.

ABSTRACT Heart failure is a chronic disease affecting millions worldwide. An efficient machine learning-based technique is needed to predict heart failure health status early and take necessary actions to overcome this worldwide issue. While medication is the primary treatment, exercise is increasingly recognized as an effective adjunct therapy in managing heart failure. In this study, we developed an approach to enhance heart failure detection based on patient health parameter data involving machine learning. Our study helps improve heart failure detection at its early stages to save patients' lives. We employed nine machine learning-based algorithms for comparison and proposed a novel Principal Component Heart Failure (PCHF) feature engineering technique to select the most prominent features to enhance performance. We optimized the proposed PCHF mechanism by creating a new feature set as an innovation to achieve the highest accuracy scores. The newly created dataset is based on the eight best-fit features. We conducted extensive experiments to assess the efficiency of several algorithms. The proposed decision tree method outperformed the applied machine learning models and other state-of-the-art studies, achieving a high accuracy score of 100%, which is admirable. All applied methods were successfully validated using the cross-validation technique. Our proposed research study has significant scientific contributions to the medical community.

INDEX TERMS Machine learning, heart failure, cross validations, feature engineering.

I. INTRODUCTION

HHeart failure is a condition in which the heart is unable to pump enough blood to meet the body's needs [1]. Cardiovascular diseases have emerged as a significant global health concern, substantially impacting public health worldwide. Heart failure is a common and serious condition affecting millions worldwide. According to a recent state, heart failure disorders cause to happen around 26 million population [2]. The causes of heart failure can be divided into two categories. First related to the heart's structure, such as a previous heart attack. Second related to the heart's function, such as high blood pressure. Symptoms of heart failure can include shortness of breath, fatigue, and swelling in the legs and ankles. Treatment options for heart failure include medications, lifestyle changes, and in some cases, surgery. Research has shown that early detection and management of heart failure can improve quality of life and prolong survival [3]. The current study focuses on developing a machine-learning model for managing heart failure to improve patient health.

Machine learning is highly involved in medical diagnoses and the healthcare industry [4]. Machine learning has many applications in the medical field, including drug discovery, medical imaging diagnosis, outbreak prediction, and heart failure prediction. Machine learning techniques can learn patterns from large medical data and perform predictive analysis. Machine learning has many advantages compared to classical medical methods, such as saving time and costs, which helps improve diagnosis. Our prominent research contributions for heart failure detection using machine learning are as follows:

The associate editor coordinating the review of this manuscript and approving it for publication was Sotirios Goudos.

TABLE 1. The performance accuracy achieved in previous studies for heart failure prediction.

Ref.	Year	Technique	Type	Accuracy (%)
[5]	2022	Random Forest	Machine Learning	96.28
[6]	2020	Random Forest	Machine Learning	95.60
[7]	2021	Random Forest	Machine Learning	86.60
[8]	2021	Random Forest	Machine Learning	88.52
[9]	2021	VAE-Two- DNN	Deep Learning	89.2
[10]	2021	DL	Deep learning	94.2
[11]	2019	DT	Machine Learning	93.19
[12]	2021	ROT	Machine Learning	91.2
[13]	2022	SVM	Machine Learning	96.72
[14]	2022	Logistic Regression	Machine Learning	85.25

- A novel PCHF feature engineering technique is proposed to select the most prominent features to enhance performance. Eight dataset features with high-importance values are selected to develop the machine learning methods using the proposed PCHF technique. We optimized the proposed PCHF mechanism by creating a new feature set as an innovation to achieve the highest accuracy scores compared to past proposed techniques.
- The nine advanced models of machine learning are used in the comparison to predict heart failure. The hyper-parameters tuning of each applied machine learning method is conducted to determine the best-fit parameters, achieving a high-performance accuracy score. To validate the performance of applied machine learning models, we have used the k-fold cross-validation technique.

This article is further divided into subsections: Section II is based on a study of the associated heart failure literature. The research working flow is analyzed in Section III. Section IV examines the machine learning methods to predict heart failure. Section V discusses and validates our study technique scientifically with experimental results. Section VI covers the research article's conclusion.

II. RELATED WORK

This section reviews the literature relevant to our proposed research study. The studies previously used to predict heart failure are analyzed. The related research results and proposed methods are discussed comparatively.

Heart disease is considered the most dangerous and deadly human disease according to the states discussed in previous studies. The increasing incidence of fatal cardiovascular diseases is a significant threat and burden to healthcare systems worldwide [15], [16]. Children are mostly affected by this critical disease [17].

This study [18] discusses the relevance of categorization models and describes the characteristics of models that have previously been applied in healthcare. The study highlights that several investigation groups have successfully tested data mining methods in clinical applications. The researchers compared the performance of several functional classifiers using two apparatuses, WEKA and MATLAB. Generally, the precision of the decision tree, logistic regression, SVM, and other algorithms reached 52% to 67.7%, which is relatively low [19].

Previous research [11] improved the accuracy from 87.27% to 93.13%, which is good but not optimal, as shown in Table 1. Past studies detect heart failure in patients using methods such as SVM, random forest, decision tree, logistic regression, and naïve bayes classifier. After comparing the results, the decision tree achieved an accuracy of 93.19%, which is good detection of heart failure in a specific dataset.

The study [20] used Cleveland data and created an ensemble model for heart disease detection. The ensemble models were built using random forest, gradient boosting, and extreme gradient boosting classifiers, achieving an accuracy of 85.71% [7]. The Cleveland data was used in the proposed study to improve the heart disease prediction by feature selection technique which helps to achieve an accuracy of 86.60%. Finally, previous studies have found significant research gaps, suggesting that the performance accuracy is not up to mark. Consequently, we thoroughly evaluate the previous study's performance analysis in this part. This related work section is based on findings summarizing the efficiency of all previously applied models. According to previous studies, different types of models still provide different prediction scores. Thus, dimensionality reduction and feature engineering can enhance the data selection, causing greater prediction accuracy [21].

We have improved our proposed study's accuracy score compared to the previous research performance score. The precise credentials and findings of heart failure are necessary for proper treatment. We used advanced machine learning techniques in this study to achieve this goal.

III. RESEARCH METHODOLOGY

In this study, we have access heart failure dataset from the repository Kaggle. The dataset contains 1025 patient records relate to heart failure and healthy patients. The data preprocessing techniques are applied to format the dataset. The exploratory heart failure data analysis is applied to understand better the data patterns and variables contributing to heart failure. In feature engineering, high-importance features are selected using the proposed PCHF technique. Then the dataset is split into two portions, train and test. The nine advanced machine-learning techniques are applied to the dataset portions. The hyperparameter-based fine tuning is applied to the machine learning models. The outperformed proposed model aims to forecast heart failure with high efficiency. Figure 1 examines the research methodology working flow.

Proposed methodology working flow steps:

- **Step 1:** The heart disease-related dataset is imported and preprocessed to remove unnecessary noise.
- **Step 2:** The exploratory data analysis is applied to understand better the heart failure data patterns.

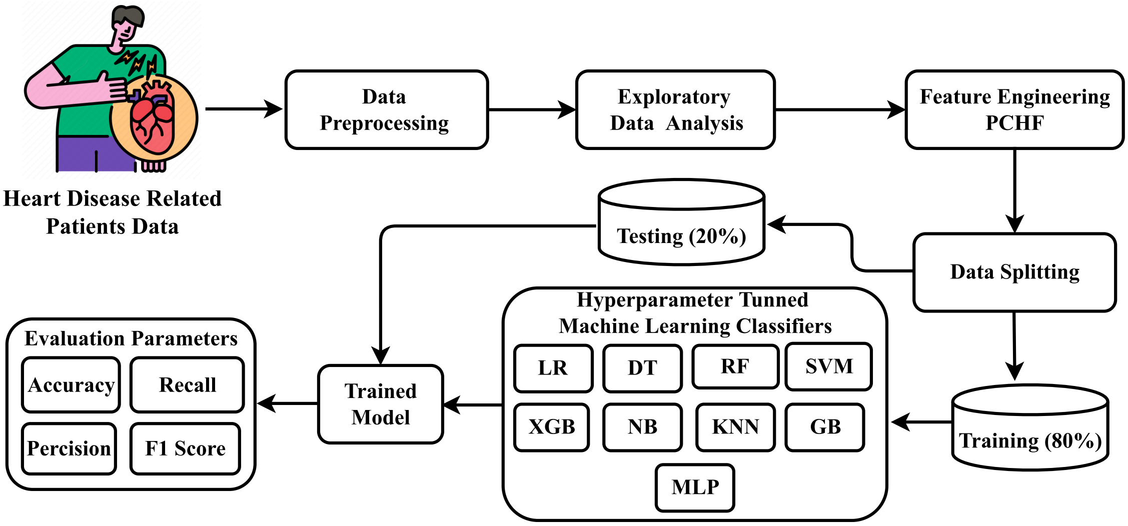

FIGURE 1. The proposed study methodology analysis for heart failure prediction.

TABLE 2. The study heart-related dataset features analysis.

Sr no.	Features	Discrete Values	Non-Null Count	Data Type
1	The age feature describes the Age of patients. The Age limit is from Min 29 to 75 Max.	Values between 29 to 75	1025	int64
2	Sex feature describes gender. 0 for Females, 1 for Males	0,1	1025	int64
3	CP feature defines chest pain. It has four different values that indicate the patient's condition according to value.	0,1,2,3	1025	int64
4	tRestBP feature describes the patient blood pressure. Ranges Min 94 to 200 Max.	values from 94 to 200	1025	int64
5	Chol feature describes the cholesterol level of patients. The Min Chol value is 126, and the Max Chol value is 564.	values between 126 to 564	1025	int64
6	FBS feature describing the fasting blood sugar of the patient. Values depend on whether the patient has more than 120 g/dl sugar=1 or less than=0.	0,1	1025	int64
7	RestECG feature shows the result of ECG from 0 to 2. Each value describes the condition of the heart pulse.	0,1,2	1025	int64
8	The thalach feature indicates the patient's maximum value count at the hospital's admission time. The Min value is 71, and the Max value is 202.	values between 71 to 202	1025	int64
9	Exang feature indicates whether the exercise encourages the Angina or not. If yes =1, Not=0.	0,1	1025	int64
10	The old Peak attribute is used to describe the depression condition of patients by assigning different values. 0 to 6.2.	values between 0 to 6.2	1025	float64
11	The slope attribute describes the patient's condition during peak exercise. These values are defined into sections [Upsloping, Flat, Down Sloping].	1,2,3	1025	int64
12	CA feature of the dataset shows the fluoroscopy status, which shows the vessels' color.	0,1,2,3,4	1025	int64
13	Thal feature is the kind of test called thallium, which is required to check when a patient has chest pain or breathing issue. Different values indicate the condition of the Thallium test.	0,1,2,3	1025	int64
14	Target is the final attribute called the label Column. This attribute describes the classes. The dataset has two classes: "0" means there is no chance of heart failure, and "1" means a strong chance of heart failure.	0,1	1025	int64

- **Step 3:** A novel feature engineering PCHF is proposed to select high-importance features which result in high-performance scores.
- **Step 4:** The heart disease-related dataset is split for training and testing.
- **Step 5:** The performance of applied techniques is evaluated. The outperformed method is used for heart disease detection.

A. DATA OVERVIEW AND PREPROCESSING

The study dataset based on heart disease-related features is used for building the applied models in this study. The dataset is publicly available online and collected from the famous repository Kaggle [22]. The study dataset has 14 features initially. Table 2 summarizes the feature details of the dataset. The dataset contains 1025 patient records of heart disease based on 713 men and 312 women samples. We have checked

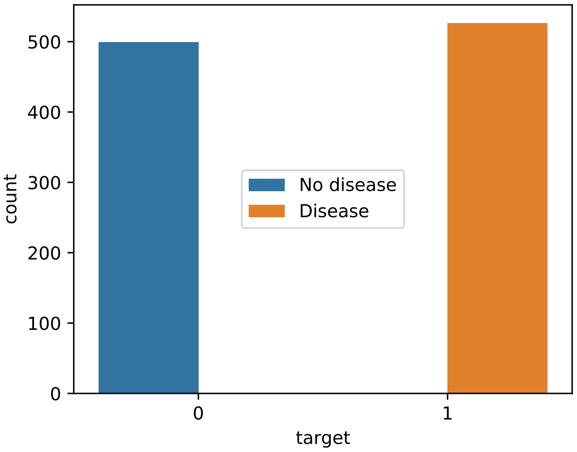

FIGURE 2. The dataset distribution analysis is based on the target column.

the dataset datatypes, which show 13 attributes have int type, and one attribute slope have float datatype. We have studied the dataset for null values, which results in the performance. The analysis shows that the dataset contains not any null values.

B. EXPLORATORY DATA ANALYSIS

The exploratory data analysis is applied to our study dataset to determine valuable statics. The exploratory data analysis is based on charts and heatmap graphs containing dataset-related patterns.

The bar chart-based dataset distribution analysis using the target feature is visualized in Figure 2. The analysis shows that the dataset contains 499 healthy patients and 526 patients with heart failure disease. In addition, 300 males and 226 females are found with heart failure disease in the dataset. This analysis demonstrates that the dataset is almost balanced to build the machine learning models.

The correlation analysis of heart failure dataset features is analyzed in Figure 3. The analysis shows that almost all features have a strong association. However, only a few have high negative correlation values. The features exang, oldpeak, ca, and thal have a high negative correlation association. The analysis concludes that the remaining dataset features have a strong association.

C. NOVEL PCHF FEATURE SELECTION TECHNIQUE

The proposed PCHF feature selection approach is analyzed in this section. The selection of highly important dataset features using the proposed PCHF technique is visualized in Figure 4. Originally the dataset contained 14 features in total. We optimized the proposed PCHF mechanism by creating a new feature set as an innovation to achieve the highest accuracy scores. The newly created dataset is based on the eight best-fit features. The PCHF technique selects the unique eight dataset features having the maximum variance compared to the original features. The proposed PCHF feature selection approach used a linear transformation mechanism. The newly created features using the PCHF method achieved high-performance accuracy for heart failure prediction in this study.

D. DATA SPLITTING

The data splitting is utilized to avoid the model fitting problem and to estimate the trained model on the unseen test data in real-time. The heart failure dataset is split into training and testing phases for machine learning classifiers. The data is split into 80% for training and 20% for testing to get results on unseen data by employed classifiers. The research classifiers we used are well-trained and conventional with excellent accuracy.

IV. APPLIED MACHINE LEARNING TECHNIQUES

Several machine-learning techniques used for heart failure prediction are studied in this section. The operational principles and basic terminology of machine learning models are explained. Our proposed research evaluates ten advanced machine-learning models for predicting heart failure.

A. LOGISTIC REGRESSION

Logistic regression (LR) is a commonly used supervised machine-learning technique that can be used for both regression and classification problems. The logistic regression method employs probability to forecast how categorical data will be labelled [23]. The process of LR for learning and making predictions is based on binary classification probability measurements. The class variables for logistic regression models must be binary classified. Similar to the target column in our study dataset with two different binary numbers. The zero is used for patients with no chance of heart failure, and the one is used for predicted heart patients in the dataset.

$$\begin{aligned} P(y=1|x) &= \frac{1}{1+e^{-z}} \quad \text{where} \\ z &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \end{aligned} \tag{1}$$

where $P(y = 1|x)$ represents the probability of a binary outcome variable y taking the value 1, given the predictor variable(s) x. The function e^{-z} is the logistic function, which maps any real value z to the range [0,1].

B. DECISION TREE

The most common supervised approach to solving classification tasks is a Decision Tree (DT). The tree-like structures are made in the DT technique [24]. The DT often includes multiple levels of nodes during tree construction. The root or parent nodes are at the top level, and the others are called child nodes. Large medical data are commonly handled via decision trees because they are easy to utilize. The data is organized as a tree, with internal nodes representing inside dataset attributes, branches for decision-making processes, and leaf nodes representing target results. The DT data is

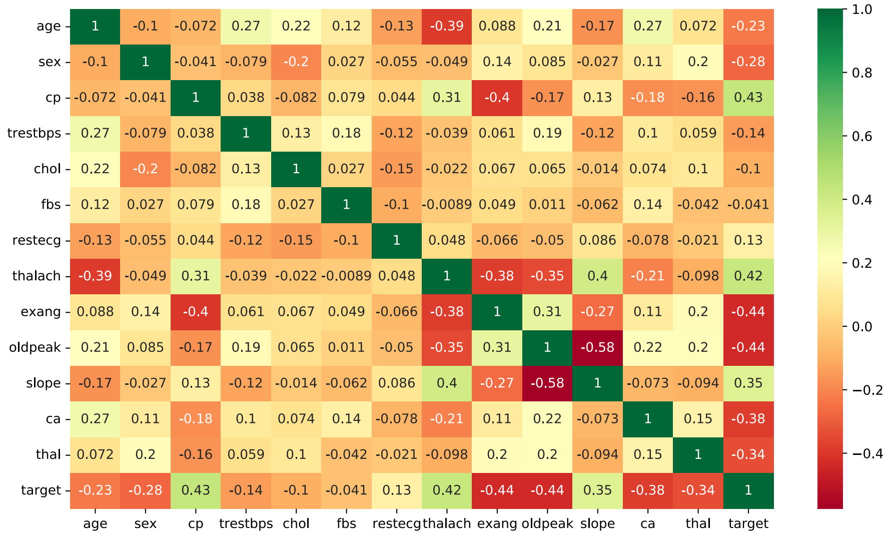

FIGURE 3. The heatmap-based correlation analysis of dataset features.

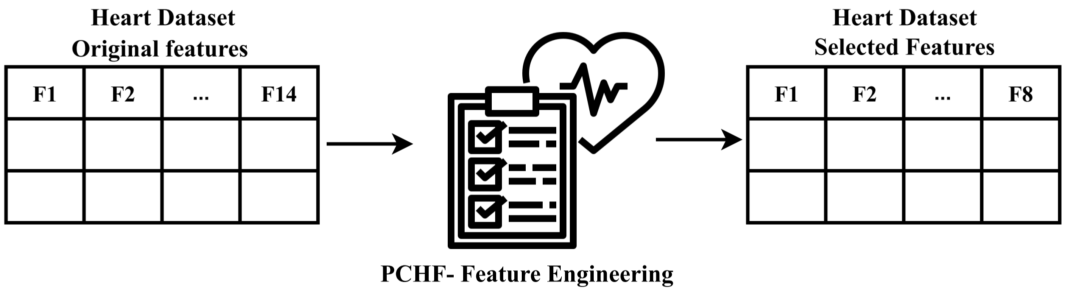

FIGURE 4. The selection of attributes from the dataset by using the PCHF technique.

divided between the nodes using the Gini index and entropy functions.

$$f(\mathbf{x}) = \sum_{m=1}^{M} c_m \mathbf{1}(\mathbf{x} \in R_m), \tag{2}$$

where $f(\mathbf{x})$ is the predicted output for input $\mathbf{x}$, M is the number of leaf nodes in the tree, R_m is the region of input space corresponding to the m-th leaf node, c_m is the prediction value associated with the m-th leaf node.

C. RANDOM FOREST

Random Forest (RF) is another supervised machine learning method commonly used to solve classification and regression problems [25]. The RF method combines several decision trees to address challenging issues and boost efficiency. An RF classifier averages many decision trees output on various subsets of a given dataset to increase the predictive accuracy. The maximum number of trees in RF gives the highest accuracy. The RF prevents overfitting and leads toward high performance.

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \tag{3}$$

where $\hat{y}$ represents the predicted output for a given input vector x. The prediction is made by taking the average of the outputs of T decision trees, denoted by $f_t(x)$.

D. SUPPORT VECTOR MACHINE

The Support Vector Machine (SVM) [26] is a well-liked supervised learning technique [27] that can be utilized to overcome classification and regression problems. Making proper decision thresholds is the aim of SVM. The SVM divides the n-dimensional space into classes using this ideal decision boundary known as a hyperplane. The hyperplane makes SVM simple to assign a new data point to the appropriate category. The hyperplane is created by SVM by choosing extreme support vectors. This technique is known as a support vector machine due to the support vectors.

$$\vec{w} \cdot \vec{x} + b = 0 \tag{4}$$

where $\vec{w}$ is the weight vector, $\vec{x}$ is the input vector, and b is the bias term.

E. EXTREME GRADIENT BOOSTING

Extreme Gradient Boosting (XGB) is a supervised ensemble machine learning model used for classification and regression analysis [28]. The ensemble learning algorithms combine multiple machine learning algorithms for a better outcome. The XGB technique combines numerous decision trees. A pair of shallow decision trees are iteratively trained by XGB, which fits the next model with each iteration, utilizing the prior model's error residuals. The final prediction output is the weighted average of each prediction in the tree. The XGB method reduces underfitting and boosts bias. The loss score in XGB is determined using the gradient descent algorithm.

$$\hat{y}i = \sum k = 1^K f_k(x_i), \quad f_k \in \mathcal{F}$$

where $\hat{y}_i$ represents the predicted value for the i-th instance, f_k represents the k-th weak learner added to the ensemble.

F. NAIVE BAYES

Naive Bayes (NB) is a supervised machine learning model to solve classification problems. The NB method is a very straightforward and effective technique capable of making accurate predictions. The NB technique is based on probability, which means the classifier makes predictions based on the likelihood of dataset variables [29]. To predict the target class for each record, NB uses the values for the independent variables. The NB model is commonly used in medical research.

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \tag{5}$$

where $P(C_k|X)$ represents the probability of a sample belonging to class C_k given the input features X. $P(X|C_k)$ is the likelihood of observing the input features X.

G. K-NEAREST NEIGHBORS

K-Nearest Neighbors (KNN) is a supervised machine learning technique primarily used for classification and regression tasks [6]. The KNN algorithm is non-parametric, which means it doesn't make any assumptions about the underlying data. The KNN algorithm places the new instance into a category comparable to the available classes, assuming that the new and available cases are similar. Most sample information is retrieved using the Euclidean distance metric in KNN.

$$y_q = \text{mode} y_{i_1}, y_{i_2}, \ldots, y_{i_k}$$

where y_q represents the predicted label for a given query point, and $i_1, i_2, \ldots, i_k$ represent the indices of the k nearest neighbors of the query point.

H. GRADIENT BOOST

One of the most often used forward-learning ensemble [30] techniques in machine learning is gradient boosting (GB). The GB is an effective method for creating forecasting analytics for situations involving regression and classification. By combining the predictions from different learner models, we may create a final predictive model that makes the right prediction. The GB approach aims to build a strong best model from several weak models. GB works on building models one at a time by training each base classifier.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

where $F_m(x)$ is the predicted output of the model at iteration m for the input x, $F_{m-1}(x)$ is the predicted output of the model at iteration $m - 1$.

I. MULTILAYER PERCEPTRON

The multilayer perceptron (MLP) [31] is a feed-forward artificial neural network that generates a set of outputs from a group of inputs. The MLP comprises several layers, the most important of which are the input layer, hidden layer, and output layer [32]. The input layer handles the input data, the hidden layer processes the data in the network, and the output layer handles the outcomes. The back-propagation algorithm is a common learning algorithm for MLP networks.

$$\boldsymbol{y} = \sigma(\boldsymbol{W}_2 \sigma(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) \tag{6}$$

where $\boldsymbol{x}$ is the input vector, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are weight matrices, $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are bias vectors, and σ is a non-linear activation function.

J. HYPERPARAMETER TUNING

The best-fit hyperparameters are selected for applied machine learning methods through an adaptive training and testing procedure [33]. The final hyperparameters are chosen on which machine learning models make the correct predictions. Table 3 reviews the selected hyperparameters for our research models. The analysis results show the parameters that we used to obtain the high performance proved beneficial for the machine learning models we used in our research.

V. RESULTS AND DISCUSSIONS

This section discusses our proposed research results and scientific validity. The machine algorithms are developed using the python programming language-based skit-learn library module. Our study performance measures are the runtime computation, accuracy, precision, recall, and f1 scores. The performance indicators of our research models are evaluated for scientific results validation.

A. PERFORMANCE RESULTS WITHOUT USING THE PCHF TECHNIQUE

The performance comparative analysis of applied machine learning models with original dataset features is analyzed in Table 4. The analysis demonstrates that the applied machine learning model achieved poor performance scores with the original dataset features. The machine learning techniques MLP, SVM, and DT achieved low-performance scores for all

TABLE 3. The optimal hyperparameters analysis for applied machine learning models is analyzed.

Models	Hyperparameters
LR	penalty=’l2’, fit intercept=True, random state=1, max iter= 100,
DT	criterion=’gini’, max_depth=300, min_samples_split=2, max_features=None, random_state=0, max_leaf_nodes=None, alpha=0.0
RF	n_estimators=300, criterion=’gini’, max_depth=300, min_samples_split=2, max_features=’sqrt’, bootstrap=True, random state=0, max_samples=None
SVM	C=1.0, kernel=’rbf’, degree=3, gamma=’scale’, probability=False, tol=0.001, cache_size=200, max_iter=-1, random_state=0
KNN	n_neighbors=3, weights=’uniform’, algorithm=’auto’, leaf_size=30, p=2, metric=’minkowski’
MLP	hidden_layer_sizes=(5, 2), activation=’relu’, solver=’lbfgs’, alpha=0.0001, learning_rate=’constant’, max_iter=300, shuffle=True, random_state=1
NB	var_smoothing=1e-09
XGB	loss=’log_loss’, learning_rate=0.1, n_estimators=100, min_samples_split=2, min_samples_leaf=1, max_depth=3, use_label_encoder=False, eval_metric=’mlogloss’
GB	loss=’log_loss’, learning_rate=1.0, n_estimators=20, subsample=1.0, criterion=’friedman_mse’,, max_depth=2,random_state=1, max_depth=1

TABLE 4. The results comparison analysis of the applied machine learning models on test data without using the PCHF technique.

Models	Runtime Computation (Seconds)	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
LR	0.062	86	87	86	86
RF	0.024	86	86	86	86
SVM	0.075	75	75	74	74
DT	0.013	75	81	75	73
XGB	0.080	92	92	92	92
NB	0.005	85	86	85	85
KNN	0.012	91	91	91	91
MLP	0.050	52	27	52	36
GB	0.010	90	90	90	90

performance metrics measures. However, the results of other applied models are acceptable, not the highest. In conclusion, the original dataset features low-performance scores for predicting heart disease.

The bar chart-based accuracy performance comparative analysis of applied machine learning models is visualized in Figure 5. The analysis demonstrates that the XGB classifier achieves the highest performance accuracy metric score of 92%. The MLP achieves the lowest accuracy score of 52%. The low-performance results are achieved by the SVM and DT models. In conclusion, the accuracy performance is not the highest in comparisons for heart disease prediction.

B. PERFORMANCE RESULTS USING THE PCHF TECHNIQUE

The results comparison analysis is performed for applied machine learning methods based on selected features using the PCHF technique. The performance metrics comparative analysis is performed in Table 5. The analysis shows that DT and RF models achieved 100% accuracy, precision, recall, and f1 scores using the PCHF technique. However, the runtime computation analysis shows that the DT model achieved 0.005 seconds which is the minimum compared to the RF and all other methods. The LR, SVM, NB, and MLP methods achieved poor accuracy performance in comparison. The analysis concludes that the DT model is the outperformed approach due to high performed accuracy with minimum runtime computations. Based on the performance metrics comparison scores, the DT is our proposed technique for predicting heart failure in real-time. The proposed PCHF technique proved very beneficial in enhancing the performance of applied methods in this study.

The bar chart-based accuracy performance comparative analysis of applied machine learning models with the proposed feature selection technique is visualized in Figure 6. The performance analysis demonstrates that RF and DT achieved 100% accuracy in comparison. In conclusion, the performance results are significantly increased using the proposed feature engineering technique.

C. K-FOLD CROSS-VALIDATION COMPARATIVE ANALYSIS

The K-fold cross-validation based on ten folds of data is applied to the machine learning models for validating the overfitting problem. Table 6 analyzes the comparative results of the K-fold cross-validation results. The analysis demonstrates that our proposed DT technique achieved a 100% accuracy score during the cross-validation. The SVM and MLP models achieve poor accuracy scores for cross-validation. Each model’s standard deviation score is also analyzed during the K-fold cross-validation. The analysis concludes that our proposed model is generalized to predict heart failure in real-time.

The comparative performance analysis of applied machine learning models during the testing and cross-validation is visualized in Figure 7. The research shows that the LR, SVM, and MLP achieved poor performance scores. The MLP

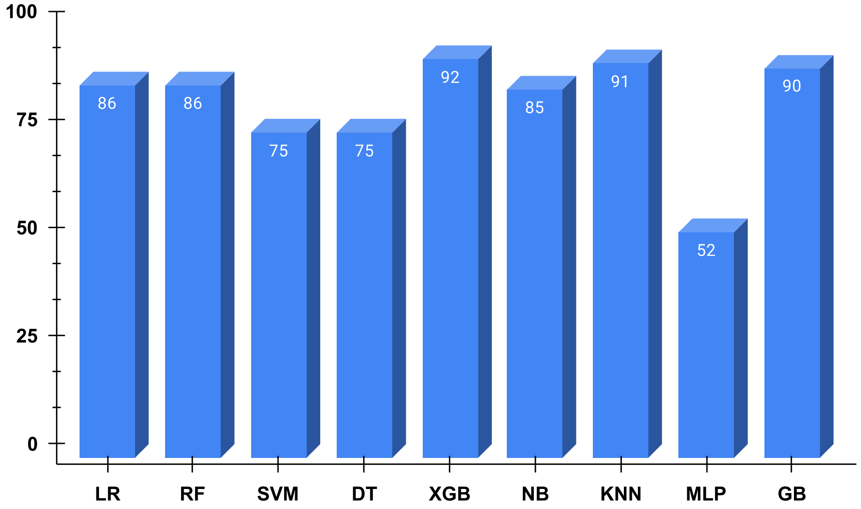

FIGURE 5. The bar chart-based results comparison analysis of the applied machine learning models without using the PCHF technique.

TABLE 5. The results comparison analysis of the applied machine learning models on test data using the PCHF technique.

Models	Runtime Computation (Seconds)	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
LR	0.044	87	87	87	87
RF	1.331	100	100	100	100
SVM	0.069	77	77	77	77
DT	0.005	100	100	100	100
XGB	0.119	98	98	98	98
NB	0.002	82	82	82	82
KNN	0.005	91	91	91	91
MLP	0.385	81	81	81	81
GB	0.046	97	97	97	97

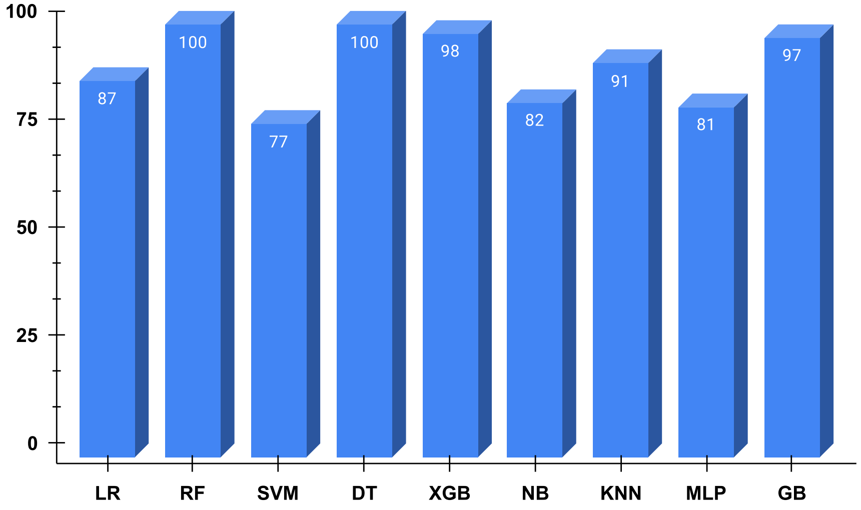

FIGURE 6. The bar chart-based results comparison analysis of the applied machine learning models using the PCHF technique.

model achieved an 81% accuracy score. However, the MLP gained a 54% cross-validation accuracy score. The analysis demonstrates that our proposed DT model is generalized and superior. This comparative analysis validates all applied models' performance for testing and cross-validation.

D. COMPARISON WITH STATE-OF-THE-ART STUDIES

The performance comparisons of the past proposed studies on our dataset are analyzed in Table 7. The parameters used for comparison are year, approach type, proposed method, accuracy, precision, and recall. The analysis demonstrates that the

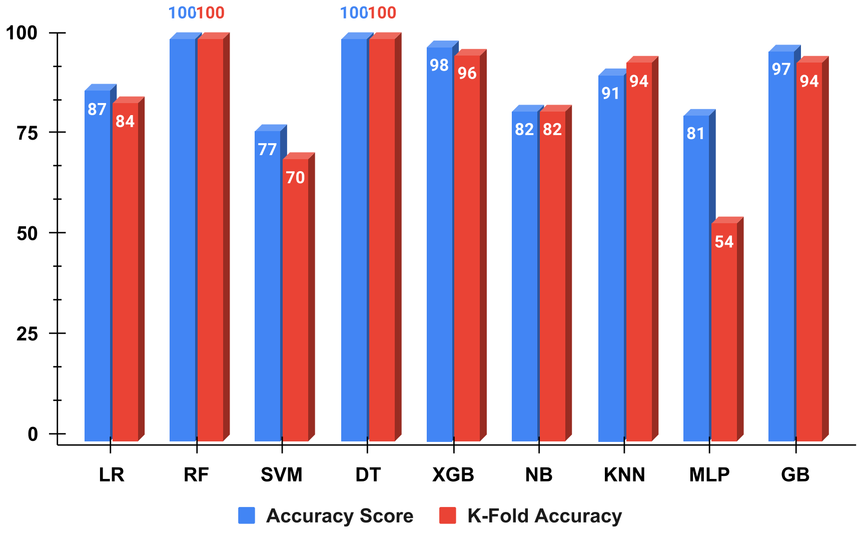

FIGURE 7. The bar chart-based performance validation analysis of applied machine learning models using the k-fold technique.

TABLE 6. The K-fold cross-validation analysis of applied machine learning models.

Models	K-Fold	Accuracy (%)	Standard Deviation
LR	10	84	0.0323 ±
RF	10	100	0.0000 ±
SVM	10	70	0.0322 ±
DT	10	100	0.0000 ±
XGB	10	96	0.0151 ±
NB	10	82	0.0289 ±
KNN	10	94	0.0375 ±
MLP	10	54	0.1064 ±
GB	10	94	0.0261 ±

TABLE 7. Comparing performance with the previously conducted approaches.

Ref.	Year	Learning Type	Proposed Technique	Accuracy (%)
[34]	2020	Deep learning	Hybrid CNN-GRU	94
[35]	2019	Machine Learning	DT	98
[36]	2022	Machine Learning	DT	98
Our	**2023**	**Machine Learning**	**PCHF+DT**	**100**

proposed DT model outperformed the previously proposed techniques. Our proposed study achieved the highest's scores for heat failure prediction with the help of the PCHF feature selection technique.

E. RESULTS COMPARISONS WITH SOTA MODELS

We have performed a performance comparison with state-of-the-art SOTA models in Table 8. The analysis shows that the SOTA model KNN-MCF [37] achieves a 92% performance accuracy score. Another model achieved decent performance results of 99% using the CMSFL-Net [38] approach for image classification problems. This analysis concludes that our proposed approach achieved high-performance scores compared to state-of-the-art models.

TABLE 8. Performance comparisons with state-of-the-art SOTA models.

Ref.	Dataset	Technique	Performance Accuracy (%)
[37]	House Prices	KNN–MCF	92%
[38]	CIFAR-10	CMSFL-Net	99%
Proposed	**Heart Failure**	**PCHF+DT**	**100%**

F. FUTURE RESEARCH

In the future, we will use defiantly deep learning-based techniques to try advanced activation functions, such as weight initialization-based rectified linear unit activation [39]. We will also increase the dataset, including data balancing. The different health parameters related to heart failure will be studied. In addition, the transfer learning-based advanced techniques will be applied.

VI. CONCLUSION

Predicting heart failure using machine learning methods is proposed in this study. The dataset based on 1025 patient records is used to build the applied models. A novel PCHF feature engineering technique is proposed, which selects the eight most prominent features to enhance performance. The logistic regression, random forest, support vector machine, decision tree, extreme gradient boosting, naive base, k-nearest neighbors, multilayer perceptron, and gradient boosting are the applied machine learning techniques in comparison. The proposed DT method achieved 100% accuracy with 0.005 runtime computations. The cross-validation technique based on 10-fold data is applied to each learning model to validate the performance. Our proposed method outperformed the state-of-the-art studies and is generalized for detecting heart failure.

REFERENCES

[1] M. Gjoreski, M. Simjanoska, A. Gradišek, A. Peterlin, M. Gams, and G. Poglajen, "Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers," in *Proc. Int. Conf. Intell. Environments (IE)*, Aug. 2017, pp. 14–19.

[2] G. Savarese and L. H. Lund, "Global public health burden of heart failure," *Cardiac Failure Rev.*, vol. 3, no. 1, p. 7, 2017.

[3] E. J. Benjamin et al., "Heart disease and stroke statistics—2019 update: A report from the American heart association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.

[4] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 156–180, 2021.

[5] C. A. U. Hassan, J. Iqbal, R. Irfan, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, and S. S. Ullah, "Effectively predicting the presence of coronary heart disease using machine learning classifiers," *Sensors*, vol. 22, no. 19, p. 7227, Sep. 2022.

[6] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: A comparative study and analysis," *Health Technol.*, vol. 11, no. 1, pp. 87–97, Jan. 2021.

[7] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *J. Reliable Intell. Environments*, vol. 7, no. 3, pp. 263–275, Sep. 2021.

[8] N. S. Mansur Huang, Z. Ibrahim, and N. Mat Diah, "Machine learning techniques for early heart failure prediction," *Malaysian J. Comput. (MJoC)*, vol. 6, no. 2, pp. 872–884, 2021.

[9] T. Amarbayasgalan, V. Pham, N. Theera-Umpon, Y. Piao, and K. H. Ryu, "An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets," *IEEE Access*, vol. 9, pp. 135210–135223, 2021.

[10] 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, pp. 1–11, Jul. 2021.

[11] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, pp. 1–8, 2019.

[12] D. K. Plati, E. E. Tripoliti, A. Bechlioulis, A. Rammos, I. Dimou, L. Lakkas, C. Watson, K. McDonald, M. Ledwidge, R. Pharithi, J. Gallagher, L. K. Michalis, Y. Goletsis, K. K. Naka, and D. I. Fotiadis, "A machine learning approach for chronic heart failure diagnosis," *Diagnostics*, vol. 11, no. 10, p. 1863, Oct. 2021.

[13] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A method for improving prediction of human heart disease using machine learning algorithms," *Mobile Inf. Syst.*, vol. 2022, pp. 1–9, Mar. 2022.

[14] S. Sarah, M. K. Gourisaria, S. Khare, and H. Das, "Heart disease prediction using core machine learning techniques—A comparative study," in *Advances in Data and Information Sciences*. Springer, 2022, pp. 247–260.

[15] C. Trevisan, G. Sergi, and S. Maggi, "Gender differences in brain-heart connection," *Brain and Heart Dynamics*. 2020, pp. 937–951.

[16] M. S. Oh and M. H. Jeong, "Sex differences in cardiovascular disease risk factors among Korean adults," *Korean J. Med.*, vol. 95, no. 4, pp. 266–275, Aug. 2020.

[17] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," *Int. J. Pharmaceutical Res.*, vol. 12, no. 4, pp. 56–66, 2020.

[18] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, no. 5, pp. 241–266, Oct. 2013.

[19] S. Ekiz and P. Erdogmus, "Comparative study of heart disease classification," in *Proc. Electr. Electron., Comput. Sci., Biomed. Eng. Meeting (EBBT)*, Apr. 2017, pp. 1–4.

[20] 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.

[21] V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: A survey," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 684–687, 2018.

[22] *Heart Disease Dataset|Kaggle*, DAVID LAPP, Atlanta, Georgia, 1988.

[23] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and KNN models for the text classification," *Augmented Human Res.*, vol. 5, no. 1, pp. 1–16, Dec. 2020.

[24] 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, Sep. 2021, Art. no. 104672.

[25] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting employee attrition using machine learning approaches," *Appl. Sci.*, vol. 12, no. 13, p. 6424, Jun. 2022.

[26] A. Raza, H. U. R. Siddiqui, K. Munir, M. Almutairi, F. Rustam, and I. Ashraf, "Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction," *PLoS ONE*, vol. 17, no. 11, Nov. 2022, Art. no. e0276525.

[27] S. Shabani, S. Samadianfard, M. T. Sattari, A. Mosavi, S. Shamshirband, T. Kmet, and A. R. Várkonyi-Kóczy, "Modeling pan evaporation using Gaussian process regression K-nearest neighbors random forest and support vector machines; comparative analysis," *Atmosphere*, vol. 11, no. 1, p. 66, Jan. 2020.

[28] 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.

[29] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving heart disease prediction using feature selection approaches," in *Proc. 16th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2019, pp. 619–623.

[30] 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, Sep. 2021.

[31] R. Pahuja and A. Kumar, "Sound-spectrogram based automatic bird species recognition using MLP classifier," *Appl. Acoust.*, vol. 180, Sep. 2021, Art. no. 108077.

[32] U. Azmat, Y. Y. Ghadi, T. A. Shloul, S. A. Alsuhibany, A. Jalal, and J. Park, "Smartphone sensor-based human locomotion surveillance system using multilayer perceptron," *Appl. Sci.*, vol. 12, no. 5, p. 2550, Feb. 2022.

[33] J. Isabona, A. L. Imoize, and Y. Kim, "Machine learning-based boosted regression ensemble combined with hyperparameter tuning for optimal adaptive learning," *Sensors*, vol. 22, no. 10, p. 3776, May 2022.

[34] A. A. Ali, H. S. Hassan, and E. M. Anwar, "Heart diseases diagnosis based on a novel convolution neural network and gate recurrent unit technique," in *Proc. 12th Int. Conf. Electr. Eng. (ICEENG)*, Jul. 2020, pp. 145–150.

[35] O. E. Taylor, P. S. Ezekiel, and F. B. D. Okuchaba, "A model to detect heart disease using machine learning algorithm," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 11, pp. 1–5, Nov. 2019.

[36] D. K. Chohan and D. C. Dobhal, "A comparison based study of supervised machine learning algorithms for prediction of heart disease," in *Proc. Int. Conf. Comput. Intell. Sustain. Eng. Solutions (CISES)*, May 2022, pp. 372–375.

[37] K. Sanjar, O. Bekhzod, J. Kim, A. Paul, and J. Kim, "Missing data imputation for geolocation-based price prediction using KNN–MCF method," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 4, p. 227, Apr. 2020.

[38] B. Olimov, B. Subramanian, R. A. A. Ugli, J.-S. Kim, and J. Kim, "Consecutive multiscale feature learning-based image classification model," *Sci. Rep.*, vol. 13, no. 1, p. 3595, Mar. 2023.

[39] B. Olimov, S. Karshiev, E. Jang, S. Din, A. Paul, and J. Kim, "Weight initialization based-rectified linear unit activation function to improve the performance of a convolutional neural network model," *Concurrency Comput., Pract. Exp.*, vol. 33, no. 22, p. e6143, Nov. 2021.

AZAM MEHMOOD QADRI received the master's degree in information technology from the Department of Computer Science and Information Technology, Virtual University, Lahore, Pakistan, in 2017. He is currently pursuing the M.S. degree in computer science with the Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan. His current research interests include machine learning, deep learning, and artificial intelligence.

ALI RAZA received the B.Sc. degree in computer science from the Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan, in 2021, where he is currently pursuing the M.S. degree in computer science. His current research interests include data science, artificial intelligence, data mining, natural language processing, machine learning, deep learning, and image processing.

KASHIF MUNIR received the B.Sc. degree in mathematics and physics from Islamia University of Bahawalpur, Pakistan, in 1999, the first M.Sc. degree in information technology from Universiti Sains Malaysia, in 2001, the second M.S. degree in software engineering from the University of Malaya, Malaysia, in 2005, and the Ph.D. degree in informatics from the Malaysia University of Science and Technology, Malaysia, in 2015. He has been in the field of higher education, since 2002. After an initial teaching experience with Binary College, Malaysia, for one semester; and Stamford College, Malaysia, for four years. Later, he relocated to Saudi Arabia. He was with the King Fahd University of Petroleum and Minerals, Saudi Arabia, from September 2006 to December 2014. He moved to the University of Hafr Al-Batin, Saudi Arabia, in January 2015. In July 2021, he joined the Khwaja Farid University of Engineering and Information Technology, Rahim Yar Khan, as an Assistant Professor with the Information Technology Department. He has published journal articles, conference papers, books, and book chapters. He was in the technical program committee of many peer-reviewed journals and conferences, where he has reviewed many research articles. His current research interests include cloud computing security, software engineering, and project management.

MUBARAK S. ALMUTAIRI received the B.Sc. degree in systems engineering from the King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, in 1997, the M.Sc. degree in industrial and systems engineering from the University of Florida, Gainesville, FL, USA, in 2003, and the Ph.D. degree in systems design engineering from the University of Waterloo, Waterloo, Canada, in 2007. From 1997 to 2000, he was an Industrial Engineer with Saudi Arabia Oil Company (Aramco). He is currently the Dean of the Applied College, University of Hafr Albatin (UHB), Hafr Albatin, Saudi Arabia, where he is also an Associate Professor with the Computer Science and Engineering Department. His current research interests include decision analysis, expert systems, risk assessment, information security, fuzzy logic, and mobile government application.