

Received 15 March 2024, accepted 15 April 2024, date of publication 30 April 2024, date of current version 8 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3395512

## RESEARCH ARTICLE

# Synergistic Feature Engineering and Ensemble Learning for Early Chronic Disease Prediction

HAMDI A. AL-JAMIMI 

Information and Computer Science Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia  
Research Excellence, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia

e-mail: aljamimi@kfupm.edu.sa

**ABSTRACT** Chronic diseases, a global public health challenge, necessitate the deployment of cutting-edge predictive models for early diagnosis and personalized interventions. This study presents an advanced methodology for early prediction of chronic diseases, including heart attack, diabetes, breast cancer, and kidney disease, leveraging a synergistic combination of cutting-edge techniques. Recognizing the challenge posed by extensive medical datasets with numerous features, we introduce a novel approach that begins with Feature Engineering using Recursive Feature Elimination (RFE) in conjunction with a Support Vector Machine (SVM). The presented methodology identifies and removes irrelevant features to simplify data complexity. The refined dataset is then input into the robust eXtreme Gradient Boosting (XGBoost) classifier, known for its efficiency and adeptness in predicting complex relationships within the data. The chosen ensemble learning algorithm demonstrates significant prowess in inducing intricate patterns crucial for chronic disease prediction. To enhance model performance, an essential phase of optimization is introduced through hyperparameter tuning using Bayesian optimization. This strategically navigates the hyperparameter space, maximizing the efficiency of the search process and fine-tuning the model for optimal predictive accuracy. The proposed approach showcases a substantial improvement in the early prediction of chronic diseases, demonstrating the effectiveness of the proposed approach.


**INDEX TERMS** Artificial intelligence, chronic disease, early prediction, machine learning, XGBoost.

## I. INTRODUCTION

The worldwide occurrence of chronic diseases constitutes a critical concern in the healthcare domain. Chronic diseases continue to exert a significant impact on global health and pose immense challenges for healthcare systems worldwide [1]. These long-term medical conditions, such as cardiovascular diseases, cancer, kidney disease, and diabetes, not only result in substantial morbidity and mortality but also impose a substantial economic burden on societies [2]. As the prevalence of chronic diseases continues to rise due to aging populations and lifestyle changes, there is an urgent need to develop innovative and accurate predictive models for early detection and personalized interventions. Early detection of chronic diseases is pivotal for preventing their progression, reducing associated complications, and improving the overall

prognosis for affected individuals. Timely recognition of such diseases during their emerging stages is of utmost importance to mitigate their potential seriousness [3]. The nature of several chronic diseases, including heart disease, diabetes, and specific cancers, is characterized by asymptomatic early stages, delaying crucial detection and intervention until significant, possibly irreversible, progression.

Recent strides in artificial intelligence (AI) have profoundly influenced medical and healthcare research [4]. AI's capacity to pinpoint patients at an elevated risk of developing specific diseases enables the implementation of early intervention and prevention strategies. This transformative capability empowers healthcare professionals to identify chronic diseases at earlier stages, subsequently alleviating the burden on both patients and healthcare systems [5], [6]. A compelling example of the advantages of early diagnosis is evident in the work of Di Biasi et al. [7], which illustrates the advantages of early detection, particularly in melanoma.

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

Furthermore, it emphasizes the critical role of applying AI techniques within the relevant clinical context.

These advancements contribute to a more comprehensive exploration of the pivotal role of early diagnosis and the strategic application of AI in medical research. By continuously learning and adapting to new data, AI models can provide personalized and precise predictions, helping healthcare providers develop tailored treatment plans for each patient's unique risk factors [8]. AI techniques provide an effective means of extracting valuable information from data. Machine Learning (ML) algorithms show substantial potential for revealing valuable insights and latent patterns in complex medical data. This capability extends to enhancing diagnostic accuracy, treatment planning, and ultimately improving patient outcomes [9], [10]. With the exponential growth in medical data availability, researchers have explored diverse ML techniques to improve early detection and prediction capabilities. This collective effort contributes to a global initiative aimed at combating chronic diseases on a broader scale [11].

XGBoost is a notable ML algorithm that builds upon the gradient-boosting decision tree algorithm. XGBoost outperforms its predecessor in both accuracy and generalization, confidently predicting even unseen data [12]. It has demonstrated exceptional performance in various domains, including healthcare [13], [14], [15]. While the XGBoost algorithm possesses notable strengths, suboptimal performance can arise from its default configuration due to insufficient parameter optimization. This limitation hinders the effective fitting of the dataset, thereby impeding its generalization capacity and adaptability [16]. Thus, in this study, Bayesian optimization (BO) was employed to effectively fine-tune the hyperparameters and enhance the model efficiency [17]. The incorporation of BO into hyperparameter optimization plays a pivotal role in achieving optimal performance, mitigating the risk of overfitting, and improving the model's robustness [18]. The combination of XGBoost and BO offers a compelling hybrid approach that harnesses the strengths of both algorithms to create sophisticated and accurate predictors for chronic diseases.

Recognizing the challenge posed by extensive medical datasets with numerous features, we introduce a novel approach that begins with Feature Engineering using Recursive Feature Elimination (RFE [19]) in conjunction with a Support Vector Machine (SVM [20], [21]). The iterative nature of RFE progressively eliminates less impactful variables based on SVM weights, yielding a subset of the most relevant features. Subsequently, the refined dataset is fed into the powerful hybrid BO–XGBoost model to predict chronic diseases by considering the impact of symptoms on disease presence. By synergizing these two techniques, the developed models achieved improved predictive capabilities and optimization efficiencies for chronic disease prediction. The breadth and depth of our validation process were ensured by using six datasets covering three distinct chronic diseases – heart attack, breast cancer, diabetes, and kidney disease – to

assess the hybrid model's practicality and efficiency. These diseases were chosen due to their widespread prevalence, significant healthcare implications, and varied manifestations. The proposed model seeks to overcome the constraints of traditional AI-based predictive models by capitalizing on the strengths of XGBoost and BO within a unified framework and incorporating a feature selection technique.

## II. AI-BASED PREDICTION IN CHRONIC DISEASES

The ML domain has garnered increasing interest in its application to predicting chronic diseases. With the ever-growing availability of medical data, researchers have explored various ML techniques to enhance early detection and prediction capabilities, contributing to a collective effort to combat chronic diseases worldwide. This section offers an overview of the extensive literature addressing the use of ML for predicting chronic diseases, with a particular focus on diseases significantly impacting public health. To navigate this vast body of research, we selected a set of representative papers encapsulating a wide spectrum of chronic diseases and ML methodologies, providing insights into the state-of-the-art and laying the foundation for our specific investigation. Table 1 summarizes the selected representative research papers, grouping them by the chronic diseases under investigation and highlighting the ML methods employed.

**TABLE 1. Summary of literature review.**

Study	Investigated disease	ML methods
[22]	Diabetes and breast cancer	SVM, NB, and DT
[23]	Diabetes, kidney, and heart attack	APD
[24]	Diabetes, heart attack, and breast cancer	CNN
[25]	Heart, hepatitis, diabetes, breast cancer	Correlation-based
[26]	Heart disease, hepatitis, diabetes, breast cancer and kidney disease	Stacked ensemble
[27]	Diabetes, breast cancer and kidney diseases	Rough K-means clustering
[28]	Breast cancer	DT, NB, k-NN, and SVM
[29]	Breast cancer	ANN
[30]	Breast cancer	Deep Learning and Light Boosting Classifier
[31]	Diabetes	PyCaret classifiers
[32]	Diabetes	CNN
[33]	Diabetes	RF, NB, and J48 DT
[34]	Diabetes	XGBoost
[35]	Heart disease	RF, SVM, k-NN, and DT
[36]	Heart disease	ANN
[37]	Heart disease	SVM, DT, RF, Gradient Boosting
[38]	Heart disease	RF, k-NN, and AdaBoost
[39]	Kidney disease	DT, k-NN and NB
[40]	Kidney disease	AdaBoost on SVM
[41]	Kidney disease	CNN
[42]	Kidney disease	Stratified Logistic Regression
[43]	Kidney disease	eXplainable AI

These studies span a wide spectrum, including diabetes, breast cancer, heart disease, hepatitis, and kidney disease. Notable ML techniques employed encompass traditional models such as SVM, Naïve Bayes (NB), Random Forest (RF), k-Nearest Neighbors (k-NN), Artificial Neural Networks (ANN), and Decision Tree (DT), to advanced methods like Adaptive Probabilistic Divergence (APD), Convolutional Neural Network (CNN), Correlation-based approaches, Stacked Ensemble, Rough K-means clustering, Deep Learning, and Light Boosting Classifier, among others.

While these studies contribute valuable insights, common limitations emerge. For instance, employing a Correlation-based approach overlooks non-linear relationships and dependencies in the data, which limits the model's ability to capture intricate patterns [25]. Similarly, the application of Rough K-means clustering to predict diabetes, breast cancer, and kidney diseases struggles with high-dimensional and noisy data, impacting the reliability of the identified clusters [27]. Additionally, employing SVM and NB for diabetes and breast cancer prediction faces challenges in handling complex relationships and non-linear patterns inherent in medical data. The rigidity of SVM and the assumptions in NB could limit their adaptability to intricate disease-related datasets [22]. Furthermore, utilizing SVM, DT, RF, and Gradient Boosting for heart disease prediction, faces interpretability challenges due to the complexity introduced by the proposed approach [37]. Identifying and addressing these limitations is crucial for enhancing the overall efficacy of disease prediction models. Identifying and addressing these limitations is crucial for enhancing the overall efficacy of disease prediction models.

The present study aims to contribute to the evolving field of chronic disease prediction by offering a robust and effective solution. The emphasis on feature selection, adaptability, and ensemble learning positions the proposed model as a valuable advancement. The proposed model offers a novel approach to disease prediction. The Advanced Feature-Selection-Based Hybrid BO–XGBoost Model stands out for several key strengths. It integrates advanced feature selection techniques to ensure the inclusion of relevant features while mitigating the impact of noise. The incorporation of Bayesian Optimization enhances adaptability to varying datasets and optimizes hyperparameters for improved performance. By adopting ensemble learning with XGBoost, our proposed model harnesses the strengths of various models while ensuring interpretability. This approach presents a promising solution to overcome the limitations identified in existing methodologies.

### III. INTEGRATED LEARNING ENSEMBLE FRAMEWORK

The illustrated framework in Figure 1 integrates a detailed methodology for early chronic disease prediction. The framework includes a feature selection method to identify relevant variables, a hybrid ensemble model that combines BO and XGBoost algorithms for improved predictive accuracy, and an experimental setup for modeling.

### A. SVM-RFE-BASED FEATURE ENGINEERING

Feature engineering holds substantial potential for advancing chronic disease prediction. Medical datasets comprise a multitude of features, some of which may not significantly contribute to chronic disease prediction and might introduce noise, RFE proves invaluable in managing high-dimensional data by selecting a subset of the most informative features [44]. Combining RFE with SVM plays a crucial role in identifying and selecting essential features, which are pivotal for constructing an efficient predictive model [45]. The application of RFE with SVM in the context of chronic disease prediction contributes to the development of a more precise, comprehensible, and clinically significant predictive model.

Given a dataset with input features  $X$  and corresponding labels  $y$ , the SVM aims to find a hyperplane that separates the data into different classes. The decision function for SVM is defined as:

$$f(X) = \text{sign}\left(\sum_{i=1}^n \alpha_i y_i K(X_i, X) + b\right) \quad (1)$$

where  $X_i$  represent the training samples in the dataset,  $y_i$  are the corresponding labels ( $-1$  or  $1$ ),  $K$  is the kernel function,  $\alpha_i$  are the Lagrange multipliers, and  $b$  is the bias term. The SVM training involves solving the optimization problem to find the optimal values of  $\alpha_i$  that maximize the margin between classes, subject to the constraint that  $\sum_{i=1}^n \alpha_i y_i = 0$ . RFE is an iterative feature selection method that eliminates the least important features based on the weights assigned by a model, which, in this case, is the SVM.

The general steps of RFE are as follows:

1. Train the SVM on the entire set of features.
2. Use the learned weights from the SVM to rank the importance of each feature.
3. Remove the least important feature(s).
4. Repeat steps 1-3 until the desired number of features is reached.

The combination of RFE with SVM involves using the SVM's decision function and weights to determine feature importance. The iterative nature of RFE allows it to progressively eliminate less impactful variables, enhancing the efficiency of feature selection. The SVM weight represents the contribution of a feature to the overall model. A higher SVM weight indicates that the feature has a stronger impact on the model's decision boundary. However, the RFE rank considers the feature's importance in a different context—it evaluates how well the model performs when the feature is removed. Therefore, a feature with a high SVM weight may have a low RFE rank if its removal doesn't significantly affect the model's performance. In contrast, a feature with a low SVM weight might have a high RFE rank if its absence noticeably impairs the model's predictive accuracy. In essence, SVM weight and RFE rank offer complementary insights—one focusing on the feature's influence within the model and the other on its importance in the absence of other features. These differences in perspective can lead to

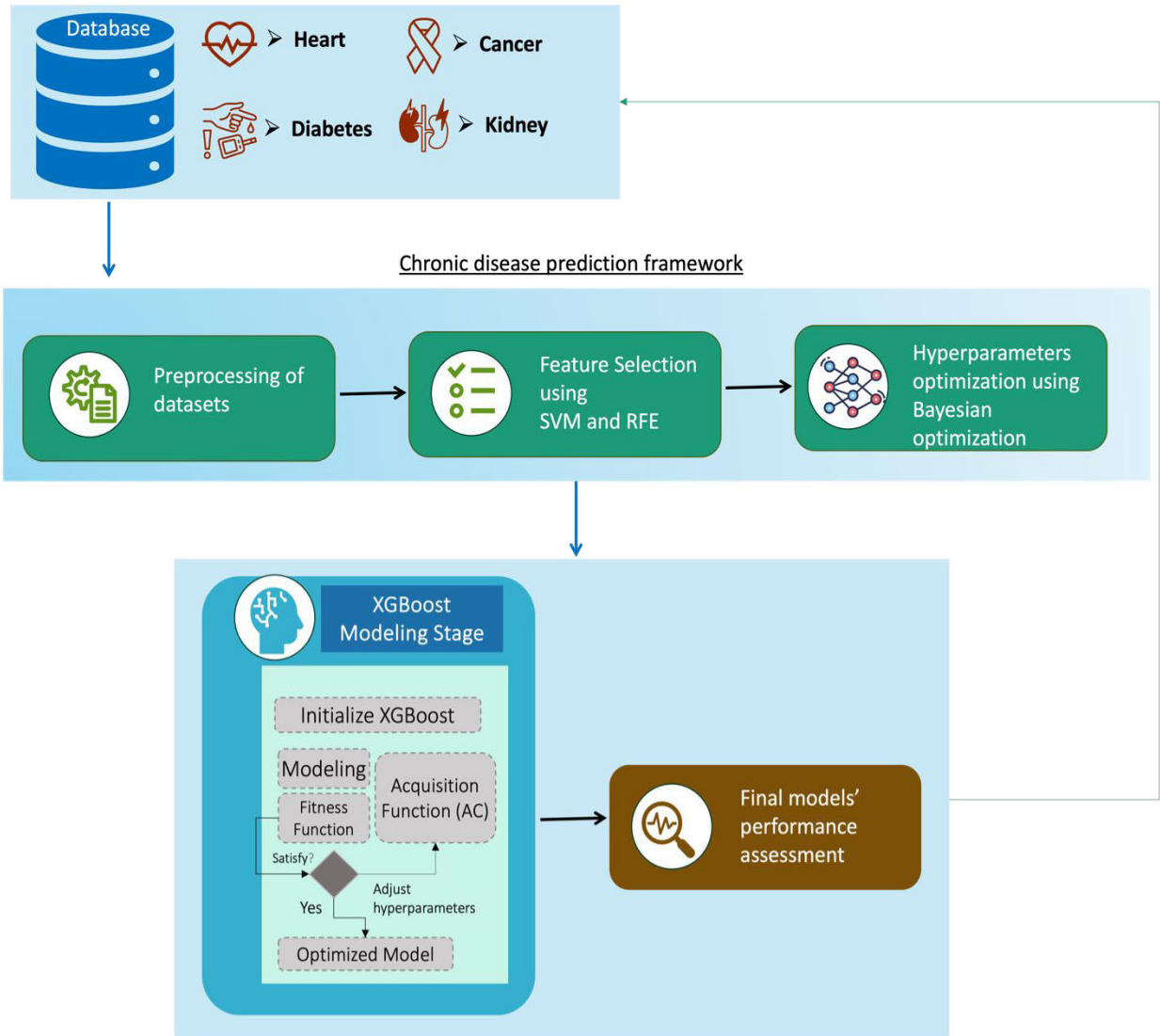


FIGURE 1. The proposed chronic disease prediction framework.

scenarios where high SVM weights align with low RFE ranks and vice versa.

**B. EXTREME GRADIENT BOOSTING ALGORITHM**

The XGBoost algorithm is an ensemble learning method based on gradient boosting [46], incorporating two crucial optimization enhancements based on gradient-boosting decision trees [47]. XGBoost has been used in data mining and has extensive applications for various problems [48]. Notably, XGBoost demonstrates advantageous traits such as fast computation, robustness, and accurate prediction [49], [50]. However, few studies have explored the potential of XGBoost in the treatment of chronic diseases.

XGBoost mitigates overfitting through two key features: its regularized objective function and second-order Taylor expansion of the loss function. XGBoost incorporates L1 (LASSO) and L2 (ridge) regularization terms into its

objective function, penalizing model complexity. In addition, XGBoost approximates the loss function using a second-order Taylor series expansion (SOT). This approximation simplifies the optimization problem and leads to more accurate loss estimates, ultimately enhancing model fitting and accuracy. These optimizations contribute to the effectiveness and robustness of XGBoost as a powerful machine-learning method.

Assume that we have a dataset denoted as  $D = \{(x_i, y_i)\} (i = 1, 2, \dots, n)$ , where  $x_i$  represents the input data and  $y_i$  represents the corresponding target or output data. In this dataset, we have  $n$  examples, numbered 1 to  $n$ , each comprising an input  $x_i$  and its corresponding target  $y_i$ . After training the model with  $K$  trees, the predicted results ( $\hat{y}_i$ ) obtained from the model are as follows:

$$\hat{y}_i = \sum_{k=1}^n f_k(x_i), f_k \in W \tag{2}$$



where  $W$  refers to the modeling space for the regression trees. Each regression tree, denoted as  $f(x)$ , represents an individual model in an ensemble. The predicted result  $y_i$  is obtained by combining the outputs of all  $K$  regression trees in the model.

$$W = \{f(x) = \mu_{l(x)}\} \tag{3}$$

where  $l(x)$ ,  $\mu$  are the leaf node and its score of the  $x$ th sample, respectively. The predicted value of the  $i$ th iteration is computed as follows:

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i) \tag{4}$$

$f_t(x_i)$  is the target function, which is optimized as follows:

$$J(f_t) = \sum_{i=1}^n J(y_i, \hat{y}_i^{t-1} + f_t(x_i) + \psi(f_t)) \tag{5}$$

where  $J(\cdot)$  and  $\psi(f_t)$  are the cost function and the model complexity, respectively.

$$\psi(f_t) = \gamma \cdot T_t + 0.5\lambda \sum_{j=1}^T \mu_j^2 \tag{6}$$

where  $\gamma$  sets the severity of the penalty for model complexity, with higher values encouraging fewer decision tree leaves  $T$ . Larger leaves will be regularized by  $\gamma$ . Equation (6) can be simplified using SOT as:

$$J(f_t) = \sum_{i=1}^n L(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) + \psi(f_t) \tag{7}$$

$$g_i = \frac{\delta L(y_i, \hat{y}_i^{t-1})}{\delta \hat{y}_i^{t-1}} \tag{8}$$

$$h_i = \frac{\delta^2 L(y_i, \hat{y}_i^{t-1})}{\delta \hat{y}_i^{t-1}} \tag{9}$$

Therefore, the computed objective function has the following form:

$$J(f_t) = \sum_{i=1}^n [g_i \mu_{l(x_i)} + \frac{1}{2} h_i \mu_{l(x_i)}^2] + \gamma T + \lambda \frac{1}{2} \sum_{j=1}^T \mu_j^2 \tag{10}$$

XGBoost offers various hyperparameters that require optimization. Optimizing these hyperparameters enables the fine-tuning of the XGBoost model to achieve optimal performance and mitigate overfitting. Each hyperparameter plays a specific role in controlling the complexity, generalization ability, and computational efficiency of the model.

Enhancing the efficiency and accuracy of models can be achieved by implementing an efficient hyperparameter optimization algorithm [51]. Grid search is widely employed as a hyperparameter optimization method because it exhaustively explores the parameter space; however, dimensionality limitations arise owing to the exhaustive nature of the search, leading to computational challenges. Given the limitations of random search for complex models, exploring alternative hyperparameter optimization techniques for efficient training becomes crucial [52]. Conventional optimization methods often struggle with the complexities of ML algorithms. Bayesian optimization [53] as a powerful method proves its

effectiveness in optimizing ANNs, SVMs, and other models [54], [55], [56] and paves the way for tackling more intricate challenges.

### C. BAYESIAN OPTIMIZATION

The optimization goal in hyperparameter tuning is often to control the maximum or minimum value of the objective function using a limited set of sampling points. Given that the function is unknown and its evaluation can be computationally expensive, an alternative approach is necessary to effectively address this challenge [51]. Heuristic optimization algorithms are commonly employed in such scenarios. These algorithms explore the search space by iteratively sampling points and evaluating objective functions at these points [57]. Examples of such algorithms include Bayesian optimization, genetic algorithms [58], particle swarm optimization [59], and simulated annealing [60]. Crucially, these methods do not rely on the mathematical form or convexity of the objective function and can handle complex, non-convex, and computationally expensive optimization problems.

By leveraging these optimization algorithms, researchers can efficiently search for optimal hyperparameter configurations through iterative sampling and evaluation of the unknown objective function to find the maximum or minimum value within the given computational resources. The position at which the optimization function is computed, denoted as  $p^+$ , is computed as follows:

$$p^+ = \arg \max_{p \in \emptyset} \vartheta(p) \tag{11}$$

where  $\vartheta$  refers to an unknown objective function, and  $p$  and  $\emptyset$  denote the sampling point and the search space of  $p$ , respectively.

Indeed, BO proved to be an exceptionally competent optimization method [53]. By incorporating prior knowledge about the objective function  $\vartheta$  with observations from strategically sampled points, BO updates its belief about the function's distribution. This dynamic posterior estimation drives its efficient search for the optimum. This approach enables the algorithm to iteratively refine its understanding of the behavior and uncertainty of a function. By utilizing this posterior information, BO can effectively evaluate the global optimal value [51], [61], making it a powerful tool for optimizing complex and computationally expensive functions encountered in various real-world applications. This includes the optimization of hyperparameters in ML models.

There are two main tasks achieved by BO [62]. To model the given data and update the posterior distribution, a Gaussian process (GP) is first chosen. The feature that defines GP is that a multivariate Gaussian distribution on  $\beta^z$  is induced by the finite collection of points  $p_z \in \emptyset_{z=1}^Z$ . The equivalent function  $\vartheta(p_z)$  is defined as the  $z^{th}$  point; from there, the marginals and conditionals of this distribution can be calculated.

Second, an acquisition function (AC) is chosen to identify the subsequent evaluation point in the search space. The posterior over function  $\vartheta(p)$  is induced if and only if the

function  $\vartheta(p)$  is derived from a GP prior and the observations take the form of  $p_z, \varepsilon_n^z_{z=1}$ , where  $\varepsilon_z \sim N(\vartheta(p_z), v)$  represents the  $z^{th}$  measured model performance and  $\vartheta$  is the variance of the noise.

The next evaluation point in the search space, denoted by  $p_{next} = \text{argmax}_p \alpha(p)$ , is then found using this posterior. The predictive variance function  $\sigma^2(p)$  and predictive mean function  $\omega(p)$  of the GP model rely on the acquisition function. The next sampling point that is most likely to produce the ideal value for the unknown objective function  $\vartheta$  can be found using these essential functions. Using the acquisition function and repeated GP model updates, BO effectively searches the search space and converges to the ideal point  $p^*$  where  $\vartheta$  is maximized.

The likelihood of improvement and the GP upper confidence bound (GP-UCB) are two methods that can be used to solve the AC optimization issue [63]. To decide whether to explore other low-confidence regions (representing the high  $\sigma(p)$  zone) that may yield better performance in hyperparameter tuning or to take advantage of the current optimal value (representing the high  $\omega(p)$  zone), the GP-UCB method is utilized in this case. To achieve a balance between the two options, parameter  $k$  is added. Here is how the function is expressed:

$$\alpha_{UCB} = \omega(p) - k\sigma(p) \tag{12}$$

Figure 2 outlines the BO algorithm, which comprises two main components: Step 2 optimizes the acquisition function and updates the posterior distribution, as illustrated in Steps 3 and 4. These steps work in tandem to iteratively refine the model's understanding of the objective function and select the next sampling point in the search space, ultimately guiding the optimization process toward the global maximum or minimum of the unknown function.

1. Repeat the following  $n$  times:
  - a. Find  $P_x$  through the optimization of AC  $\alpha(p)$   
 $- P_x = \text{argmax}_p \alpha(p)$
  - b. Perform sampling of the function  $\vartheta(P_x)$
  - c. Increase the data  $\phi_{1:x} = \phi_{1:x-1}, (P_x, \vartheta(P_x))$
  - d. Update the posterior of the function  $\vartheta$

FIGURE 2. Bayesian optimization algorithm.

#### D. HYBRID BO-XGBOOST MODEL

This section presents a comprehensive analysis of the proposed BO-XGBoost model, outlining its three core stages: data preparation, feature set development, and model construction and validation.

To enhance the feature selection process, XGBoost was employed using a gain metric to determine the optimal split nodes. This feature selection step enables the identification of the most influential features, contributing to improved model

performance and efficiency.

$$\begin{aligned} & \text{gain} \\ &= \frac{1}{2} \left[ \frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] \gamma \end{aligned} \tag{13}$$

In the XGBoost feature selection process, following segmentation, the left node samples are denoted as  $I_L$ , and the right node samples are denoted as  $I_g$ . The total sample  $I$  is the sum of  $I_L$  and  $I_g$ , representing all samples considered. The gain score is a metric used to measure feature importance in XGBoost. A higher gain score indicates a higher feature importance score, indicating that the corresponding feature is more crucial and effective in the model [64]; in other words, features with higher gain scores have a more substantial impact on the predictive performance of the model and play a more significant role in determining the target variables. Thus, during the feature selection process, XGBoost prioritizes features with higher gain scores because they contribute more to improving the model's accuracy and overall performance.

The second stage, the BO-XGBoost model, consists of four distinct steps:

- 1) Establish the objective function for optimization and initialize the XGBoost model using (2) to (12).
- 2) Define the search domain for hyperparameters through BO optimization.
- 3) Develop a probabilistic model during optimization using a GP, based on the conducted search iterations. Subsequent hyperparameter selection is guided by maximizing the acquisition function until a predetermined number of iterations is reached.
- 4) Record the optimization results, including hyperparameter values and XGBoost verification errors, for each candidate parameter set.

#### IV. EXPERIMENTAL SETUP

The current research was initiated by carefully selecting datasets and applying preprocessing techniques. This section presents the experimental setups, including dataset selection, data processing, and performance evaluation metrics.

##### A. DATA COLLECTION

Efficient evaluation of AI systems in healthcare requires a thorough examination of their performance across diverse datasets [65], [66]. Due to the diverse nature of disease prevalence, manifestations, and risk factors across regions and ethnicities, there exists a potential for biased or inaccurate predictions when models are trained on limited data from underrepresented populations. It is imperative to address this critical issue [67]. Recognizing and mitigating regional differences in patient data is essential to ensure the fairness and reliability of AI-driven healthcare insights. This approach is crucial for minimizing health disparities and promoting an

inclusive and equitable application of advanced technologies in global healthcare practices.

This study employed a rigorous data-driven approach, utilizing multiple publicly available datasets for heart disease, breast cancer, diabetes, and kidney disease. These datasets encompassed pertinent features such as demographics, laboratory tests, and disease-specific indicators.

The selection of datasets for this study was guided by a meticulous approach aimed at ensuring diversity and representativeness across various investigated chronic diseases. The objective was to construct a comprehensive evaluation framework for the proposed methodology, taking into consideration the multifaceted nature of real-world medical datasets. For a thorough assessment, datasets with varying numbers of features, ranging from 9 to 32, were selected. This diversity enables examining the adaptability of the model to datasets with differing levels of complexity, reflecting the diverse nature of medical data in practice.

Moreover, the significant variation in sample sizes among the chosen datasets, ranging from 333 to 1025 samples, was carefully considered. This selection allows for a thorough evaluation of the methodology's scalability and performance across datasets with distinct scales, mirroring the variability encountered in real-world scenarios.

Recognizing the prevalence of imbalanced data in medical datasets, datasets with imbalanced class distributions, such as the Abu Dhabi dataset for kidney disease, were intentionally included. This strategic inclusion enables a thorough evaluation of the methodology's effectiveness in scenarios where positive cases may be considerably outnumbered, a frequent situation in medical diagnostics. Furthermore, the selection of datasets placed significant emphasis on geographical and demographic diversity. Datasets such as Sylhet and Abu Dhabi represent various regions and demographics, ensuring a comprehensive examination of the proposed methodology's applicability across diverse contexts.

Table 2 presents a summarized view of the key characteristics of the selected datasets employed in this study. In the following, a more detailed description of each dataset has been presented, offering comprehensive insights into their distinctive features and contextual relevance.

- The Heart Attack - Erbil dataset was collected at the Medical Help Centre, a private hospital and heart center in Erbil, Iraq. Comprising 21 features and 333 patient records, this open dataset serves the primary objective of utilizing native patients' information to predict the likelihood of heart disease. The gathered information is categorically classified into five groups, encompassing demographic details, medical history, physical examinations and symptoms, medical lab tests, and diagnostic features. The selection of dataset features is guided by the recommendations of medical professionals, ensuring the inclusion of relevant and meaningful information for heart disease prediction [68].
- The CHSLB dataset, comprising patients' records from Cleveland, Hungary, Switzerland, and Long Beach,

encompasses both male and female subjects. The dataset comprises a total of 1025 entries, distributed across 13 features, with the class distribution represented as the 14th attribute. Among the individuals studied, 499 individuals were identified as healthy and free from heart disease, while the remaining 526 individuals were categorized as sick. Importantly, the dataset indicates the absence of missing values. The data was sourced from the Kaggle database [69].

- The Wisconsin Diagnostic Breast Cancer (WDBC) dataset captures essential features computed from digitized images of fine needle aspirates (FNA) of breast masses, specifically characterizing cell nuclei. With 569 data points, the dataset classifies instances into 212 malignant and 357 benign cases. The ten features include radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, each having three attributes: mean, standard error, and worst. The dataset incorporates 30 features, offering a representation of cell nuclei characteristics crucial for breast cancer diagnosis [70].
- The Wisconsin Breast Cancer Dataset (WBCD) comprises 699 instances obtained from FNA of human breast tissue. Each record in the database encompasses 10 attributes, as elaborated in Section V. These attributes are assigned integer values ranging from 1 to 10, with 1 indicating proximity to benign and 10 signifying the highest degree of anaplasia. Each sample is accompanied by its corresponding class label, designated as either benign or malignant [71].
- The Diabetes Sylhet Dataset encompasses sign and indication data from individuals who are newly diabetic or at risk of developing diabetes. Collected through direct questionnaires administered by healthcare professionals at Sylhet Diabetes Hospital in Sylhet, Bangladesh, the dataset includes 17 features. With a total of 520 samples, it highlights 320 positive cases and 200 negative cases, providing valuable insights into the characteristics associated with diabetes in the local population [72].
- The Diabetes Pima Dataset comprises data from 768 female diabetic patients belonging to the Pima Indian community in Phoenix, Arizona, all aged 21 years or older. With 9 attributes, the dataset offers a detailed exploration across 768 samples. Among the patients, 500 do not have diabetes, while 268 have been diagnosed with diabetes, resulting in a distribution of 35% for diabetes and 65% for non-diabetic individuals [73].
- The Kidney Disease Abu Dhabi Dataset comprises electronic medical records of 491 patients gathered at Tawam Hospital in Al-Ain City (Abu Dhabi, United Arab Emirates) [74]. This dataset encompasses 22 features and 491 samples, with a predominance of 435 positive cases and 56 negative cases. The patient demographics include 241 women and 250 men, with an average age of 53.2 years, providing a comprehensive

**TABLE 2. Datasets used for classification.**

Chronic Disease	Dataset Name	No. of Features	No. of Samples	No. of Positive Cases	No. of Negative Cases	Dataset Link
Heart	Erbil	20	333	118	215	[68]
	CHSLB	14	1025	526	499	[69]
Breast Cancer	WDBC	32	569	212	357	[70]
	WBCD	10	699	240	459	[71]
Diabetes	Sylhet	17	520	320	200	[72]
	Pima	9	768	268	500	[73]
Kidney	Abu Dhabi	22	491	435	56	[75]
	India	25	400	250	150	[76]

representation of the kidney health profile within this specific population [75].

- The Kidney Disease India Dataset consists of 400 records and 25 features. Among these features are class attributes denoted as CKD and NOTCKD, indicating the presence or absence of chronic kidney disease in the patients. The dataset encompasses 14 categorical and 11 numerical features. Notably, there is a substantial number of missing values, with only 158 records being complete. Moreover, the dataset demonstrates a notable imbalance, with 250 observations (62.5%) classified as CKD and 150 (37.5%) as NOTCKD. This distribution offers a nuanced perspective on the prevalence of chronic kidney disease statuses within the dataset [76].

## B. DATA PREPROCESSING

In the domain of predictive analytics, the significance of data quality cannot be overstated, particularly in the context of real-world medical datasets. The diverse nature and sources of these datasets introduce challenges, including outliers, missing data points, and irrelevant features. These factors can have a substantial impact on the accuracy of subsequent analyses and model training [77]. Recognizing the inherent complexities, our study placed a strong emphasis on rigorous data pre-processing to ensure the reliability and robustness of our predictive models [78].

The pre-processing pipeline involved several crucial steps, each tailored to address specific challenges posed by real-world medical data. Key features were meticulously extracted, taking into consideration their relevance to disease prediction. Simultaneously, measures were implemented to safeguard patient anonymity, aligning with ethical considerations [79]. Cleaning the data involved the identification and handling of outliers, a process vital for accurate model training.

Outliers, recognized as potential sources of bias in the analysis, were addressed through a robust replacement strategy. For numerical features, outliers were replaced with representative values, specifically the mean, to maintain data integrity. For categorical features, a similar approach was employed, replacing outliers with the mode to preserve the categorical distribution [80].

Handling missing data was another critical aspect of our data pre-processing strategy. Imputation methods were

employed based on the type of data — mean imputation for numerical features and mode imputation for categorical features. This approach ensured a comprehensive treatment of missing entries, minimizing the impact on subsequent analyses and predictions.

Challenges encountered during the data pre-processing phase, such as variations in data quality and the presence of anomalies, were diligently addressed to maintain the integrity of the datasets. Our rigorous data pre-processing efforts not only contributed to the accuracy of disease prediction but also ensured the reliability of our findings.

After preprocessing the datasets, they were systematically divided through random sampling, allocating 80% for training and 20% for testing. This division ensures robust model learning with the extensive training set, while the 20% testing set enables a thorough evaluation of the model's generalization performance on unseen data.

## C. PERFORMANCE EVALUATION

In the model validation process, we utilized a separate set of testing data comprising 20% of the dataset. These testing samples were not included in the model training phase, allowing us to evaluate how well the model could make accurate predictions when presented with instances representing a class it hadn't encountered during training. Essentially, this assessment aimed to gauge the model's ability to handle new, previously unseen data during validation, ensuring its robustness and performance beyond the initial training datasets.

A confusion matrix assessed the congruence between actual and predicted outcomes generated by the developed BO-XGBoost models. The analysis involved four key elements commonly used in binary classification metrics: True Positives, True Negatives, False Positives, and False Negatives [81]. These components were employed for further examination, and their detailed metrics are presented in Table 3 for further examination.

Precision measures the accuracy of positive predictions and is vital for avoiding consequences such as unnecessary treatments or missed interventions. Recall measures the model's ability to correctly identify individuals with the disease, which is crucial to preventing delayed treatment and serious complications. The model's performance has been evaluated with accuracy for a general view and F1-score for a balanced perspective on precision and recall. The area under



**TABLE 3. Metrics used to evaluate the performance of ML-based models for classification.**

Metric	Equation
Accuracy	$\frac{TP + TN}{(TP + TN + FP + FN)}$
Precision	$\frac{TP}{(TP + FP)}$
Recall	$\frac{TP}{(TP + FN)}$
F1-Score	$2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$
AU-ROC	$\frac{1}{2} \left( \frac{TP}{(TP + FN)} + \frac{TN}{(TN + FP)} \right)$
MCC	$\frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$

the receiver operating characteristic curve (AU-ROC) is a widely used metric in medical diagnosis and ML that offers a comprehensive evaluation of model performance, especially in the face of class imbalances. A score of 1 signifies a perfect model, while 0.5 indicates performance equivalent to random guessing [82]. Finally, Matthew’s correlation coefficient (MCC) serves as a balanced measure, particularly valuable for imbalanced datasets, ranging from -1 (total disagreement) to 1 (perfect prediction), with 0 indicating random prediction. The MCC is widely utilized in ML research for its ability to assess model performance across various scenarios.

**V. RESULTS AND DISCUSSION**

In this research, predictive models were employed to predict the onset of diverse medical conditions, encompassing heart attacks, breast cancer, diabetes, and kidney disease. Each disease was examined using two datasets, a strategic choice aimed at supporting the robustness, and applicability of the predictive models. Utilizing multiple datasets for each chronic disease takes into account the inherent diversity in healthcare data. This approach establishes a more comprehensive foundation for precise predictions, thereby enhancing the efficacy of strategies aimed at managing and preventing chronic diseases.

Notably, the chosen datasets underscore the issue of class imbalance, where one dataset (CHSLB) exhibits balanced classes (51% and 49%), while six out of eight datasets have one class representing over 60% of the samples. In the Kidney-Abu Dhabi dataset, one class comprises a striking 89% of the data, highlighting the significance of addressing this common challenge for accurate predictions in medical contexts [83], [84]. In addressing the inherent imbalance within the datasets, the proposed methodology incorporated a combination of strategic adjustments within the XGBoost framework. Specifically, the `scale_pos_weight` parameter, a key feature of XGBoost, was fine-tuned to assign appropriate weights to the minority class. This adjustment aimed to mitigate the challenges posed by class imbalance, enhance the model’s learning from instances of the minority class,

and contribute to its overall robustness in predicting chronic diseases. Moreover, a comprehensive assessment of the developed models’ performance on imbalanced datasets was conducted by employing a diverse set of evaluation metrics, aimed at providing detailed insights. Precision, recall, and F1-score were selected to evaluate the model’s ability to accurately identify positive instances while minimizing false positives. These metrics are crucial considerations, especially in the context of imbalanced datasets. Additionally, the use of AU-ROC provided a holistic perspective on the trade-off between true positive and false positive rates. The MCC, as a robust metric accounting for both sensitivity and specificity, further enhanced the evaluation, ensuring a balanced assessment of the model’s performance even in the presence of imbalanced class distributions. By incorporating these metrics into the evaluation process, the aim is not only to demonstrate the model’s effectiveness in handling imbalanced data but also to provide a comprehensive understanding of its impact on early prediction accuracy for chronic diseases.

Table 4 presents the optimal hyperparameter settings for the BO-XGBoost-based models developed in this study; this table serves as a comprehensive reference; offering insights into the hyperparameter choices that result in optimal model configurations. The hyperparameters, namely `colsample_bytree`, `learning_rate`, `maximum_depth`, `N_estimators`, and `subsample` play a crucial role in fine-tuning the models’ performance. The hyperparameter settings were determined through a systematic optimization process that utilized Bayesian optimization techniques. The values presented in the table represent the culmination of an iterative refinement process aimed at enhancing the predictive capabilities and overall effectiveness of the BO-XGBoost models in the context of this study.

The conducted evaluation employed a diverse array of performance metrics to assess the efficacy of the model. The following subsection provides detailed findings from these experiments, with a summary presented in Table 5.

**A. RESULTS ON HEART DISEASE DATASETS**

This experiment, focusing on heart disease prediction, implemented an AI-based model on two distinct datasets related to heart health: the Erbil dataset, which comprises 20 features, and the CHSLB dataset, which comprises 14 features. To optimize the BO-XGBoost model and ensure its ability to accurately predict outcomes in various contexts, feature selection processes were implemented. The feature selection process, utilizing both SVM weights and RFE ranks, aimed to identify the most relevant features for predicting the presence of heart diseases.

Table 6 presents the SVM weights and RFE ranks obtained through the feature selection process for the Erbil dataset. Each feature has an associated weight, indicating its contribution to the SVM model. The weights are typically used to understand the influence of each feature on the model’s decision-making. Each feature also has an RFE rank, which

**TABLE 4. Optimal settings for hyperparameters of the developed BO-XGBoost-based models.**

	Colsample bytree	Learning rate	Max depth	N estimators	Subsample
Heart-Erbil	0.90	0.29	5.19	119	0.94
Heart- CHSLB	0.62	0.02	7.39	190	0.67
Breast Cancer - WDBC	0.98	0.69	5.25	101	0.52
Breast Cancer WBCD	0.83	0.27	6.98	72	0.59
Diabetes- Sylhet	0.50	0.01	10.0	165	0.63
Diabetes- Pima	0.98	0.92	6.92	108	0.78
Kidney-Abu Dhabi	0.95	0.82	8.64	146	0.98
Kidney-India	0.85	0.49	6.68	181	0.65

**TABLE 5. Classification accuracy results obtained from BO-XGBoost classifiers.**

DATASET	Precision	Recall	F1-Score	MCC	Accuracy
Heart-Erbil	1.00	1.00	1.00	1.00	1.00
Heart- CHSLB	1.00	1.00	1.00	1.00	1.00
Breast Cancer - WDBC	1.00	0.97	0.99	0.98	0.99
Breast Cancer- WBCD	0.98	0.98	0.98	0.97	0.99
Diabetes- Sylhet	1.00	1.00	1.00	1.00	1.00
Diabetes- Pima	0.72	0.71	0.71	0.57	0.81
Kidney - Abu Dhabi	0.83	0.62	0.71	0.70	0.96
Kidney-India	1.00	1.00	1.00	1.00	1.00

**TABLE 6. SVM weights and RFE ranks of selected features of the erbil dataset.**

Feature	SVM WEIGHT (RFE RANK)
Age	0.0046, 1
Gender	0.0157, 1
Smoking Status	0.0488, 1
Chest Pain	0.0179, 2
Family History of Heart Disease	0.0248, 5
Lifestyle	0.0129, 1
Diabetes Mellitus	0.0042, 7
Blood Pressure Diastolic	0.0021, 6
Hypertension	0.0655, 1
Interventricular Septal Thickness	0.0440, 4
Electrocardiographic Pattern	1.1044, 1
Presence of Q-Wave	1.1044, 3

indicates its importance after the recursive feature elimination process. A lower rank generally implies higher importance.

The process has identified 12 features out of the original 20 as relevant for the prediction task. The selected features encompass a range of factors, including demographics (age, gender), lifestyle choices (smoking, hypertension), and various medical indicators (chest pain, diabetes, electrocardiographic patterns). For instance, a feature with a high SVM weight, such as Electrocardiographic Pattern, holds substantial influence in the overall model decision, as evidenced by its RFE rank of 1. This aligns with the general understanding that a lower RFE rank implies higher importance. In the context of the Erbil dataset, the removal of the Electrocardiographic Pattern feature significantly impacts predictive accuracy, underscoring its importance in both SVM and RFE assessments. In contrast, the Diabetes Mellitus feature has a modest influence in the SVM model, as indicated by its low SVM weight and higher RFE rank. While it contributes

to the model, its importance is relatively lower compared to features with higher SVM weights and lower RFE ranks. The SVM-RFE analysis offers a valuable initial assessment of feature importance, underscoring the necessity for more in-depth investigations. It suggests the potential incorporation of additional modeling techniques for a comprehensive understanding of the relationships between these features and the target variable. These findings highlight that a combination of demographic, lifestyle, and medical factors significantly contributes to predicting heart disease in the Erbil dataset.

The feature selected from the CHSLB heart dataset unveiled varying degrees of significance in predicting heart-related outcomes. The process has identified 10 features out of the original 14 as the most important features for the prediction task. Table 7 presents the results from the SVM weights and RFE ranks. It provides valuable insights into feature importance in the CHSLP dataset. Features like Chest Pain Type exhibit substantial SVM weights (0.6780) and an RFE rank equal to 1, indicating their significant influence on the model’s decision-making even after recursive feature elimination. Similarly, Thalassemia demonstrates a high SVM weight (0.7260) and an RFE rank of 1, emphasizing its critical role in predictive accuracy. These findings underscore the model’s ability to integrate traditional cardiovascular indicators and specific medical conditions for accurate predictions. Features with high SVM weights and low RFE ranks, such as Exercise-Induced Angina, Maximum Heart Rate Achieved, and Serum Cholesterol, are prioritized in the model, highlighting their importance in capturing nuanced patterns in cardiovascular health. Despite Resting Blood Pressure and Fasting Blood Sugar > 120 mg/dl having low SVM weights, these features have RFE ranks, suggesting their importance in the model.

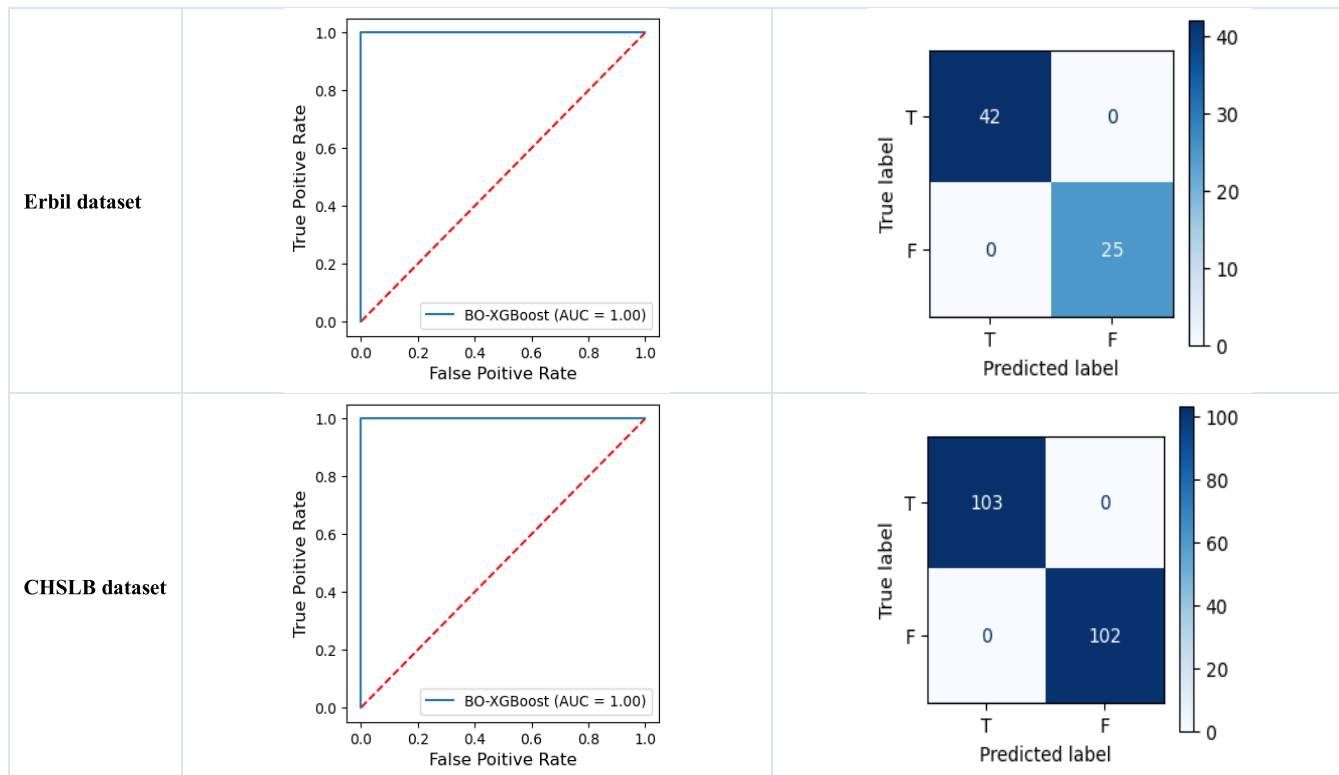


FIGURE 3. Prediction results: Erbil heat disease dataset and CHSLB dataset.

TABLE 7. SVM weights and RFE ranks of selected features of the CHSLB dataset.

Feature	SVM WEIGHT (RFE RANK)
Gender	0.8236, 4
Chest Pain Type	0.6780, 1
Resting Blood Pressure	0.0113, 1
Fasting Blood Sugar > 120 mg/dl	0.0420, 1
Resting Electrocardiographic Results	0.3259, 3
Exercise-Induced Angina	0.7746, 1
Maximum Heart Rate Achieved	0.3968, 1
Slope of the Peak Exercise ST Segment	0.5678, 2
Serum Cholesterol	0.5587, 1
Thalassemia	0.7260, 1

The selection of these features implies that they collectively contribute significantly to the predictive ability of the model for heart-related conditions in the CHSLB heart dataset.

The BO-XGBoost model, tailored for heart attack prediction and depicted in Figure 3, exhibits promising outcomes across both datasets. The comprehensive evaluation presented in Table 5 indicates that all employed classification metrics, particularly the AUC measure highlighted in Figure 3, consistently achieved 100% accuracy. This robust performance is particularly noteworthy for its accurate prediction of both True Positive and True Negative cases.

The exceptional level of accuracy and precision suggests the model’s potential benefits in real-life scenarios, where

early and accurate prediction of heart attacks is critical for timely intervention and patient well-being. The model’s ability to achieve perfect classification across various metrics underscores its reliability. It indicates its potential utility as an effective tool for proactive heart attack risk assessment and prevention in clinical settings.

### B. BREAST CANCER

The second experiment centered on predicting breast cancer using two datasets: WDBC with 32 features and WBCD with 10 features.

In the context of the WDBC dataset, the feature selection process identified only 18 out of the 32 features to be utilized in the modeling phase. This allows the prioritization of influential features while potentially excluding less informative ones. Table 8 displays the SVM weights and RFE ranks of selected features from the WDBC dataset, presumably related to breast cancer prediction given the features’ names. Features such as Mean Concavity (SVM weight: 0.8301, RFE rank: 1) exhibit substantial importance, with a high SVM weight and a top RFE rank, emphasizing their crucial roles in the model’s decision-making process. These features maintain significance even after recursive feature elimination. Conversely, features like Worst (Largest) Concavity (SVM weight: 1.7257, RFE rank: 6) showcase high SVM weights but relatively low RFE ranks, indicating the low importance of the feature, but still contributing to the overall model.

This nuanced evaluation underscores the varied influence of different features on the model’s predictive accuracy, providing valuable insights into the intricacies of the WDBC dataset. Emphasizing the significance of considering both SVM weights and RFE ranks enhances our understanding of the dataset’s complexities. It is noteworthy to highlight that the feature selection process has revealed that removing the remaining 14 discarded features will not have a detrimental effect on the predictive accuracy of the model.

**TABLE 8. SVM weights and RFE ranks of selected features of the WDBC dataset.**

Feature	SVM WEIGHT (RFE RANK)
Mean Radius	1.6656, 1
Mean Texture	0.1065, 1
Mean Perimeter	0.2069, 1
Mean Smoothness	0.2692, 8
Mean Compactness	0.2129, 1
Mean Concavity	0.8301, 1
Mean Concave Points	0.4128, 1
Mean Symmetry	0.3227, 1
Standard Error of Texture	1.2874, 1
Standard Error of Perimeter	0.3003, 5
Standard Error of Concavity	0.1838, 7
Worst (Largest) Radius	0.9137, 1
Worst (Largest) Texture	0.2627, 1
Worst (Largest) Smoothness	0.5336, 2
Worst (Largest) Compactness	0.2778, 3
Worst (Largest) Concavity	1.7257, 6
Worst (Largest) Concave Points	0.6804, 1
Worst (Largest) Symmetry	0.5080, 4

**TABLE 9. SVM weights and RFE ranks of selected features of the breast cancer wisconsin dataset.**

Feature	SVM WEIGHT (RFE RANK)
Clump Thickness	0.3392, 1
Uniformity of Cell Shape	0.1021, 2
Marginal Adhesion	0.1554, 1
Epithelial Cell Size	0.0208, 1
Bare Nucleoli	0.2581, 1
Bland Chromatin	0.2939, 1
Normal Nucleoli	0.0943, 1
Mitoses (Number of Mitotic Figures)	0.2537, 1
Clump Thickness	0.3392, 1
Uniformity of Cell Shape	0.1021, 2

In the breast cancer prediction experiments depicted in Figure 4, the implemented BO–XGBoost model demonstrated an impressively high level of accuracy, consistently reaching 99% for both datasets. The AUC value, a critical measure of model performance, achieved a perfect score of 1, further affirming the robustness of the model. Although the model encountered a misprediction in one positive case across the two datasets, this discrepancy was effectively captured by the recall values detailed in Table 5, which were 97% and 98% for the WDBC dataset and WBCD dataset, respectively. These findings indicate the remarkable ability of the model to effectively identify and classify breast cancer instances with high accuracy and sensitivity. Such precise predictions have substantial real-life implications, as they can

significantly contribute to early detection and intervention, enhance patient outcomes, and aid in the development of targeted treatment strategies.

Table 9 provides insights into the SVM weights and RFE ranks of selected features from the Breast Cancer Wisconsin dataset (WBCD). These features, encapsulating diverse cell characteristics, are pivotal contributors to the model’s predictive power. Notably, all selected features hold top positions with RFE ranks of 1 or 2, reinforcing their collective significance in shaping the model’s predictions. This thorough evaluation underscores the influential role of these factors in enhancing the overall predictive capability of the model. It is important to highlight that the prediction model is built using the 10 original features identified through this comprehensive analysis.

**C. DIABETES**

This experiment investigated diabetes prediction, utilizing two datasets: Sylhet with 17 features and Pima with nine features. For the first dataset, 11 features have been identified. The results of the feature selection process for the Sylhet dataset are shown in Table 10, which sheds light on the statistical significance of these features in predicting diabetes. The selected features in the Sylhet diabetes dataset reflect symptoms and conditions commonly associated with diabetes. The SVM weights and RFE ranks in the Sylhet Diabetes dataset highlight the individual contributions of selected features to the predictive model. Polyuria and Polydipsia exemplify crucial factors with high SVM weights (1.6662 and 1.9996, respectively) and top RFE ranks (1). These features stand out as pivotal contributors, emphasizing their significance in diabetes prediction. On the other hand, the Gender feature, with an SVM weight of 1.9997, highlights the nuanced impact of features, as it possesses a high weight, but a lower RFE rank of 5. This suggests that, while the Gender feature plays a role, other features take precedence during the feature elimination process. Additionally, Alopecia, with a low SVM weight (0.3337) but a top RFE rank (1), signifies its importance despite its lower impact on the overall model. This comprehensive evaluation provides valuable insights into the varying importance of features and the differential influence of RFE on their rankings.

The SVM weights and RFE ranks for the selected features from the Pima Diabetes dataset, as presented in Table 11, offer valuable insights into their contributions to the predictive model. It shows that the eight features presented in the dataset were selected to get the best performance. The Pima diabetes dataset’s selected features cover a range of health indicators, including reproductive history, glucose levels, blood pressure, BMI, diabetes pedigree function, and age. Several features, including Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age, exhibit both high SVM weights (ranging from 0.0004 to 0.4386) and top RFE ranks (all ranked 1). These features play crucial roles in shaping the predictive model, underscoring their significance in diabetes prediction.



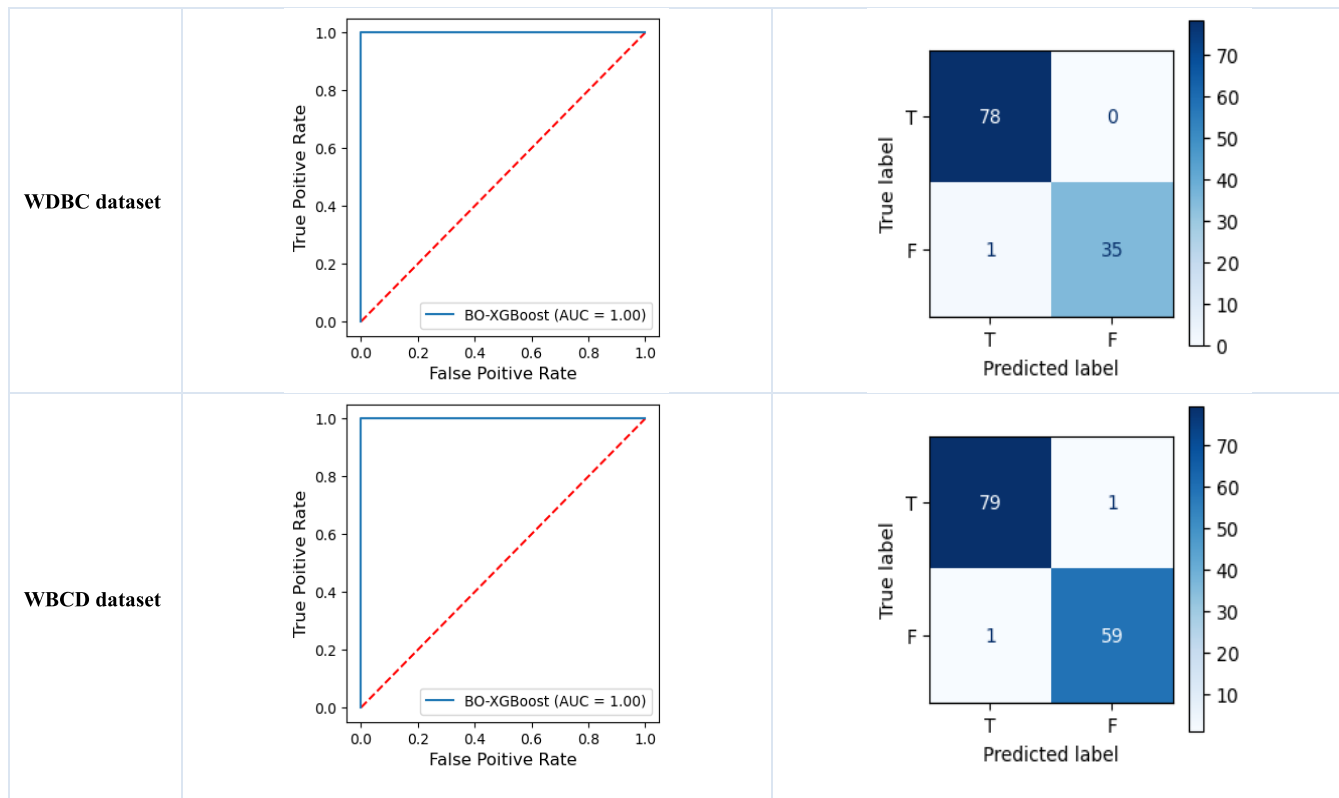


FIGURE 4. Prediction results: Breast cancer WDBC dataset and breast cancer WBCD dataset.

TABLE 10. SVM weights and RFE ranks of selected features of the sylhet diabetes dataset.

Feature	SVM WEIGHT (RFE RANK)
Gender	1.9997, 5
Polyuria	1.6662, 1
Polydipsia	1.9996, 1
Polyphagia	0.6666, 1
Genital thrush	0.3340, 4
Itching	1.3332, 3
Irritability	1.6661, 1
delayed healing	0.3332, 1
Partial paresis	0.6667, 2
Muscle stiffness	0.6668, 1
Alopecia	0.3337, 1

While individual SVM weights are low, the recursive feature elimination process identifies each feature as important for predicting diabetes.

The consistently top RFE ranks indicate that, after recursive feature elimination, these features maintain their importance and contribute significantly to the overall model accuracy. The model appears to rely on a combination of factors rather than heavily emphasizing a single feature.

In the experiments focusing on diabetes prediction, the BO- XGBoost model underwent evaluation using two distinct datasets. The Sylhet dataset exhibited exceptional accuracy, achieving a perfect score of 100% across all classification measures, including the Area Under the Curve

TABLE 11. SVM weights and RFE ranks of selected features of the Pima diabetes dataset.

Feature	SVM WEIGHT (RFE RANK)
Pregnancies	0.0952, 1
Glucose	0.0303, 1
Blood Pressure	0.0088, 1
Skin Thickness	0.0018, 1
Insulin	0.0004, 1
BMI	0.0653, 1
Diabetes Pedigree Function	0.4386, 1
Age	0.0018, 1

(AUC). Conversely, in the case of the Pima dataset, the predictive performance of the model fell short of reaching 90% across all measures, as detailed in Table 5. The associated confusion matrix in Figure 5 reveals that out of the 106 positive cases, 17 were misclassified as negative, and 13 negative cases out of 48 were erroneously classified as positive. In the context of medical diagnosis, the significance of recall values is evident, as a lower recall may impact the early identification and treatment of potential patients requiring immediate attention. Despite this, it is noteworthy that the results obtained for the Pima dataset remain competitive when benchmarked against other models, showcasing the model’s competence in the challenging task of diabetes prediction. The imbalanced nature of the dataset, with 65% negative and 35% positive cases, poses challenges for

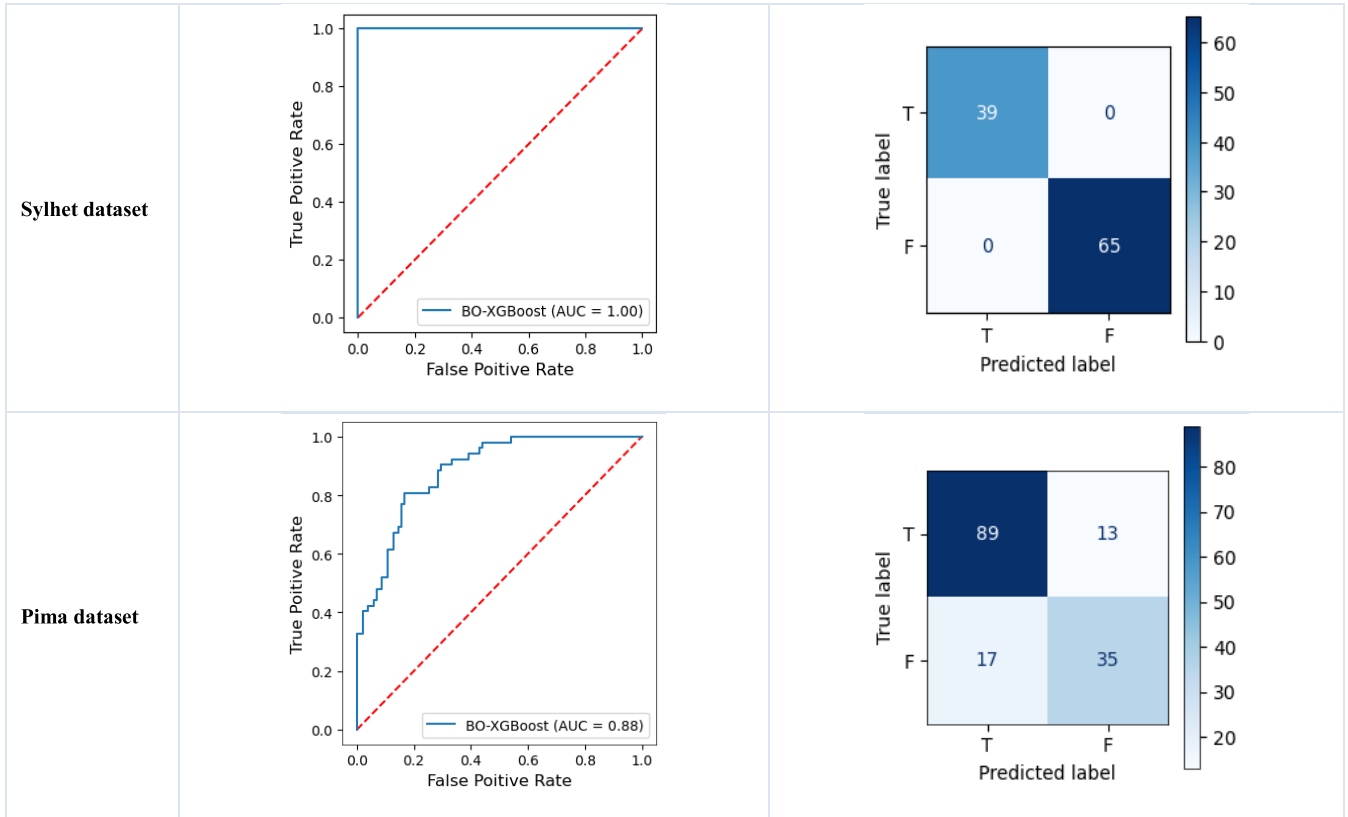


FIGURE 5. Prediction results: Diabetes-Sylhet dataset and diabetes- Pima dataset.

diabetes prediction, potentially leading to a bias toward the majority class.

**D. KIDNEY DISEASE**

In the final set of experiments, attention shifted to kidney disease prediction, involving two datasets: Abu Dhabi with 22 features and India with 25 features. The feature selection using the SVM-RFE method provides insights into the features’ statistical significance in predicting kidney-related outcomes. Table 12 presents the SVM weights and RFE ranks of selected features from the Abu Dhabi Kidney dataset. The process identified 14 features to be used in the modeling phase. It is noteworthy that eight features have been eliminated, highlighting the model’s focus on retaining the most informative variables. The selected features include a combination of demographic information, medical history, medication history, and baseline measures. Features such as Gender, History of Coronary Heart Disease (CHD), History of Vascular, History of Smoking, History of HTN, History of Obesity, Cholesterol Baseline, eGFR Baseline, and TIME\_YEAR demonstrate substantial SVM weights (ranging from 0.0290 to 0.4589) and consistently maintain RFE rank 1, underscoring their pivotal roles in kidney disease prediction. These features remain influential even after recursive feature elimination, emphasizing their enduring importance. On the other hand, features like History Diabetes, DLDmeds,

DMmeds, HTNmeds, and ACEIARB exhibit notable SVM weights (ranging from 0.1041 to 0.7500) but have higher RFE ranks (2-3), suggesting their importance diminishes after feature elimination. This nuanced evaluation sheds light on the relative contributions of individual features in predicting kidney disease, considering both their SVM weights and RFE ranks.

Table 13 provides information on the SVM weights and RFE ranks of selected features from the India Kidney dataset. Notably, only 12 features out of 25 were identified through the SVM-RFE method. The 12 identified features from the India Kidney dataset encompass a range of medical indicators related to urine analysis, blood parameters, and medical conditions. Notably, features such as Sugar, Red Blood Cells, Pus Cell, Serum Creatinine, Potassium, and Anemia exhibit substantial SVM weights (ranging from 0.3136 to 1.7706) and consistently maintain top RFE ranks (1), emphasizing their crucial roles in kidney disease prediction. These features retain their importance even after recursive feature elimination, highlighting their enduring impact on the model’s decision-making. On the other hand, Specific Gravity, Albumin, Red Blood Cell Count, Hypertension, and Diabetes Mellitus also display noteworthy SVM weights (ranging from 0.1752 to 0.8462) but low RFE ranks (3-7). These features collectively contribute to predicting kidney-related conditions. The comprehensive assessment of these features

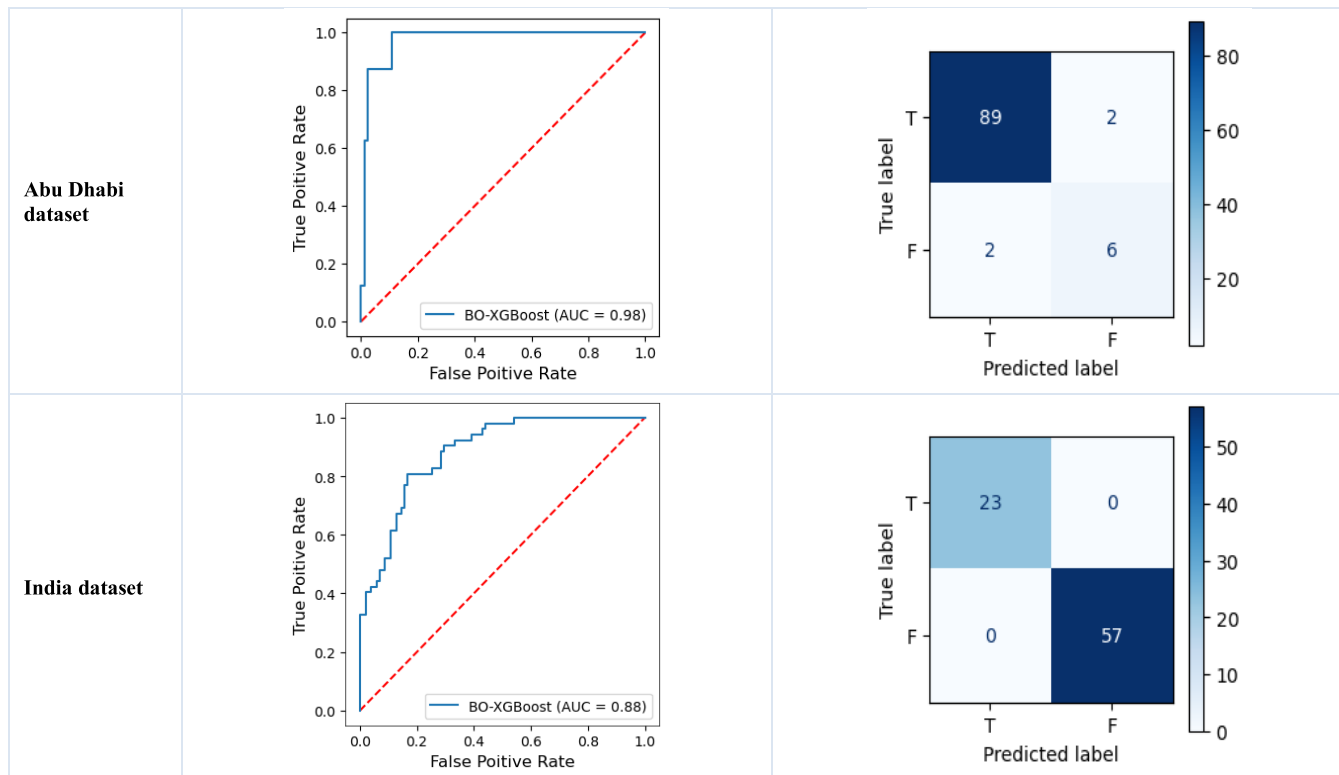


FIGURE 6. Prediction results: kidney-Abu Dhabi dataset and kidney-india dataset.

TABLE 12. SVM weights and RFE ranks of selected features of the abu dhabi kidney dataset.

Feature	SVM WEIGHT (RFE RANK)
Gender	0.4589, 1
History of Diabetes	0.7500, 2
History of CHD	0.3324, 1
History of Vascular	0.3811, 1
History of Smoking	0.1954, 1
History of Hypertension (HTN)	0.1869, 1
History of Obesity	0.0290, 1
DLDmeds	0.1282, 3
DMmeds	0.3053, 1
HTNmeds	0.5004, 1
ACEIARB	0.2657, 1
Cholesterol Baseline	0.1041, 1
eGFR Baseline	0.0381, 1
TIME YEAR	0.2257, 1

provides valuable insights into the complex relationships between various factors and kidney-related conditions, contributing to a deeper understanding of the India Kidney dataset.

High SVM weights for Red Blood Cells, Pus Cell, Serum Creatinine, Potassium, and Sugar emphasize their significance in the model. Albumin and Diabetes Mellitus have lower RFE ranks but still substantial SVM weights. Their inclusion suggests that while not the top predictors, they contribute significantly to the predictive power of the model.

In the chronic disease experiment outlined earlier, two datasets were employed to evaluate the performance of

TABLE 13. SVM weights and RFE ranks of selected features of the india kidney dataset.

Feature	SVM WEIGHT (RFE RANK)
Specific Gravity	0.6658, 5
Albumin	0.8029, 2
Sugar	0.5433, 1
Red Blood Cells	1.7706, 1
Pus Cell	0.3136, 1
Serum Creatinine	0.7569, 1
Potassium	0.1823, 1
Red Blood Cell Count	0.1752, 6
Hypertension	0.8462, 7
Diabetes Mellitus	0.8462, 3
Pedal Edema	0.7436, 4
Anemia	0.3860, 1

BO-XGBoost models in predicting kidney chronic disease. Although the model achieved an accuracy of 96% on the Abu Dhabi dataset, the recall value was notably lower, at approximately 62%.

The difference in measures can be attributed to the imbalanced nature of the dataset, exemplified by the confusion matrix in Figure 6. The testing set, in particular, included eight negative cases and 91 positive cases, contributing to the observed differences. However, despite this imbalance, the proposed model achieves outstanding performance. For the second dataset, the BO-XGBoost model demonstrated excellent performance, achieving 100% accuracy across all measures. These accurate predictions hold signifi-

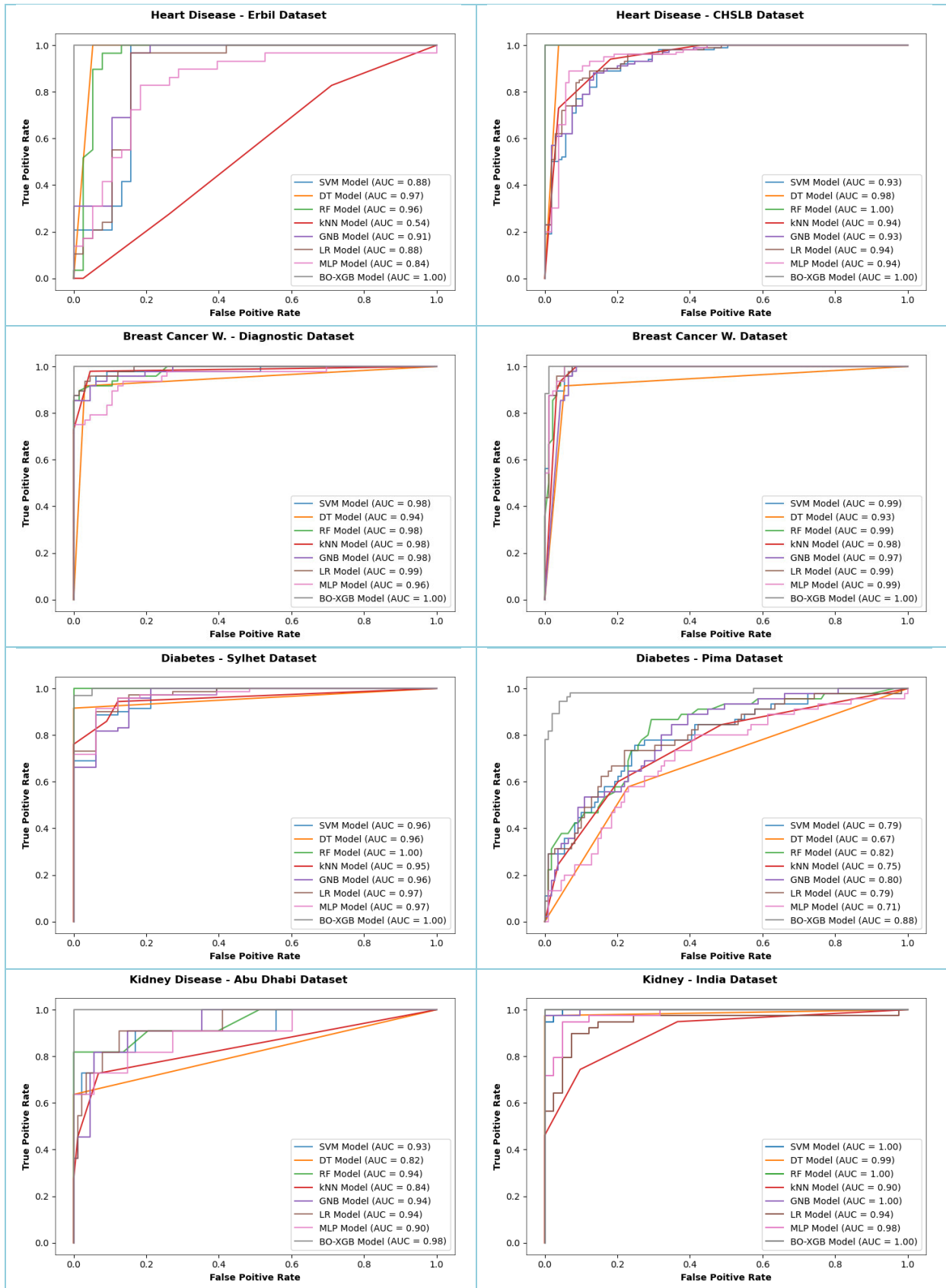


FIGURE 7. Comparison results of different ML methods against BO-XGBoost.



cant real-life implications, contributing to the early detection, intervention, and development of targeted treatment strategies, ultimately enhancing patient outcomes.

### E. COMPARATIVE ANALYSIS

Among a diverse ensemble of machine learning algorithms, including established methods like SVM, DT, RF, K-NN, GNB, LR, and MLP, the hybrid BO–XGBoost model reigned supreme. Our comprehensive evaluation, crafted using the AUC metric for classification accuracy, painted a definitive picture: the BO–XGBoost model stood head and shoulders above its competitors. Its consistently superior AUC values show the exceptional predictive power of BO–XGBoost in predicting chronic diseases like heart attack, breast cancer, diabetes, and kidney disease. Figure 7 demonstrates the performance results of the applied ML methods.

### VI. CONCLUSION AND FUTURE WORK

The early detection of chronic diseases remains a formidable obstacle for researchers, driving the exploration of various AI techniques for analyzing medical data and predicting disease onset. The findings reported in this study demonstrated the effectiveness of the proposed approach for predicting chronic diseases (i.e., heart attack, breast cancer, diabetes, and kidney disease). The integration of XGBoost, and Bayesian optimization contributes to a robust and efficient model for the accurate identification of individuals at risk, paving the way for precision healthcare and proactive disease management.

In addition, the proposed approach incorporates feature selection methods to refine the analysis. The SVM-RFE was used to identify and prioritize relevant features within datasets, discarding irrelevant ones, and assigning scores to those with the most significant predictive power. This targeted approach ensures the model focuses on the most impactful information, enhancing its effectiveness.

This study demonstrated the remarkable performance of the proposed hybrid model, combining Bayesian Optimization with XGBoost, in comparison to a range of well-established ML models. A thorough comparative analysis included SVM, DT, RF, KNN, GNB, LR, and MLP. The evaluation encompassed two datasets for each disease, ensuring a comprehensive assessment of the BO–XGBoost model's predictive capabilities. The outcomes underscore the adaptability of the proposed hybrid model across diverse datasets and various disease types. Through this extensive comparison, our study not only showcases the impressive performance of the BO–XGBoost model but also positions it favorably in the context of existing ML models for chronic disease prediction.

The model's effectiveness in handling multiple chronic diseases underscores its potential for real-world application in healthcare settings. Its ability to leverage diverse datasets and prioritize high-impact features suggests it could be readily integrated into medical practice, empowering healthcare professionals with valuable prediction tools for early disease detection and intervention.

While this study paved the way for significant advancements in chronic disease prediction, it also identifies research gaps that need further exploration. The development of universally applicable prediction systems remains a challenge, and the study acknowledges the potential discrepancies between research findings and real-world clinical use due to data limitations and variations in medical datasets. Addressing these challenges will involve continuous research and collaboration to refine and validate AI-powered prediction models for robust and widespread implementation in healthcare systems, ultimately benefiting both patients and healthcare professionals.

### ACKNOWLEDGMENT

The authors would like to thank the support provided by the King Fahd University of Petroleum and Minerals (KFUPM) and also would like to thank the expert physicians with the KFUPM Medical Center for their invaluable insights into the patient data analyzed in this study and their guidance on result interpretation.

### CONFLICTS OF INTEREST

The author declares that there are no conflicts of interest.

### REFERENCES

- [1] W. H. O. Diet, "Chronic diseases," World Health Org., Geneva, Switzerland, 2003.
- [2] R. Sawhney, A. Malik, S. Sharma, and V. Narayan, "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease," *Decis. Anal. J.*, vol. 6, Mar. 2023, Art. no. 100169.
- [3] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.
- [4] U. Ullah and B. Garcia-Zapirain, "Quantum machine learning revolution in healthcare: A systematic review of emerging perspectives and applications," *IEEE Access*, vol. 12, pp. 11423–11450, 2024.
- [5] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl.*, Apr. 2008, pp. 108–115.
- [6] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.
- [7] L. Di Biasi, F. De Marco, A. A. Citarella, M. Castrillón-Santana, P. Barra, and G. Tortora, "Refactoring and performance analysis of the main CNN architectures: Using false negative rate minimization to solve the clinical images melanoma detection problem," *BMC Bioinf.*, vol. 24, no. 1, p. 386, Oct. 2023.
- [8] O. Khan, J. H. Badhiwala, G. Grasso, and M. G. Fehlings, "Use of machine learning and artificial intelligence to drive personalized medicine approaches for spine care," *World Neurosurgery*, vol. 140, pp. 512–518, Aug. 2020.
- [9] A. Bohr and K. Memarzadeh, *Artificial Intelligence in Healthcare*. New York, NY, USA: Academic, 2020.
- [10] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, pp. S36–S40, Jan. 2017.
- [11] S. Agarwal and D. M. G. C. Prabha, "Chronic diseases prediction using machine learning—A review," *Ann. Rom. Soc. Cell Biol.*, vol. 25, no. 1, pp. 3495–3511, 2021.
- [12] L. Yang, L. Han, Z. Chen, J. Zhou, and J. Wang, "Growing trend of China's contribution to haze research," *Scientometrics*, vol. 105, no. 1, pp. 525–535, Oct. 2015.
- [13] K. Panagiotopoulos, A. Korfiati, K. Theofilatos, P. Hurwitz, M. A. Deriu, and S. Mavroudi, "MEVA-X: A hybrid multiobjective evolutionary tool using an XGBoost classifier for biomarkers discovery on biomedical datasets," *Bioinformatics*, vol. 39, no. 7, Jul. 2023, Art. no. btad384.

- [14] A. Maleki, M. Raahemi, and H. Nasiri, "Breast cancer diagnosis from histopathology images using deep neural network and XGBoost," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105152.
- [15] M. J. Raihan, M. A.-M. Khan, S.-H. Kee, and A.-A. Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP," *Sci. Rep.*, vol. 13, no. 1, p. 6263, Apr. 2023.
- [16] X. Xiong, X. Guo, P. Zeng, R. Zou, and X. Wang, "A short-term wind power forecast method via XGBoost hyper-parameters optimization," *Frontiers Energy Res.*, vol. 10, May 2022, Art. no. 905155.
- [17] P. I. Frazier, "Bayesian optimization," in *Recent Advances in Optimization and Modeling of Contemporary Problems*, Inform. MD, USA, 2018, pp. 255–278.
- [18] Y. Yang, K. Wang, Z. Yuan, and D. Liu, "Predicting freeway traffic crash severity using XGBoost-Bayesian network model with consideration of features interaction," *J. Adv. Transp.*, vol. 2022, pp. 1–16, Apr. 2022.
- [19] X.-W. Chen and J. C. Jeong, "Enhanced recursive feature elimination," in *Proc. 6th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2007, pp. 429–435.
- [20] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Aug. 1998.
- [21] I. Steinwart and A. Christmann, *Support Vector Machines*. New York, NY, USA: Springer, 2008.
- [22] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using Machine learning algorithms," *Mater. Today Proc.*, vol. 80, pp. 3682–3685, Jan. 2023.
- [23] S. Hegde and M. R. Mundada, "Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence based feature selection approach," *Int. J. Pervasive Comput. Commun.*, vol. 17, no. 1, pp. 20–36, Feb. 2021.
- [24] S. Sandhiya and U. Palani, "An effective disease prediction system using incremental feature selection and temporal convolutional neural network," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 11, pp. 5547–5560, Nov. 2020.
- [25] M. Z. Alam, M. S. Rahman, and M. S. Rahman, "A random forest based predictor for medical data classification using feature ranking," *Informat. Med. Unlocked*, vol. 15, 2019, Art. no. 100180.
- [26] E. Maini, B. Venkateswarlu, D. Marwaha, and B. Maini, "Upgrading the performance of machine learning based chronic disease prediction systems using stacked generalization technique," *Int. J. Comput. Digit. Syst.*, vol. 10, no. 1, pp. 1031–1039, Nov. 2021.
- [27] T. H. H. Aldhiani, A. S. Alshebami, and M. Y. Alzahrani, "Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms," *J. Healthcare Eng.*, vol. 2020, pp. 1–16, Mar. 2020.
- [28] A. Sinha, B. Sahoo, S. S. Rautaray, and M. Pandey, "Improved framework for breast cancer prediction using frequent itemsets mining for attributes filtering," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS)*, May 2019, pp. 979–982.
- [29] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast cancer prediction: A comparative study using machine learning techniques," *Social Netw. Comput. Sci.*, vol. 1, no. 5, pp. 1–14, Sep. 2020.
- [30] S. Sharmin, T. Ahammad, M. Alamin Talukder, and P. Ghose, "A hybrid dependable deep feature extraction and ensemble-based machine learning approach for breast cancer detection," *IEEE Access*, vol. 11, pp. 87694–87708, 2023.
- [31] P. Whig, K. Gupta, N. Jiwani, H. Jupalle, S. Kouser, and N. Alam, "A novel method for diabetes classification and prediction with pycaret," *Microsyst. Technol.*, vol. 29, no. 10, pp. 1479–1487, Oct. 2023.
- [32] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023.
- [33] K. H. Reddy and G. Saranya, "Prediction of cardiovascular diseases in diabetic patients using machine learning techniques," in *Artificial Intelligence Techniques for Advanced Computing Applications*. Cham, Switzerland: Springer, 2021, pp. 299–305.
- [34] S. Gündoğdu, "Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique," *Multimedia Tools Appl.*, vol. 82, no. 22, pp. 34163–34181, Sep. 2023.
- [35] M. Abood Kadhim and A. M. Radhi, "Heart disease classification using optimized machine learning algorithms," *Iraqi J. Comput. Sci. Math.*, vol. 4, no. 2, pp. 31–42, Feb. 2023.
- [36] A. Gazali, K. Debasis, and R. M. Sahoo, "A novel system based on artificial neural network for heart disease classification," in *Proc. 3rd Int. Conf. Artif. Intell. Signal Process. (AISP)*, Mar. 2023, pp. 1–5.
- [37] H. A. Al-Jamimi, "Intelligent methods for early prediction of heart disease," in *Proc. 9th Int. Conf. Control, Decis. Inf. Technol. (CoDIT)*, Jul. 2023, pp. 2574–2578.
- [38] N. Absar, E. K. Das, S. N. Shoma, M. U. Khandaker, M. H. Miraz, M. R. I. Faruque, N. Tamam, A. Sulieman, and R. K. Pathan, "The efficacy of machine-learning-supported smart system for heart disease prediction," *Healthcare*, vol. 10, no. 6, p. 1137, Jun. 2022.
- [39] O. A. Jongbo, A. O. Adetunmbi, R. B. Ogunrinde, and B. Badeji-Ajisafe, "Development of an ensemble approach to chronic kidney disease diagnosis," *Sci. Afr.*, vol. 8, Jul. 2020, Art. no. e00456.
- [40] E. Listiana, R. Muzayanah, M. A. Muslim, and E. Sugiharti, "Optimization of support vector machine using information gain and AdaBoost to improve accuracy of chronic kidney disease diagnosis," *J. Soft Comput. Explor.*, vol. 4, no. 3, pp. 152–158, 2023.
- [41] A. S. Hannan and P. Pal, "Detection and classification of kidney disease using convolutional neural networks," *J. Neurol. Neurorehab. Res.*, vol. 8, p. 136, Jan. 2023.
- [42] D. Chicco, C. A. Lovejoy, and L. Oneto, "A machine learning analysis of health records of patients with chronic kidney disease at risk of cardiovascular disease," *IEEE Access*, vol. 9, pp. 165132–165144, 2021.
- [43] P. A. Moreno-Sánchez, "Data-driven early diagnosis of chronic kidney disease: Development and evaluation of an explainable AI model," *IEEE Access*, vol. 11, pp. 38359–38369, 2023.
- [44] W. You, Z. Yang, and G. Ji, "Feature selection for high-dimensional multi-category data using PLS-based local recursive feature elimination," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1463–1475, Mar. 2014.
- [45] K. Duan, S. S. Keerthi, and A. N. Poo, "Evaluation of simple performance measures for tuning SVM hyperparameters," *Neurocomputing*, vol. 51, pp. 41–59, Apr. 2003.
- [46] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
- [47] Y. Xu, X. Zhao, Y. Chen, and Z. Yang, "Research on a mixed gas classification algorithm based on extreme random tree," *Appl. Sci.*, vol. 9, no. 9, p. 1728, Apr. 2019.
- [48] H. Chen, "Enterprise marketing strategy using big data mining technology combined with XGBoost model in the new economic era," *PLoS ONE*, vol. 18, no. 6, Jun. 2023, Art. no. e0285506.
- [49] V. Jain and M. Agrawal, "Heart failure prediction using XGB classifier, logistic regression and support vector classifier," in *Proc. Int. Conf. Advancement Comput. Technol. (InCACCT)*, May 2023, pp. 1–5.
- [50] Y. Kong, Y. Wang, S. Sun, and J. Wang, "XGB and SHAP credit scoring model based on Bayesian optimization," *J. Comput. Electron. Inf. Manage.*, vol. 10, no. 1, pp. 46–53, Feb. 2023.
- [51] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *J. Electron. Sci. Technol.*, vol. 17, pp. 26–40, Mar. 2019.
- [52] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.
- [53] B. Betrò, "Bayesian methods in global optimization," *J. Global Optim.*, vol. 1, no. 1, pp. 1–14, 1991.
- [54] H. A. Al-Jamimi, G. M. BinMakhashen, and T. A. Saleh, "Artificial intelligence approach for modeling petroleum refinery catalytic desulfurization process," *Neural Comput. Appl.*, vol. 34, no. 20, pp. 17809–17820, Oct. 2022.
- [55] N. Basha, G. Ibrahim, H. A. Choudhury, M. S. Challiwala, R. Fezai, B. Malluhi, H. Nounou, N. Elbashir, and M. Nounou, "Bayesian-optimized neural networks and their application to model gas-to-liquid plants," *Gas Sci. Eng.*, vol. 113, May 2023, Art. no. 204964.
- [56] E. C. Garrido-Merchán, D. Fernández-Sánchez, and D. Hernández-Lobato, "Parallel predictive entropy search for multi-objective Bayesian optimization with constraints applied to the tuning of machine learning algorithms," *Expert Syst. Appl.*, vol. 215, Apr. 2023, Art. no. 119328.
- [57] T. Yu and H. Zhu, "Hyper-parameter optimization: A review of algorithms and applications," 2020, *arXiv:2003.05689*.
- [58] H. Albrahim and S. A. Ludwig, "Hyperparameter optimization: Comparing genetic algorithm against grid search and Bayesian optimization," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2021, pp. 1551–1559.

- [59] P. R. Lorenzo, J. Nalepa, M. Kawulok, L. S. Ramos, and J. R. Pastor, "Particle swarm optimization for hyper-parameter selection in deep neural networks," in *Proc. Genetic Evol. Comput. Conf.*, Jul. 2017, pp. 481–488.
- [60] A. Gülcü and Z. Kuş, "Multi-objective simulated annealing for hyper-parameter optimization in convolutional neural networks," *PeerJ Comput. Sci.*, vol. 7, p. e338, Jan. 2021.
- [61] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, and A. G. Doyle, "Bayesian reaction optimization as a tool for chemical synthesis," *Nature*, vol. 590, no. 7844, pp. 89–96, Feb. 2021.
- [62] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.
- [63] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proc. IEEE*, vol. 104, no. 1, pp. 148–175, Jan. 2016.
- [64] J. Chen, F. Zhao, Y. Sun, and Y. Yin, "Improved XGBoost model based on genetic algorithm," *Int. J. Comput. Appl. Technol.*, vol. 62, no. 3, p. 240, 2020.
- [65] D. S. W. Ting, C. Y. Cheung, Q. Nguyen, C. Sabanayagam, G. Lim, Z. W. Lim, G. S. W. Tan, Y. Q. Soh, L. Schmetterer, Y. X. Wang, J. B. Jonas, R. Varma, M. L. Lee, W. Hsu, E. Lamoureux, C.-Y. Cheng, and T. Y. Wong, "Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy: A multi-ethnic study," *NPJ Digit. Med.*, vol. 2, no. 1, p. 24, Apr. 2019.
- [66] G. Liew, T. Tsang, B. Marshall, M. Saw, L. M. Khachigian, S. Ong, I.-V. Ho, and V. Wong, "Proportion of people with diabetic retinopathy and macular oedema varies by ethnicity in a tertiary retinal clinic in australia: Findings from the Liverpool eye and diabetes study (LEADS)," *BMJ Open*, vol. 13, no. 2, Feb. 2023, Art. no. e055404.
- [67] J. Konttila and K. Väyrynen, "Challenges of current regulation of AI-based healthcare technology (AIHT) and potential consequences of the European AI act proposal," in *Proc. 13th Scand. Conf. Inf. Syst.*, 2022, p. 7.
- [68] *Erbil Heart Disease Dataset*. Accessed: Nov. 2023. [Online]. Available: <https://www.kaggle.com/datasets/hangawqadir/erbil-heart-disease-dataset>
- [69] *Heart Disease Dataset*. Accessed: Nov. 2023. [Online]. Available: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?select=heart.csv>
- [70] *Breast Cancer Wisconsin (Diagnostic) Data Set*. Accessed: Nov. 2023. [Online]. Available: <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>
- [71] *Wisconsin Breast Cancer Database*. Accessed: Nov. 2023. [Online]. Available: <https://www.kaggle.com/datasets/roustekbio/breast-cancer-csv>
- [72] *Early Stage Diabetes Risk Prediction Dataset*. Accessed: Nov. 2023. [Online]. Available: <https://archive.ics.uci.edu/dataset/529/early>
- [73] *Pima Indians Diabetes*. Accessed: Nov. 2023. [Online]. Available: <https://data.world/uci/pima-indians-diabetes>
- [74] S. Al-Shamsi, D. Regmi, and R. D. Govender, "Chronic kidney disease in patients at high risk of cardiovascular disease in the united Arab emirates: A population-based study," *PLoS ONE*, vol. 13, no. 6, Jun. 2018, Art. no. e0199920.
- [75] L. O. D. Chicco and C. A. Lovejoy. (2021). *Chronic Kidney Disease EHRs Abu Dhabi*. [Online]. Available: <https://www.kaggle.com/datasets/davidechicco/chronic-kidney-disease-ehrs-abu-dhabi>
- [76] *Kidney Disease Dataset*. Accessed: Nov. 2023. [Online]. Available: <https://www.kaggle.com/datasets/akshayksingh/kidney-disease-dataset>
- [77] D. G. T. Arts, "Defining and improving data quality in medical registries: A literature review, case study, and generic framework," *J. Amer. Med. Inform. Assoc.*, vol. 9, no. 6, pp. 600–611, Nov. 2002.
- [78] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, vol. 72. New York, NY, USA: Springer, 2015.
- [79] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Comput. Methods Programs Biomed.*, vol. 220, Jun. 2022, Art. no. 106773.
- [80] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *IEEE Access*, vol. 7, pp. 107964–108000, 2019.
- [81] L. M. Rudner, "Expected classification accuracy," *Pract. Assessment, Res. Eval.*, vol. 10, no. 1, p. 13, 2005.
- [82] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation," in *Proc. Australas. Joint Conf. Artif. Intell.* Hobart, TS, Australia: Springer, Dec. 2006, pp. 1015–1021, 2006.
- [83] A. Auriemma Citarella, L. Di Biasi, F. De Marco, and G. Tortora, "ENTAIL: Yet another amyloid fibrils cClassifier," *BMC Bioinf.*, vol. 23, no. 1, pp. 1–15, Dec. 2022.
- [84] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *Int. J. Mach. Learn. Comput.*, vol. 3, no. 2, pp. 224–228, 2013.



**HAMDI A. AL-JAMIMI** received the Ph.D. degree in computer science and engineering, in 2015. He is currently a Distinguished Scientist and an Academician with the King Fahd University of Petroleum & Minerals (KFUPM), Saudi Arabia. He has since built an impressive publication portfolio in renowned journals and premier conferences. Specializing in the intersection of artificial intelligence and data science, he explores cutting-edge applications and methodologies, with a particular emphasis on the healthcare, petrochemicals, and engineering sectors. His research interests include advancing knowledge and fostering technological innovation in these crucial domains. Committed to collaborative research, he actively cultivates interdisciplinary partnerships that yield meaningful societal impact. His unwavering dedication to excellence and insatiable quest for knowledge underscore his substantial contributions to the scientific community and broader academic landscape.

...