Analysis of Cancer-Associated Mutations of POLB Using Machine Learning and Bioinformatics

Razan Alkhanbouli, Amira Al-Aamri, Maher Maalouf, Kamal Taha, Andreas Henschel, and Dirar Homouz

***Abstract*—DNA damage is a critical factor in the onset and progression of cancer. When DNA is damaged, the number of genetic mutations increases, making it necessary to activate DNA repair mechanisms. A crucial factor in the base excision repair process, which helps maintain the stability of the genome, is an enzyme called DNA polymerase β (Polβ) encoded by the POLB gene. It plays a vital role in the repair of damaged DNA. Additionally, variations known as Single Nucleotide Polymorphisms (SNPs) in the POLB gene can potentially affect the ability to repair DNA. This study uses bioinformatics tools that extract important features from SNPs to construct a feature matrix, which is then used in combination with machine learning algorithms to predict the likelihood of developing cancer associated with a specific mutation. Eight different machine learning algorithms were used to investigate the relationship between POLB gene variations and their potential role in cancer onset. This study not only highlights the complex link between POLB gene SNPs and cancer, but also underscores the effectiveness of machine learning approaches in genomic studies, paving the way for advanced predictive models in genetic and cancer research.**

***Index Terms*—POLB, DNA damage repair, SNPs, bioinformatics, machine learning.**

I. Introduction

CANCER is a major global health concern that caused more than ten million deaths worldwide in 2020, as reported by the World Health Organization [1]. The development and progression of cancer are influenced by various factors, one of the key contributors being DNA damage caused by internal and external genotoxic agents. In the United States alone, approximately 2 million cancer cases were diagnosed in 2022, and global incidence rates are similarly high [2]. DNA damage can increase mutations and genomic instability, leading to cancer development [3]. To ensure DNA integrity, organisms have developed mechanisms to repair damaged DNA, and one such mechanism is called Base Excision Repair (BER), which plays a crucial role in removing damaged bases before they can cause mutations [4]. Within the BER pathway, the enzyme DNA polymerase β (Polβ) is essential for its proper functioning. Accurate prediction of cancer risk requires a comprehensive understanding of the molecular mechanisms involved in DNA repair. BER corrects DNA damage caused by oxidative and deamination events by identifying abnormal bases, removing the damaged base, and filling the resulting gap with the correct nucleotide, as shown in Fig. 1.

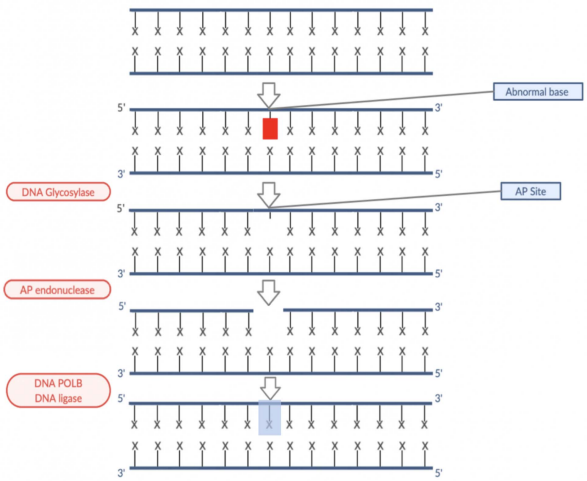


Fig. 1. This diagram illustrates the Base Excision Repair (BER) pathway, which repairs DNA damage from abnormal or damaged bases. The steps include removing the damaged base with DNA glycosylase, cleaving the AP site with AP endonuclease, filling it in with DNA polymerase β, and sealing it with DNA ligase for functional DNA.

Mutations in the POLB gene, responsible for encoding the Polβ enzyme, can have a detrimental effect on the DNA repair mechanism, which could lead to the development of cancer [5]. Several types of cancer, such as gastric, oral squamous, and colorectal cancers, have been associated with POLB gene mutations and overexpression [5]. When the DNA repair machinery is deficient, it generates DNA lesions and increases the likelihood of mutations. Human studies have been conducted to investigate the impact of POLB gene mutations on cancer development.

A study by Kiwerska and Szyfter [3] has shown that most tumors harbor mutated versions of the POLB gene, with an estimated occurrence rate of 30-40%. Fig. 2 illustrates the functional pathway of the POLB gene under normal conditions and the consequent pathway leading to cancer development when mutations

Manuscript received 20 July 2023; revised 19 April 2024; accepted 20 April 2024. Date of publication 1 May 2024; date of current version 9 October 2024. *(Corresponding authors: Maher Maalouf; Dirar Homouz.)*

Razan Alkhanbouli and Maher Maalouf are with the Management Science and Engineering, Khalifa University of Science, Technology, Abu Dhabi 127788, UAE (e-mail: maher.maalouf@ku.ac.ae).

Amira Al-Aamri is with the Center for Biotechnology, Khalifa University of Science, Technology, Abu Dhabi 127788, UAE.

Kamal Taha and Andreas Henschel are with the Department of Computer Science, Khalifa University of Science, Technology, Abu Dhabi 127788, UAE.

Dirar Homouz is with the Department of Physics, Khalifa University of Science, Technology, Abu Dhabi 127788, UAE (e-mail: dirar.homouz@ku.ac.ae).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCBB.2024.3395777, provided by the authors.

Digital Object Identifier 10.1109/TCBB.2024.3395777

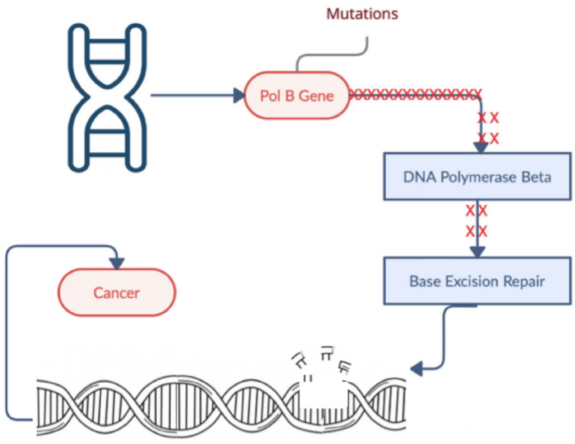


Fig. 2. The diagram illustrates the impact of mutations in the POLB gene on the DNA repair process and how this can lead to cancer. Normally, the POLB gene produces polymerase β, essential for base excision repair. Mutations in the POLB gene disrupt this function, resulting in accumulated DNA damage and an increased cancer risk.

occur. The figure highlights the crucial role of POLB in DNA repair and how its impairment through mutation can prevent the repair of damaged bases. When discussing genetic variations within a gene's DNA sequence, it is important to consider single nucleotide polymorphisms (SNPs), alterations that occur in a single base pair of the DNA sequence. In the human genome, scientists have identified approximately 4 to 5 million SNPs [3]. These variations can be inherited or arise spontaneously and are found in various genes, including those involved in DNA repair, such as POLB. It should be noted that SNPs have the potential to significantly affect the expression and activity of proteins encoded by these genes [3]. Consequently, this can profoundly affect the efficacy of DNA repair mechanisms such as base excision repair (BER). The specific location and functional impact of these SNPs play a crucial role in determining their effects. In some cases, SNPs can decrease the efficiency of DNA repair, accumulate genomic mutations, and increase the vulnerability of an individual to developing cancer [6].

The primary objective of this research is to investigate the use of bioinformatics tools to extract features to generate vectors for single nucleotide polymorphisms (SNP) and then to construct machine learning (ML) models capable of predicting the probability of the participation of a specific SNP in cancer development. By shedding light on the underlying mechanisms of cancer, this study can potentially improve our understanding of its progression. Furthermore, it could facilitate the development of more precise predictive models to assess cancer risk. The structure of the paper is as follows: Section II includes a comprehensive review of the literature, followed by the methodology in Section III. Section IV presents the results along with the discussion, and finally, Section V states the conclusion.

II. Related Work

The identification of POLB mutations in tumors has sparked interest in their potential role in cancer development and progression. Numerous POLB variants have been detected in multiple types of cancer, such as colorectal, prostate, lung, breast, bladder, and esophageal cancer [7]. Research indicates that POLB mutations are present in up to 40% of tumors [8]. Specifically, missense mutations that affect the protein-coding region have been associated with approximately 35% of cases of prostate cancer [9]. In the context of prostate cancer, somatic mutations in the POLB gene affect enzyme function, leading to microsatellite instability and loss of heterozygosity, which contribute to the progression of the disease. Furthermore, a study of the POLB gene in prostate cancer uncovered more than 20 mutations, most of which were present on more than 50% of the tumor chromosomes, indicating their significant involvement in cancer growth [10]. In particular, a study investigating mutations in the promoter region of the POLB gene revealed that one of these mutations significantly decreases the transcriptional activity of the gene [11]. Within the realm of colorectal cancer, mutations that occur in the catalytic domain of POLB result in a truncated protein that hinders POLB's ability to fill gaps in DNA strands [12]. Furthermore, in gastric cancer, POLB gene mutations were found to affect BER rate and DNA affinity, thus playing an essential role in cancer progression [13]. These mutations and single nucleotide polymorphisms (SNPs) serve as valuable indicators for understanding the development and progression of cancer [14]. SNPs constitute an essential source of variability within the human genome [15]. In particular, they have been recognized as pivotal markers in the association of diseases with specific genes [16]. Furthermore, SNPs have been implicated in various human diseases, including cancer [17]. Multiple mutations have been found to increase chromosomal alterations while reducing POLB gap-filling activity, ultimately causing genomic instability [18]. By investigating the impact of POLB SNPs on cancer development, we could improve the accuracy of cancer diagnosis and treatment options. SNPs have become significant factors in the pathogenesis and progression of various diseases, including cancer.

Bioinformatics tools play a crucial role in the prediction and detection of cancer mutations. These computer-aided tools and methods effectively analyze and interpret genetic information to identify various abnormalities, including mutations and alterations in the DNA sequence, associated with cancer [19]. Using these tools, researchers can uncover genetic markers closely related to cancer development, progression, and treatment. In a specific study [20], a family of breast cancer lacking BRCA mutations was examined to explore genetic mutations. The researchers used three bioinformatics tools to assess the impact of single nucleotide polymorphisms (SNPs) on protein function and their potential role in diseases. SIFT, PolyPhen-2, and Mutation Taster. Through this analysis, they successfully discovered seven variations that posed potential risks. This discovery highlights the practicality of bioinformatics tools in identifying and characterizing the influence of SNPs on the development and progression of diseases. Moreover, another study [21] employed a bioinformatics tool called Functional Analysis through Hidden Markov Models (FATHMM), which has shown superior accuracy compared to traditional SNP prediction tools. The study further validated the effectiveness of

this tool in predicting the impact of SNPs on protein function. Furthermore, a study [6] focused on the prediction of the functional impact of a single nucleotide polymorphism (SNP) on the DNA polymerase β (POLB) protein. Five different bioinformatic tools, SIFT, PolyPhen, CADD, REVEL, and Provean, were used to achieve this. These tools assess the substitution of amino acids caused by SNPs and predict whether they are likely to harm protein function. By integrating the results obtained from these tools, the authors successfully identified several SNPs that could potentially affect the functionality of POLB in DNA repair processes. Additionally, machine learning algorithms can be trained using various data sources, including genomic and clinical data, to predict the probability of cancer development or patient outcomes [22]. These algorithms can analyze and interpret large datasets, allowing the identification of patterns and associations that may not be easily recognizable by traditional analytical methods. Using machine learning techniques, researchers and healthcare professionals can gain valuable insights into cancer prognosis, treatment responses, and personalized medicine approaches. Machine learning techniques, such as Artificial Neural Networks (ANN), Bayesian Networks (BN), Support Vector Machines (SVM), and Decision Trees (DT), have become indispensable in the field of cancer research [23]. Machine learning has been used primarily as a valuable aid in cancer diagnosis and detection [24]. These powerful tools have played a pivotal role in the detection of cancer-related mutations and in the prediction of the occurrence of cancer-associated mutations. Integrating bioinformatics tools with machine learning has revolutionized the identification of cancer-associated Single Nucleotide Polymorphisms (SNPs), improving accuracy in pinpointing clinically significant SNPs. Subsequently, this breakthrough has contributed to better cancer diagnosis and treatment.

III. Methodology

The proposed methodology for identifying cancer-related POLB mutations is presented in Fig. 3 and consists of four stages: a) an assembly stage for the SNP data set, which forms the basis of our analysis, b) a feature extraction stage for the identification and selection of features, c) a negative SNP identification stage, and d) a classification stage to differentiate between cancer-linked and non-cancer-linked mutations.

A. SNPs Dataset

The primary focus of the study was to examine the POLB gene, which is located on chromosome 8. As a result, all data collection was specifically limited to this chromosome. To gather the necessary data sets for Homo sapiens, the information was sourced from four distinct databases. Cancer mutations (positive data) were extracted from the Catalogue of Somatic Mutations in Cancer (COSMIC), which is known as the most extensive and comprehensive database for assessing the impact of somatic mutations on human cancer [25]. In addition to COSMIC, the National Cancer Institute (NCI) was used as a supplementary resource for the acquisition of positive cancer mutations. Furthermore, the number of positive mutations obtained from

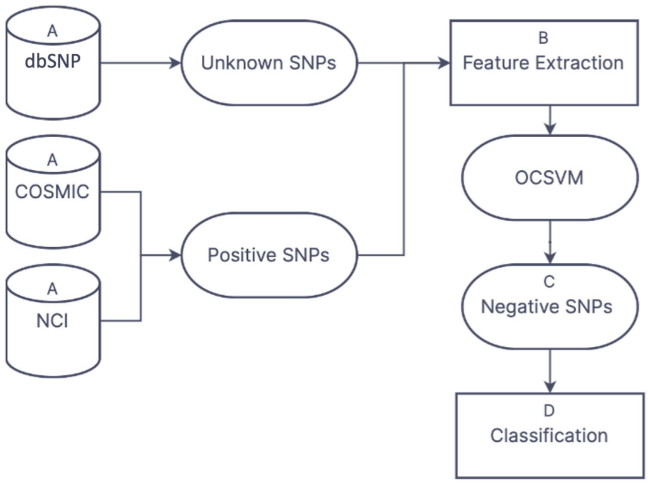


Fig. 3. Methodology flowchart shows the bioinformatics data analysis process followed by the classification. The analysis includes approximately 12,000 POLB unknown single nucleotide polymorphisms (SNPs) sourced from the National Center for Biotechnology Information and the National Human Genome Research Institute's Single Nucleotide Polymorphism Database (dbSNP), 170 positive mutations from COSMIC, and 62 from NCI.

COSMIC was 170, while 62 were from NCI. Additionally, approximately 12,000 unknown single nucleotide polymorphisms (SNPs) from POLB were collected from the National Center for Biotechnology Information and the National Human Genome Research Institute collaborated to establish and maintain the Single Nucleotide Polymorphism Database (dbSNP) [26]. This particular database serves as a publicly accessible repository for genetic variations found within and between different species, providing a valuable resource for the study.

B. Feature Extraction

Feature extraction is a technique used in bioinformatics to pre-process data and reduce its dimensionality. In this particular study, three tools, namely Mutation Taster [27], Fathmm-MKL [28], and the Database for functional predictions of non-synonymous SNPs (dbNSFP) [29], were used to extract SNP features and generate a vector for each SNP. The Mutation Taster tool played a crucial role in retrieving PhyloP scores, which assess the conservation and acceleration of an individual alignment site. These scores provide insight into how conserved a particular site is between different species. However, the Fathmm tool was employed to identify coding and noncoding SNPs and predict their functional consequences. This tool helps distinguish between SNPs that occur within the coding regions of genes, potentially affecting the protein sequence, and those that occur in the non-coding regions, which can influence gene regulation or other non-proteincoding functions. To further improve the analysis, the dbNSFP tool was used to collect additional information for each SNP. This included PolyPhen2 and SIFT prediction scores, which provide estimates of the possible impact of a variant on protein function. The tool also provided the REVEL score and the CADD raw score, which serve as estimates of the pathogenicity and deleteriousness of a variant. These

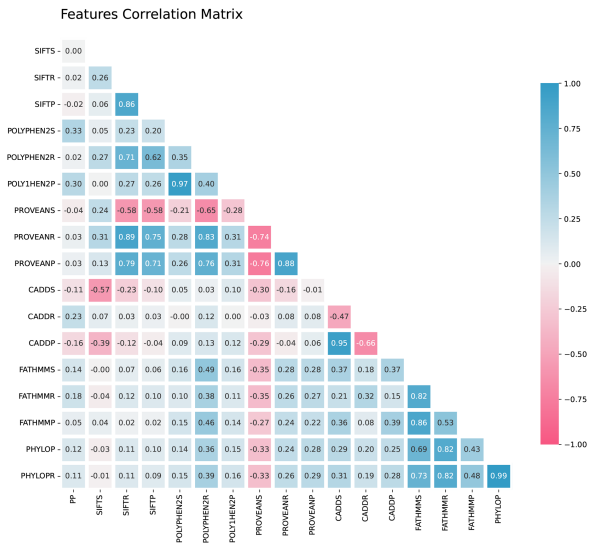


Fig. 4. This heatmap illustrates the correlation matrix of the original 18 SNP features. The variation in color represents the strength and direction of correlations between pairs of features, with darker shades indicating stronger correlations.

scores help determine the possible disease-causing effects of an SNP. Furthermore, the data feature set included rank scores and prediction scores for some features. Despite the extensive feature extraction capabilities of these tools, we encountered a limitation with the initial dataset of 12,000 POLB SNPs due to incomplete data. Not all SNPs had the 18 essential features for our analysis, leading us to focus on a subset with complete information. This decision was crucial to ensure the scientific validity of our study.

1) Feature Selection: To enhance the analysis, various feature selection techniques were applied to determine the importance of each feature. This process effectively solves the dimensionality problem by identifying essential features for classification [30]. As a result, the number of features was reduced from 18 to 14. To visually represent the identified important features and their correlations, a feature selection heatmap was generated. This heatmap, depicted in Fig. 4, illustrates the correlation among all features. The reduction was implemented by identifying and removing features from highly correlated pairs.

C. Negative SNPs

The One-Class Support Vector Machine (OCSVM) model is a type of machine learning algorithm used primarily for anomaly detection. It is trained on data from only one class in this case, the positive data, and aims to test unknown data to identify and classify data points that deviate significantly from the positive pattern. With this framework in mind, we used SNPs from unknown data that were represented with all sets of features, resulting in 616 SNPs. Our approach involved the use of the One-Class Support Vector Machine (OCSVM) model [31], which was used to train positive data to establish a baseline for mutations, then test the unknown data and identify negative mutations. Subsequently, we integrated the results of this process into our classification models to improve our ability to predict mutations. The decision to employ the OCSVM model was motivated by its ability to effectively handle imbalanced data sets, making it suitable for our task where negative data was limited. By training the model with positive data and testing unknowns, we were able to extract patterns and characteristics associated with specific negative mutations in the POLB gene. This enabled us to identify a set of 577 potential negative mutations.

D. Classification Models

In this study, we used eight different binary classification algorithms to classify our data set including Logistic Regression (LR), Weighted Logistic Regression (WLR), Random Forest (RF), Weighted Random Forest (WRF), Rare Event Weighted Kernel Logistic Regression (REWKLR), Multilayer Perceptron (MLP), eXtreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM). Each is characterized by its distinct mathematical model. The mathematical equations for some of the classifiers used are summarized below. In these equations, $f(X_i)$ denotes the prediction outcome for the vector X_i presented by n features $(x_1, x_2, \ldots, x_n)$.

Logistic Regression (LR): Mathematically [32] the underlying LR can be represented by the following equation:

$$f(X_i) = \frac{1}{1 + e^{-(b_0 + b_1 \cdot x_1 + \cdots + b_n \cdot x_n)}} \tag{1}$$

The binary outcome or response variable for the ith observation in the data set is represented by $f(X_i)$. This is what we aim to predict based on the given predictor variables $(x_1, x_2, \ldots, x_n)$ and their corresponding coefficients $(b_0, b_1, \ldots, b_n)$.

Weighted Logistic Regression (WLR): WLR [33] incorporates the weight w_i into the logistic regression to represent larger populations using a small data sample. Two main equations represent WLR:

The logit transformation function that predicts the outcome for vector X_i:

$$f(X_i) = X_i \beta \tag{2}$$

where, $X_i =< x_1, x_2, \ldots, x_n >$, as stated before. The log-likelihood equation that should be maximized to determine the best β vector, which is also inclusive of the regularization parameter λ:

$$lnL\,(\beta) = \sum_{i=1}^{j} w_i ln\left(\frac{e^{y_i X_i \beta}}{1 + e^{X_i \beta}}\right) - \frac{\lambda}{2}\,|\beta|^2 \tag{3}$$

Random Forest (RF) The mathematical model underlying RF [34] for classification tasks is represented as follows:

For a new point X_i, the class prediction $\hat{C}_B(X_i)$ is obtained by majority vote on all trees in the forest:

$$f(X_i) = \hat{C}_B(X_i) = \text{majority vote}\{\hat{C}_b(X_i)\}_{b=1}^{B} \tag{4}$$

In this equation, $\hat{C}_b(X_i)$ represents the class prediction of the b-th random forest tree for the point X_i. The majority vote on the predictions $\hat{C}_b(X_i)$ for b ranging from 1 to B (the total number

of trees) determines the final class prediction $\hat{C}_B(X_i)$ for the point X_i.

Weighted Random Forest (WRF): Following the model presented in [35], the mathematical model underlying the WRF is represented by the following equation:

$$f(X_i) = \sum_{b=1}^{B} w_b \cdot v_{X_i b} \tag{5}$$

Let $X_i b$ be the vote for tree b for subject X_i in the data set. This equation represents the weighted prediction $f(X_i)$ for the ith observation in the data set. The prediction is the sum of votes $v_{X_i b}$ from all trees (B) in the forest, each weighted by a factor w_b determined during training.

Rare Event Weighted Kernel Logistic Regression (REWKLR): Introduces a kernel to WLR to represent rare events nonlinearly [37]. The logit transformation function for REWKLR is defined as:

$$f(X_i) = X_i \alpha \tag{6}$$

The kernel parameter α indicates the width of the kernel. Whereas the regularized log-likelihood is:

$$lnL_W(\alpha) = \sum_{i=1}^{j} w_i ln\left(\frac{e^{y_i X_i \alpha}}{1 + e^{X_i \alpha}}\right) - \frac{\lambda}{2}\alpha^T X \alpha \tag{7}$$

X here is the whole matrix of observations and their feature values. The parameter λ is the regularization variable that ensures that no overfitting of the data is observed.

Multilayer Perceptron (MLP): MLP model for binary classification [38] uses input features X_i to predict the outcome class. Mathematically:

$$f(X_i) = \sigma\left(W_2 g\left(W_1^T X_i + b_1\right) + b_2\right), \tag{8}$$

where X_i represents the input features, W_1, W_2 are the weight matrices, b_1, b_2 are the bias vectors, g is an activation function and (σ) is a logistic function. The result $f(X_i)$ is the probability of X_i being classified into a particular class, according to the threshold of the logistic function.

eXtreme Gradient Boosting (XGBoost): The equation for XGBoost binary classification [39] is represented as:

$$f(X_i) = \sigma \sum_{b=1}^{B} f_b(X_i) \tag{9}$$

where X_i is the feature vector for the i-th instance, B is the total number of trees, $f_b(X_i)$ is the prediction of the b-th decision tree for the i-th instance, and σ is the logistic sigmoid function, which is used to convert the output into a probability.

The model is trained by iteratively adding trees, f_b, which are chosen to minimize the objective function, which is a combination of a specific loss function and a regularization term.

Support Vector Machine (SVM): The kernel underlying the SVM is the Radial Basis Function (RBF) kernel [40]. The prediction function is defined as:

$$f(X_i) = \Phi(X_i).w + b \tag{10}$$

TABLE I
MODEL PERFORMANCE

Model	Optimal Parameters
LR	$\mathcal{C} = 5.65$
WLR	$\mathcal{C} = 5.65$, weight = 3:1
RE-WKLR	$\sigma = 2.8$, $\lambda = 2.2$
RF	Max Depth = 10, # of trees = 40
WRF	Max Depth = 25, # of trees = 60,weight = 5:1
MLP	Hidden layer = (100,) Activation function: ReLU
XGBOOST	Learning rate = 0.5, Max Depth = 3, # of trees = 150
SVM	Kernel: RBF, $\mathcal{C} = 3$, $\gamma = 0.001$

where w is:

$$w = \sum_{i=1}^{j} y_i \alpha_i \Phi(X_{i,train}) \tag{11}$$

and where $f(X_i)$ is the decision function that predicts the class of the input sample X_i, α_i is the Lagrange multiplier associated with the ith support vector. y_i is the class label of the ith support vector (-1 or 1), $\Phi(X_i)$ is the RBF kernel function, measuring the similarity between the input sample and the support vector, and b is the bias term. The RBF kernel depends on the kernel coefficient γ [41].

These models are used to accomplish the task of predicting SNPs based on the application of machine learning classifiers to specific data on cancer genomics.

1) Data Representation and Hyperparameter Tuning: Our study involved 813 SNPs, consisting of 236 positive SNPs and 577 negative SNPs. The SNPs were represented using a binary classification scheme, each SNP was represented by an 18 feature vector, facilitating a comprehensive understanding of their characteristics. This multidimensional representation helped us extract patterns and relationships from the data. The application of each classification algorithm encompassed hyperparameter tuning, in which a set of value ranges was tested and optimized for maximum accuracy using a randomized search. Details of the hyperparameters used for each method are summarized in Table I

2) Data Training: The study's data collection was restricted due to its exclusive focus on the polymerase beta gene, leading to a limited amount of available data. The training process involved the use of 650 non-synonymous single nucleotide polymorphisms (SNPs) and 14 feature scores. To address the issue of imbalanced data, we implemented stratified sampling, which aimed to maintain the proportional distribution of zeros (negative data) and ones (positive data) in both the training and testing data sets. Data were divided in an 80:20 ratio, with 80% allocated for model training and 20% for testing purposes. Specifically, the training data consisted of 461 instances classified as zeros and 189 instances classified as ones.

IV. RESULTS AND DISCUSSION

A. System Specification and Data Preparation/Testing

The classification model was implemented using the Python programming language, specifically using the Anaconda3 environment [42]. The model was run on a MacBook Air equipped

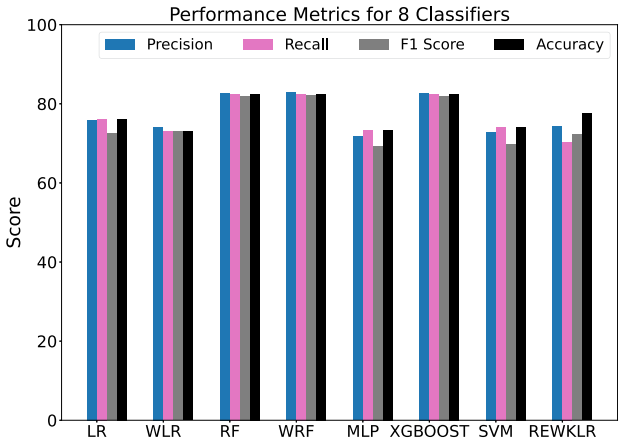


Fig. 5. The performance metrics reported for all classifiers are testing accuracy, f1 score, precision, and recall.

with an M1 chip, 8 GB of memory, and MacOS version 13.1. For the development of the REWKLR algorithm, we used MatLab [36], following the procedure described in [37]. On the other hand, the scikit-learn library [43] was used in Python to design the remaining classification models. All experiments and evaluations were performed on these systems, considering the respective platforms' capabilities. To assess the performance of the classification model, it was tested using a dataset consisting of 163 non-synonymous SNPs along with their corresponding 14 feature scores. In this dataset, the ratio of instances classified as zeros to those classified as ones was 116 to 47. This ratio provided an understanding of the distribution of the classes within the data, which was taken into account during the evaluation process. To evaluate the performance of each model, a comprehensive set of metrics was calculated, including accuracy, precision, recall and the F1 score as shown in table. These metrics are widely used in classification problems to assess the ability of the models to correctly identify positive cancer SNPs and unknown SNPs. We used bootstrapping on the test data to generate the confidence interval for the accuracies and performance metrics of the classifiers. We performed 1000 iterations, with each iteration using a bootstrap sample of size 40 from the test dataset for each classifier.

B. Classification Report

This study aimed to assess the performance of eight binary machine learning classifiers using a dataset consisting of non-synonymous single nucleotide polymorphisms (SNPs) of polymerase Beta. The primary objective was to accurately classify these SNPs as positive or negative based on their association with cancer development. Our findings indicate that all classifiers demonstrated strong performance in this task, suggesting that the data set is high quality and contains crucial features to accurately predict the association of cancer. Specifically, the classifiers exhibited excellent F1 scores, precision and recall rates, highlighting their efficacy in distinguishing between positive and negative SNPs. These performance metrics for each classifier are illustrated in Fig. 5.

Almost all classifiers for predicting cancer mutations within the POLB gene exhibit nearly balanced recall and precision values. This indicates that these classifiers perform well in both correctly identifying positive cases and avoiding false positives (negative mutations). Balanced recall and precision suggest that classifiers effectively capture relevant patterns and features associated with cancer mutations while maintaining a low rate of misclassifications. This balance is important to ensure accurate identification of cancer-related genetic variations and minimize potential false positives, thus improving the reliability of classifiers in practical applications. Additionally, this indicates that our assumption may hold since we used the OCSVM model to extract potentially negative data from the unknown data found in the POLB gene. Furthermore, by adopting this approach, we successfully identified cancer mutations and differentiated them from harmless variations, even in situations where negative data were absent. This suggests that the integration of OCSVM with classification models can offer a robust tool to predict mutations, especially in scenarios where access to negative data is restricted or insufficient. Consequently, this novel methodology holds great promise for improving cancer research and diagnostic efforts. F1 scores obtained from the classifiers to predict mutations within the POLB gene indicate a relatively close performance range of 69.45% to 82.12% among most models. This suggests that classifiers exhibit similar abilities in terms of precision and recall to distinguish between positive and negative mutations. However, despite similar performance, the inclusion of weighting schemes did not lead to significant improvements in LR and WLR F1 scores or RF and WRF. The lack of improvement may be attributed to the nature of the data set and the impact of stratified sampling. The initial balanced class distribution achieved through stratified sampling could already mitigate the need for the weighting scheme in WLR and WRF, resulting in comparable F1 scores for both models. Furthermore, while the WRF classifier achieved the highest F1 score of 82.12% and the MLP classifier had the lowest score of 69.45%, the overall performance differences between these classifiers were not statistically significant. This suggests that, when considering only F1 scores, classifiers show comparable abilities in identifying cancer-related genetic variations within the POLB gene. These findings highlight the importance of considering multiple factors beyond the inclusion of weighting schemes when evaluating the performance of the classifier. The nature of the dataset, sampling techniques, and the specific characteristics of the POLB gene can affect the effectiveness of weighting schemes. It is crucial to explore and optimize other aspects to achieve substantial improvements in differentiating between cancer-linked and non-cancer-linked mutations within the POLB gene. Furthermore, as shown in Fig. 5, the classifiers demonstrated good performance with ensemble classifiers such as RF and its weighted version WRF achieving accuracies of 82.38% and 82.44%, respectively, showcasing the effectiveness of ensemble methods. In addition, XGB and REWKLR also delivered strong results, with accuracies of 82.35% and 77.53%, respectively. In comparison, simpler models such as LR and WLR reported lower accuracy of 76.31% and 72.93%. While SVM and MLP recorded accuracies of 74.34% and 73.58%.

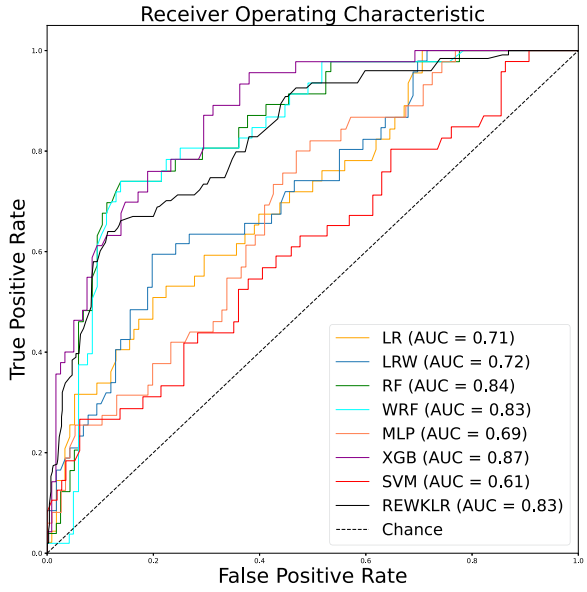


Fig. 6. ROC curve for testing data showing the recall is increased at a low fallout for most classifiers.

This highlights the superior performance of complex models in predictive accuracy.

Additionally, to understand the performance of the model, we examine the Receiver Operating Characteristics (ROC) curve on the testing data, as illustrated in Fig. 6. We used resampling techniques on the testing data to construct these curves and made multiple measurements for each threshold value. This approach allowed us to comprehensively evaluate the performance of the model at various decision thresholds. The results show variations in the AUC of the ROC curve between classifiers. XGBoost, RF, REWKLR, and WRF achieved the highest accuracy with AUC values of 87%, 84%, 83% and 83%, respectively, demonstrating their effectiveness in accurately predicting cancer mutations. On the contrary, LR, WLR, MLP, and SVM achieved a lower accuracy ranging from 61% to 72%. However, it is important to note that the accuracy of the testing data set may be somewhat diminished due to the limited size of the data set. These findings highlight the efficacy of ensemble-based models in enhancing the accuracy of predicting POLB gene mutations. The machine learning classifiers evaluated showed strong performance in accurately classifying cancer-related variations within the POLB gene. These findings provide valuable information for the accurate prediction of cancer mutations using machine learning classifiers.

C. Benchmark Analysis

In this study, we conducted a benchmark analysis to assess the predictive accuracy of our method against the bioinformatics tools, including SIFT, CADD, Polyphen2, Provean, Fathmm, and PhyloP. These tools are widely recognized for their predictive capabilities across various bioinformatics applications. Our evaluation specifically utilized the predicting score of each bioinformatics tool to ensure a fair and direct comparison. The benchmark analysis centered on the classifier with the highest accuracy in our model, the weighted random forest. As illustrated in Table II, our approach (Combined) outperformed the conventional tools, achieving an accuracy of 82%. PhyloP and SIFT were the next most accurate tools, with scores of 75% and 73%, respectively, while Fathmm and CADD reported similar accuracies of 71%. This comparison highlights the effectiveness of our method in providing more accurate predictions than those currently available in the field. The improved performance of our approach is due to our methodology of combining all features of the bioinformatics tools into a comprehensive feature vector. By integrating these diverse features, our method leverages the strengths of each tool, while the weighted random forest classifier effectively synthesizes these inputs to improve prediction accuracy. This approach enables our method to capture a more holistic view of the data, resulting in significantly more precise and reliable predictions of POLB mutations.

TABLE II
BENCHMARK ACCURACY

Tool	Accuracy
SIFT	73%
CADD	71%
Polyphen2	66%
Provean	55%
Fathmm	71%
PhyloP	75%
Combined	82%

V. CONCLUSION

In conclusion, prevention of cancer development is highly dependent on DNA repair mechanisms, particularly base excision repair facilitated by the POLB gene and its corresponding DNA polymerase beta enzyme, which play a crucial role in maintaining genomic stability. Our research emphasizes the essential role of these mechanisms and the potential implications of their mutations in the onset of cancer. In this study, feature extraction tools were used to compile a comprehensive single nucleotide polymorphism (SNP) matrix. This comprehensive matrix of feature vectors served as the basis for subsequent analysis using advanced machine learning algorithms, helping to accurately predict the likelihood of cancer association with specific mutations. Our approach combined Python and MATLAB programming languages along with specialist bioinformatics tools to create a system that was highly efficient in extracting pertinent features from a dataset populated with non-synonymous SNPs and their associated feature scores. The machine learning classifiers used in this study have proven their ability to differentiate successfully between positive and negative cancer mutations. The excellent performance of these classifiers is demonstrated by their impressive F1 scores, precision, and recall rates. The results of this study also draw attention to the merits of ensemble-based models such as XGBoost, RF, and WRF. These models achieved accuracy of 82%, underscoring the importance of thoughtful classifier selection to improve accuracy in the prediction of cancer mutations. Interestingly, our study found that the weighting schemes had a negligible impact on the F1 scores, suggesting that the properties of the data set and the

sampling techniques are critical to determining the performance of the classifier. In our effort to provide a robust and reliable analysis of POLB mutations related to cancer, we have integrated advanced machine learning methodologies, particularly the REWKLR algorithm. This integration showed promise for an accurate identification of cancer-related genetic variations within the POLB gene. Additionally, combining the OCSVM model with other classification techniques effectively predicted cancer mutations and addressed the challenge of limited access to negative data. These results underscore the potential of our research to improve the ability to predict an individual's likelihood of developing cancer based on their unique SNP feature vector. Through effective analysis and interpretation of the complex relationships between specific mutations and cancer development, we can leverage machine learning techniques to significantly influence personalized medicine. To further validate the effectiveness of our approach, we conducted a benchmark analysis comparing our model performance against the predictive accuracy of individual feature scores. This analysis revealed that our integrated method significantly outperforms predictions based on single features, demonstrating a substantial improvement in accuracy. Additionally, we acknowledge the limitations of our study, including the specific sample size and scope that may restrict the broader applicability of our findings. Future research will aim to expand our investigation to other cancer-related genes and employ comparative analyzes with existing mutation classification models. This expansion is crucial for enriching our understanding of cancer development and improving the predictive precision of our models. Addressing the challenge of limited negative data sets and integrating more diverse data with experimental validation are key objectives in our future work. This comprehensive approach is essential to improve the reliability and applicability of our research in the field of cancer genetics and prediction. Our research represents a pioneering effort to apply machine learning prediction methods to POLB gene mutations. By integrating bioinformatics tools with advanced machine learning techniques, we have significantly improved the accuracy of predicting cancer-associated mutations. This novel approach fills a crucial gap in the scientific literature and advances our understanding of polymerase beta mutations in cancer development. Our work not only contributes to the evolution of cancer research, but also paves the way for early detection and personalized treatment strategies.

Data Availability Statement

The datasets used is this study are available on Mendeley Data repository at [44], and the code is available on GitHub at https://github.com/Razan-Alkhanbouli/POLB-Dataset-and-Analysis.

References

[1] Cancer, World Health Organization, Feb. 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *Cancer J. Clinicians*, vol. 72, no. 1, pp. 7–33, 2022.

[3] K. Kiwerska and K. Szyfter, "DNA repair in cancer initiation, progression, and therapy—A double-edged sword,," *J. Appl. Genet.*, vol. 60, no. 3/4, pp. 329–334, 2019.

[4] J. F. Alhmoud, J. F. Woolley, A.-E. Al Moustafa, and M. I. Malki, "DNA damage/repair management in cancers," *Cancers*, vol. 12, no. 4, 2020, Art. no. 1050.

[5] X. Tan et al., "Clinical significance of a point mutation in DNA polymerase beta (POLB) gene in gastric cancer," *Int. J. Biol. Sci.*, vol. 11, no. 2, pp. 144–155, 2015, doi: 10.7150/ijbs.10692.

[6] O. A. Kladova, O. S. Fedorova, and N. A. Kuznetsov, "The role of natural polymorphic variants of DNA polymerase B in DNA repair," *Int. J. Mol. Sci.*, vol. 23, no. 4, 2022, Art. no. 2390.

[7] D. Starcevic, S. Dalal, and J. B. Sweasy, "Is there a link between DNA polymerase beta and cancer?," *Cell Cycle*, vol. 3, no. 8, pp. 996–999, 2004.

[8] S. Ray, M. R. Menezes, A. Senejani, and J. B. Sweasy, "Cellular roles of DNA polymerase beta," *Yale J. Biol. Med.*, vol. 86, no. 4, pp. 463–469, 2013.

[9] C. L. An, D. Chen, and N. M. Makridakis, "Systematic biochemical analysis of somatic missense mutations in DNA polymerase B found in prostate cancer reveal alteration of enzymatic function," *Hum. Mutat.*, vol. 32, no. 4, pp. 415–423, 2011.

[10] N. M. Makridakis, L. F. Caldas Ferraz, and J. K. V. Reichardt, "Genomic analysis of cancer tissue reveals that somatic mutations commonly occur in a specific motif," *Hum. Mutat.*, vol. 30, no. 1, pp. 39–48, 2009, doi: 10.1002/humu.20810.

[11] Q. Wu et al., "Polymorphic mutations in the polb gene promoter and their impact on transcriptional activity," *Thoracic Cancer*, vol. 13, no. 6, pp. 853–857, 2022.

[12] R. Silvestri and S. Landi, "DNA polymerases in the risk and prognosis of colorectal and pancreatic cancers," *Mutagenesis*, vol. 34, pp. 363–374, 2019.

[13] S. Dalal, A. Chikova, J. Jaeger, and J. B. Sweasy, "The Leu22Pro tumor-associated variant of DNA polymerase beta is DRP lyase deficient," *Nucleic Acids Res.*, vol. 36, no. 2, pp. 411–422, 2007, doi: 10.1093/nar/gkm1053.

[14] J. Huszno and E. Grzybowska, "TP53 mutations and SNPs as prognostic and predictive factors in patients with breast cancer (review)," *Oncol. Lett.*, vol. 16, pp. 34–40, 2018.

[15] R. Alzubi, N. Ramzan, H. Alzoubi, and A. Amira, "A hybrid feature selection method for complex diseases SNPs," *IEEE Access*, vol. 6, pp. 1292–1301, 2018, doi: 10.1109/access.2017.2778268.

[16] R. Alzubi, N. Ramzan, and H. Alzoubi, "Hybrid feature selection method for autism spectrum disorder SNPs," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol.*, 2017, pp. 1–7.

[17] E. Capriotti and R. B. Altman, "A new disease-specific machine learning approach for the prediction of cancer-causing missense variants," *Genomics*, vol. 98, no. 4, pp. 310–317, 2011, doi: 10.1016/j.ygeno.2011.06.010.

[18] R. W. Sobol, "Genome instability caused by a germline mutation in the human DNA repair gene POLB," *PLoS Genet.*, vol. 8, no. 11, 2012, Art. no. e1003086.

[19] H. Zheng et al., "Comprehensive review of web servers and bioinformatics tools for cancer prognosis analysis," *Front. Oncol.*, vol. 10, 2020, Art. no. 68.

[20] J. M. Noh, J. Kim, D. Y. Cho, D. H. Choi, W. Park, and S. J. Huh, "Exome sequencing in a breast cancer family without BRCA mutation," *Radiat. Oncol. J.*, vol. 33, no. 2, 2015, Art. no. 149.

[21] H. A. Shihab et al., "Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models," *Hum. Mutat.*, vol. 34, no. 1, pp. 57–65, 2012.

[22] B.-J. Kim and S.-H. Kim, "Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 6, pp. 1322–1327, 2018.

[23] K. Kourou and T. P. Exarchos, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.

[24] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Inform.*, vol. 2, 2006, Art. no. 117693510600200, doi: 10.1177/117693510600200030.

[25] Cosmic, "COSMIC - catalogue of somatic mutations in cancer," Nov. 29, 2022. [Online]. Available: https://cancer.sanger.ac.uk/cosmic

[26] E. M. Smigielski, "DbSNP: A database of single nucleotide polymorphisms," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 352–355, 2000.

[27] J. M. Schwarz, D. N. Cooper, M. Schuelke, and D. Seelow, "MutationTaster2: Mutation prediction for the deep-sequencing age," *Nature Methods*, vol. 11, no. 4, pp. 361–362, Apr. 2014.
[28] M. F. Rogers, H. A. Shihab, M. Mort, D. N. Cooper, T. R. Gaunt, and C. Campbell, "FATHMM-XF: Enhanced accuracy in the prediction of pathogenic sequence variants via an extended feature set," *Bioinformatics*, vol. 34, pp. 511–513, Feb. 2018.
[29] X. Liu, X. Jian, and E. Boerwinkle, "dbNSFP: A lightweight database of human non-synonymous SNPs and their functional predictions," *Hum. Mutat.*, vol. 32, pp. 894–899, 2011.
[30] B. Wei, Q. Peng, X. Kang, and C. Li, "A hybrid feature selection algorithm used in disease association study," in *Proc. 8th World Congr. Intell. Control Automat.*, 2010, pp. 2931–2935, doi: 10.1109/wcica.2010.5554442.
[31] Y. Guerbai, Y. Chibani, and B. Hadjadji, "The effective use of the one-class SVM classifier for reduced training samples and its application to handwritten signature verification," in *Proc. Int. Conf. Multimedia Comput. Syst.*, 2014, pp. 362–366.
[32] M. Maalouf et al., "Logistic regression in data analysis: An overview," *Int. J. Data Anal. Techn. Strategies*, vol. 3, no. 3, pp. 281–299, 2011.
[33] M. Maalouf and M. Siddiqi, "Weighted logistic regression for large-scale imbalanced and rare events data," *Knowl. Based Syst.*, vol. 59, pp. 142–148, 2014.
[34] T. Hastie, R. Tibshirani, and J. Friedman, "Random Forests. In: The elements of statistical learning. Springer series in statistics," Springer, New York, NY, 2009. [Online]. Available: https://doi.org/10.1007/978-0-387-84858-7_15
[35] S. Winham, R. Freimuth, and J. Biernacka, "A weighted random forests approach to improve predictive performance," *Stat. Anal. Data Mining: ASA Data Sci. J.*, vol. 6, pp. 496–505, 2013.
[36] The MathWorks Inc., "MATLAB version: 9.13.0 (R2022b)," Natick, Massachusetts: The MathWorks Inc., 2022. [Online]. Available: https://www.mathworks.com
[37] M. Maalouf, D. Humouz, and A. Kudlicki, "Robust weighted kernel logistic regression to predict gene-gene regulatory association," in *Proc. IIE Annu. Conf.*, 2014, pp. 1356–1360.
[38] B. Müller, J. Reinhardt, and M. T. Strickland, *Neural Networks: An Introduction*, Springer Science & Business Media, 2012.
[39] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, New York, NY, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[40] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[41] J. E. Johnson, V. Laparra, A. Pérez-Suay, M. D. Mahecha, and G. Camps-Valls, "Kernel methods and their derivatives: Concept and perspectives for the earth system sciences," *PLoS One*, vol. 15, no. 10, Oct. 2020, Art. no. e0235885.
[42] Anaconda Software Distribution, "Anaconda documentation, Anaconda Inc. Vers. 2–2.4.0," 2020. [Online]. Available: https://docs.anaconda.com
[43] F. Pedregosa et al.,, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[44] R. Alkhanbouli, A. Alaamri, M. Maalouf, K. Taha, A. Henschel, and D. Homouz, "POLB SNPs dataset," *Mendeley Data*, vol. 1, 2024. [Online]. Available: https://doi.org/10.17632/d6385g6kv6.1

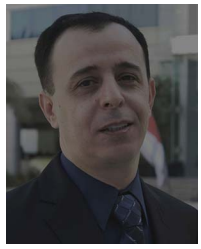
Razan Alkhanbouli received the bachelor's degree in chemical engineering and the master's in industrial and systems engineering from Khalifa University, in 2020 and 2023. She is currently working toward the PhD degree with a research focus on the development of advanced machine learning and data analysis to improve disease prediction. Her research interests include machine learning for rare events, and enhancing healthcare predictive models.

Amira Al-Aamri received the MSc degree from Khalifa University, in 2017, and her research focused on gene-gene network prediction using text mining. She is a research associate with more than 5 years of experience in research, project management, and the development of innovative solutions for various projects at Khalifa University (KU). She specializes in genetic Big Data analysis and is currently a member of the Biotechnology Center, KU, actively participating in diverse projects covering human, animal, and plant genomic data.

Maher Maalouf received the MS and PhD degrees in industrial engineering from the University of Oklahoma, OK, USA. He joined the Department of Industrial and Systems Engineering, Khalifa University, Abu Dhabi, UAE in 2011 and is currently an associate professor. His research interests include applied operations research, machine learning methods, and applications, regression and classification methods, in addition to imbalanced and rare-events data models.

Kamal Taha (Senior Member, IEEE) received the PhD degree in computer science from the University of Texas at Arlington, USA. He is an associate professor with the Department of Electrical and Computer Engineering, Khalifa University, UAE, since 2010. He has more than 100 refereed publications that have appeared in prestigious top-ranked journals, conference proceedings, and book chapters. More than 30 of his publications have appeared in IEEE Transactions journals. He was as an instructor of computer science with the University of Texas at Arlington, USA, from August 2008 to August 2010. He worked as an engineering specialist for Seagate Technology, USA, from 1996 to 2005 (Seagate is a leading computer disc drive manufacturer in the US). His research interests span information retrieval, data mining, databases, bioinformatics, information forensics & security, and defect characterization of semiconductor wafers, with an emphasis on making data retrieval and exploration in emerging applications more effective, efficient, and robust. He serves as a member of the Program Committee, editorial board, and review panel for several international conferences and journals, some of which are IEEE and ACM journals.

Andreas Henschel received the MSc and PhD degrees in computer science from Technical University Dresden, Germany, in 2002 and 2008, respectively. He joined Khalifa University in 2018 and his current position is associate professor with the Department of Electrical Engineering and Computer Science. As a bioinformatician, he is frequently developing algorithms such as Machine Learning based tools to solve problems in the Life Sciences. A common theme in his research is mining large amounts of genetic data, while dealing with hierarchical feature spaces.



Dirar Homouz received the PhD degree from the University of Houston, TX, USA, in 2007. He is an associate professor of physics with the Department of Applied Math and Sciences, Khalifa University. His current research is in the area of computational biophysics where he uses Molecular Dynamics simulations to study protein conformational dynamics in a cell-like environment. In addition, he uses these methods to investigate the confinement effects in many important nano applications. He is also interested in studying gene expression regulation and gene-gene networks. He also works on developing efficient machine learning algorithms to handle high throughput data to infer direct causal relationships in biological systems.