IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received September 20, 2019, accepted October 5, 2019, date of publication October 9, 2019, date of current version October 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2946314

Predicting Drug Interaction With Adenosine Receptors Using Machine Learning and SMOTE Techniques

ABDELRAHMAN I. SAAD, YASSER M. K. OMAR, AND FAHIMA A. MAGHRABY
Arab Academy for Science, Technology and Maritime Transport (AASTMT), Cairo 002022, Egypt
Corresponding author: Abdelrahman I. Saad (abdelrahman.saad@aast.edu)

ABSTRACT Cancer is one of the most influential factors causing death in the world. Adenosine which is a molecule, found in all human cells by coupling with G protein it turns into an adenosine receptor. Adenosine receptor is an important target for cancer therapy. Adenosine stops the growth of malignant tumor cells such as lymphoma, melanoma and prostate carcinoma. Adenosine is activated by interacting with drugs to stop tumor cells from spreading and cure cancer disease. This research aims to predict drugs and potential drug candidates that interact with adenosine receptors. We built a machine learning model using three different classification techniques: Random Forest (RF), Decision Tree (DT) and Support Vector Machine (SVM) then we chose the best technique after comparing the results. Unlike other researches, we used the drug side effect integrated into drug fingerprint as a feature to train our model to classify drugs (interacting and non-interacting) with adenosine receptors. We ranked the interacting drugs with adenosine receptors based on drug side effects to find the most preferred drug (least side effect) among several drugs, which helps in drug design. Most existing datasets contain drugs, targets and the interactions between them, neglecting drug side effects. We formed a new dataset that has the drug side effect. The new dataset is composed of 400 drugs, 794 targets and 3990 drug side effects. Since the dataset was imbalanced we applied Synthetic Minority Oversampling Technique (SMOTE). After conducting experiments, RF achieved the best classification performance with an accuracy of 75.09%.

INDEX TERMS Adenosine, classifier, drug, DTI, drug fingerprint, receptor, side effect, target.

I. INTRODUCTION

Cancer occurs in the form of malignant tumor cause spreading abnormal cells throughout the whole body. There are several types of cancer affecting body organs such as Leukemia cancer forming blood tissues in bone marrow, Myeloma and Lymphoma, which attacks the immune system and weakens it. Finally, the carcinoma that affects the skin or the tissues of the body organs such as the prostate. There are also other types of cancer like sarcoma (affects connecting tissues ex. cartilage), brain and spinal cord cancers. In this study, we focus on blood, skin and immune system cancer types [1], [2].

According to the National Cancer Institute, 1,735,350 new cases are going underdiagnosis in USA and 609,640 people are going to die as a result of the disease [3]. There are several ways to cure cancer such as radiotherapy and chemotherapy. A patient is subjected to radiotherapy. Radiotherapy is therapy using radio waves to control or kill malignant cells while chemotherapy (hormonal thereby) is using chemical drugs to treat the damaged cells which are our concern in this study. Studies prove the presence of high-level ratios of adenosine molecules in cancer tissues. Chemical drugs are a good option to treat these molecules [4]. This adenosine molecule showed a great impact on the growth of tumor cells, which presented an important medical field called drug discovery (drug repositioning) [5].

Diseases are cured by drugs such as cancer in our study by interacting with the target (adenosine). Drugs are designed and tested before using them this process is called drug discovery. Discovering drugs and make use of them require huge time and cost [6]. Machine learning facilitates predicting drug-target interactions and enhances the drug discovery process in addition to developing new applications for the existing drugs [7].

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

Algorithms and machine learning models help a lot in predicting drug-target interaction by reducing cost and time contrast to the molecular docking which simulates the targets in a 3D form, but it cannot simulate all the targets since they should have special features as an input which don't exist for all targets [8]. Computational models used in predicting drug-target interaction are classified to supervised machine learning and semi-supervised machine learning. Supervised machine learning where input and output data are known for classification. In drug-target interaction (DTI) known drug- target interacting pairs are considered positively labeled while the non-interacting ones are considered negatively labeled. Classification models use these labels in training. In semi-supervised machine learning, only some of the data is labeled while the majority are unlabeled. These unlabeled data can reduce the accuracy of the classifier, which leads to bad results.

Adenosine molecule effects appear when it interacts with G-protein coupled, as a result, Adenosine A3, A1 and A2a are formed [9]. Gi and Gq proteins interact with adenosine molecule to form A3 while pertussis toxin-sensitive G proteins (Gi0, Gi1, Gi2 and G3) form A1 and finally the A2a results from interacting with Gs and Golf proteins [10]. A3 receptors are found in tumor cells in the form of HL60 and K562 leukemia while A1 receptors are found in human melanoma A375 cell lines and finally, A2a receptors are found in various cells such as Jurkat lymphoma. Every type of these receptors has a significant role in treating cancer. These receptors are activated using drugs which in turn fight cancer.

The Drug side effect has a great influence on the process of drug design. According to DrugBank, the total drugs are 10562 [11] but only the approved drugs are 3254 which are eligible to be used by patients due to their accepted side effects. In 2016, Coelho *et al.* [12] proposed that integrating drug side effects to other features would enhance drug-target interaction prediction.

Since previous studies, focused on matching drugs and targets in terms of interaction and neglecting their relation (application) to the medical field. Also, they generated drug descriptors from the compound's chemical structure and neglected an important feature such as drug side effects. In addition, we ranked the predicted drugs based on the number of side effects that will help pharmaceutical and doctors in drug design.

Based on the literature survey most of the existing drug target datasets are imbalanced as the count of non-interacting drugs is more than the count of the interacting ones so we applied SMOTE technique to balance our dataset.

The rest of the paper is organized as follows. Section II views the previous studies of DTI and the drug side effects. Section III discusses the dataset, drug features, used machine learning classification models and model's performance evaluation. Section IV illustrates the proposed framework. Section V states the experiments and results. Section VI discusses the experimental results. Finally, section VII concludes the paper and suggest future approaches.

II. LITERATURE SURVEY

In 2008, Campillos *et al.* [13] predicted targets (proteins) using side-effect similarity. In their study, they proved that there was a relationship between drugs and targets connected through drug side effect as two unrelated drugs may have similar side effects by interacting with the same target. In other words, this strong relation helped in predicting new targets for old drugs. Their dataset was collected from Matador [14], DrugBank [15] and Ki DB [16] public databases, which contained 746 drugs, 4857 drug-target relations and a side-effect network, formed of drug-drug relations. They developed a side-effect similarity measure using weighting schemes then they classified drug side-effects using Unified Medical Language System (UMLS). By constructing ontology network, they concluded that there was an inversely proportional relationship between the recurrence of drug side effect and two drugs sharing the same target (protein), finally they predicted 2903 drug-target interacting pairs with a probability of 25%.

In 2016, Coelho *et al.* [12] used two machine learning classification models SVM and RF to predict DTI. The first model predicted drugs with reference to the target's type while the second model predicted drugs without referring to the target's type. They collected their dataset from DrugBank [15] and Yamanishi *et al.* [7] research, and consisted of 927 drugs, 1370 targets, and 5127 drug interactions. SVM model with reference to the target's type (protein) showed a great result in terms of AUC (Area Under The Curve). They suggested a future approach to enhance the prediction of DTI by using both network centrality metrics and expanding the area of proteomic space.

In 2016, Galeano and Paccanaro [17] presented the idea of chemical similarity prediction which was built on the theory of the drugs that are similar in their chemical structures help in predicting targets near to them. They collected their dataset from Biogrid [18] and DrugBank [15] databases, which contained 9336 drugs, and 4612 targets. They built two networks, the first network consisted of nodes and each node represented a drug and similarity between drugs was calculated by using Tanimoto Coefficient. The second network represented the interactions between proteins and called interactome to detect the relationship between them. Finally, they measured the similarity between the two networks to predict similar targets. The similarity between two networks reached 85% in terms of AUC, they suggested that enhancing the similarity ratio would occur when integrating side effect similarity.

In 2017, Sinha *et al.* [19] used Decision Tree, Support Vector Machine, Random Forest and Naïve Bayes as classification techniques to inspect Leishmania Donovani membrane a special kind of protein. The aim of their study was to predict the usability of the protein whether to be a drug target or a vaccine. They used four classification techniques and 28 proteins [20] as an input to their model then they evaluated each technique and used the best one. Finally, they used another 37 proteins and decided on the role of each protein (drug-target or vaccine). The best result was obtained by Naïve Bayes with an accuracy of 76.17%.

In 2017, Hao et al. [21] used Dual-Network Integrated Logistic Matrix Factorization (DNILMF) to predict DTI. They proposed that similar drugs and targets could help in predicting nearby drugs and targets. They formed a new dataset which contained 829 drugs, 733 targets and 3688 interactions. They used kernel construction techniques to build drug and target profiles, calculated the matrix profiles using kernel techniques, then similar classes were diffused. They predicted DTI using DNILMF which was better than Neighborhood Regularized Logistic Matrix Factorization (NRLMF); they said that using genetic algorithm could enhance their proposed model.

In 2017, Wen *et al.* [22] predicted new drug and target interactions without considering the type of targets by using a deep learning methodology. They formed their dataset of 1412 drugs, 520 targets and 2146240 interaction pairs between drug and target. The data was extracted from DrugBank [15] database. They used Extended Connectivity Fingerprints (ECFPs) to generate drug descriptors and Protein Sequence Compositions (PSCs) to generate target descriptors. A neural network called Deep Belief Networks (DBN) was implemented. They tested their model using an external dataset from DrugBank [23] containing 4383 drugs, 2528 targets and 7352 interaction pairs between drug and target. They compared their model to Random Forest (RF), Decision Trees (DT) and Bernoulli Naïve Bayes (BNB) classifiers. The accuracy of DBN, BNB, DT and RF were 85%, 72%, 76% and 83% respectively.

In 2018, Manoochehri and Nourani [24] used Deep Matrix Factorization (DMF) to predict drug-target interaction. They discussed two approaches. The first approach was to build a predictive model based on identifying non-interacting negative pairs (drug-target) in the unlabeled data then using both positive and negative pairs to build the model. The second approach was predicting data using Ranking on Top methods which rank the positive interacting pairs higher than the non-interacting negative ones. Their model was divided into two steps. They used K-Nearest Neighbor technique (KNN) classification technique to extract negative samples form data then they used DMF i.e. a deep learning approach to generate latent vectors. They used golden benchmark dataset constructed by Yamanishi *et al.* [25] where there were four different target classes Ion Channels (IC), Enzymes, G-Protein-Coupled Receptors (GPCR) and Nuclear Receptors (NR) that contains 204, 445, 95 and 54 drugs in IC, Enzyme, GPCR and NR respectively and 210, 664, 223 and 26 targets in IC, Enzymes, GPCR and NR respectively. They evaluated their model using Area Under the Precision-Recall (AUPR) curve, Area Under the Curve (AUC) and 10-fold cross-validation. Finally, they compared their model with Neural Matrix Factorization (NeuMF) and DMF with random sampling. The results were higher using their proposed model (DMF+KNN) with an average accuracy of 73.65% using Hit Ratio Metric.

In 2019, Saad *et al.* [26] used KNN, RF and DT machine-learning classification techniques to predict DTI. They formed their dataset from drug central and spider version 4.1 public databases. They built two matrices. The first matrix was the drug side effect matrix and the second was the drug-target matrix used in training and testing. They did three experiments to study the effect of using drug features. The aim of the first experiment was to use drug fingerprints to classify drugs and identify their interaction with the corresponding targets, KNN achieved an accuracy of 95.6%. The aim of the second experiment was to use drug side effects to classify drugs and identify their relation to targets, KNN achieved an accuracy of 91.28%. Finally, the aim of the third experiment was to classify drugs based on using both drug side effects and drug fingerprint KNN achieved an accuracy of 97.63%. They came to the conclusion that using drug fingerprints besides drug side effects enhanced the accuracy of the used classifiers. TABLE 1 summarizes the previous related work.

In this study, we extended our work to deeply focus and concentrate on finding a medical application for the previous study as it was a general case study. So, we worked on a specific types of targets (adenosine receptors) and studied the effect of drugs on these targets in terms of interaction. We made use of the drug side effect to rank drugs to help in the drug design process and choose the best drug alternatives for the patient. We also used new techniques in our experiments such as SMOTE to balance the dataset and SVM for classification.

III. MATERIALS AND METHODS

A. DATASETS

We used the dataset in the study conducted by Saad *et al.* [26], it was a combination of two datasets where the first dataset was a drug dataset extracted from drug central and contained 2736 drugs, 1938 targets and 14521 interactions between drugs [27] while the other one was a drug side effect dataset extracted from spider version 4.1 [28] and contained 1430 drugs, 5868 side effects and 139756 drug side effect pairs.

The new dataset was compiled using joining and merging techniques. We used the compound ID to join the two datasets. The resulted dataset is 400 drugs in common that has multiple targets and drug side effects as shown in TABLE 2 [26]. After that, we focused on our area of interest which is the adenosine receptors and their associated interacting drugs with their corresponding side effects.

We used SMOTE to generate new instances from the new samples from the classes having minor cases (24 drugs in A3 and 11 drugs in both A1 and A2a) by taking instances from the features in space for the target classes and the nearby classes then created new samples based on combining the features of the target classes with the features of the nearby classes as shown in TABLE 3.

B. DRUG FINGERPRINT

Drugs are represented by a feature vector called drug fingerprint. Drug fingerprint is obtained by simulating the

TABLE 1. Summary of related work.

Paper authors	Objectives	Approach	Dataset	Accuracy
Monica Campillos et al. [13], 2008	Predicted targets (proteins) using side effect similarity	1. Similarity networks	1. Matador 2. DrugBank 3. Ki DB	Predicted drug-target interacting pairs with a probability of 25%
Edgar D. Coelho et al. [12], 2016	Predicted drug target interactions without considering the type of target	1. Random Forests 2. Support-Vector Machine 3. Logistic regression	1. Yamanishi 's dataset 2. DrugBank	1. Average SVM: 92.75% 2. Average RF: 92.25%
Diego Galeano et al. [18], 2016	Predicted targets based on drug chemical structure	1. Tanimoto similarity	1. Biogrid 2. Drugbank	The AUC (Area Under The Curve) reached 85 % in similarity
Arvind Sinha et al. [20], 2017	Predicted usability of the protein whether being a drug-target or a vaccine	1. SVM 2. DT 3. RF 4. Naïve Bayes	Research by Kumar et al., 2015 titled: "Proteomic analyses of membrane enriched proteins.."	1. SVM: 63% 2. RF: 73% 3. DT: 56.33% 4. Naïve Bayes: 76.17%
Ming Hao et al. [22], 2017	Predicted interactions between drug and target using DNILMF	1. DNILMF 2. NRLMF	Compiled a drug-target interaction dataset using compound ID	1. Average DNILMF: 97.57% 2. Average NRLMF: 96.9%
Ming Wen et al. [23], 2017	Predicted new DTI without considering the type of targets	1.Deep learning	1.Yamanishi 's dataset 2. DrugBank	1. DBN: 85% 2. BNB: 72% 3. DT: 76% 4. RF: 83%
Hafez Manoochehri et al. [25], 2018	Predicted drug-target interactions using DMF	1. KNN 2. DMF	1.Yamanishi 's dataset	1. Average DMF+KNN: 73.65% 2. Average DMF with random sampling: 71.95% 3. NeuMF: 72.47%
Abdelrahman Saad et al. [27], 2019	Predicted drug-target interactions using machine learning classification techniques	1. RF 2. DT 3. KNN	1. Drug cental 2. Spider 4.1	1. Using drug fingerprint: 93.57% (DT), 93.84% (RF) and 95.16% (KNN) 2. Using drug side effect: 89.97% (DT), 90.23% (RF) and 91.28% (KNN) 3. Using drug fingerprint and drug side effect: 96.89% (DT), 96.97% (RF) and 97.63% (KNN)
Our research	Predicted drugs interacting with adenosine receptors using machine learning and ranking predicted drugs based on drug side effects	1.SVM 2. DT 3. RF 4. SMOTE	1. Drug central 2. Spider 4.1	1. Adenosine A3: 70.53% (SVM), 70.26% (DT) and 73.68% (RF) 2. Adenosine A1: 61.90 (SVM), 66.48% (DT) and 66.30% (RF) 3. Adenosine A2a: 69.78% (SVM), 74.36% (DT) and 75.09% (RF)

TABLE 2. Summary of datasets.

No. of features	Drugs	Targets	Side-effect
Dataset 1	2376	1938	-
Dataset 2	1430	-	5868
New dataset	400	794	3990

TABLE 3. Number of drugs before and after SMOTE.

No. of drugs	Before SMOTE			After SMOTE		
	A3	A1	A2a	A3	A1	A2a
Interacting	24	11	11	205	278	278
Non-interacting	377	390	390	377	390	390

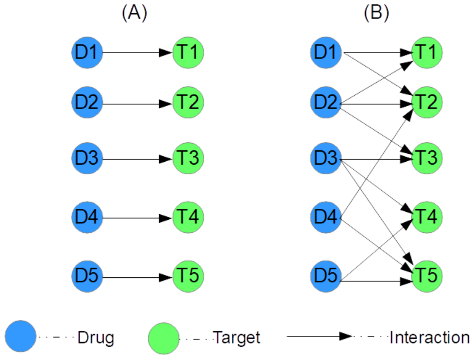


FIGURE 1. Prediction scenario.

molecules forming the drug. The simulation is based on molecule information such as the atom numbers and the bonds between these atoms. This information then used to generate encoded fingerprints (binary bits) to be used later as a strong features. Drug fingerprint is used in classification and drug similarity techniques [29] that help in predicting new potential drugs for the existing targets and vice versa as shown in Figure 1. In part (A) each drug interacts with one corresponding target as pairs [(D1, T1), (D2, T2), (D3, T3), (D4, T4), (D5, T5)] i.e., drug 1 interacts only with target 1, this can be used as a valuable input in drug-target interaction but it is not sufficient for prediction and discovering hidden interactions between drugs and targets. In part (B) we can find that not only every drug interacts with only one target as pairs [(D1, (T1, T2)), (D2, (T1, T2, T3)), (D3, (T3, T4, T5)), (D4, (T2, T5)), (D5(T4, T5))], i.e. drug 1 interacts with two targets named target 1 and target 2, but each drug could interact with multiple targets with the helping of drug fingerprint similarity, for instance, two drugs sharing same drug fingerprint could interact with the same targets, thus will help in finding new applications for the existing drugs.

C. DRUG SIDE EFFECT

Drugs have side effects that cause unpleasant symptoms to patients e.g. skin rash and dizziness. Side effects highly impact drug discovery as it limits the use of drugs and decreases its value. Drug side effects vary from one person to another depending on the reaction between the chemical substances in the drug and the targeted cells in the human body. It has been reported that the severity of side effects is the second cause for drug manufacturing failure and the fourth cause leading to death in USA [30], [31].

D. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is one of the famously used machine learning techniques. SVM is a machine learning classification method. It is summarized as follows: inputs, represented by input vectors are non-linearly mapped to a high dimensional feature space. A decision surface and a quadratic formula are constructed to classify between those input features while ensuring high generalization ability of the learning machine. It is considered as a robust and powerful method in data analysis and pattern recognition [32]. Support Vector Machine (SVM) was proposed by Vapnik and Chervonenkis in the 1990s. There are two types of patterns linear and nonlinear. The basic idea of SVM is to construct a decision plane (hyperplane) to separate set of objects belonging to different classes [33], given the following data set $(x_i y_i)$ for $i = 1 \ldots N, \mathbf{x}_i \in R^d$ and $y_i \in \{-1, 1\}$ for training a classifier of $f(x)$ as in equation (1)

$$f\left(\mathrm{x}_i\right)\begin{cases}\geq 0 & y_i = +1\\ < 0 & y_i = -1\end{cases} \tag{1}$$

classes are correctly classified when $y_i f\left(\mathbf{x}_i\right) > 0$ in case of binary classification, but for linear classification classifier has an equation (2) in the form of

$$f(x) = w^{\tau} x + b \tag{2}$$

Since w represents the weight of the vector and b represents the bias (SVM parameters), for better classification, performance the margin is maximized using equation (3)

$$f(x) = \sum_i \alpha_i y_i \left(\mathbf{x}_i^{\top}\mathbf{x}\right) + b \tag{3}$$

where X_i are supporting vectors that support the algorithm and it is defined when the value of α_i (weight of the point) is not zero.

E. DECISION TREE (DT)

Decision Tree (DT) learning is one of the most used methods for inductive inference. It is a classification method that approximates discrete-valued target function. Decision Trees are constructed using only those attributes best able to differentiate the concepts to be learned [34]. A DT is built by initially selecting a subset of instances from a training set. This subset is then used by the algorithm to construct a DT. The remaining training set instances test the accuracy of the constructed tree. If the DT classifies the instances correctly, the procedure terminates. If an instance is incorrectly classified, the instance is added to the selected subset of training instances and a new tree is constructed. This process continues until a tree that correctly classifies all non-selected instances are created or the DT is built from the entire training set. A statistical property, called Information Gain, is used. Information Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected. In order to define Information Gain, first, we have to define an idea from an information theory called Entropy. Entropy measures the amount of information in an attribute using equation (4)

$$Entropy(S) = -\sum_{c\in C} p_{c\,\log_2} p_c \tag{4}$$

Given a collection S of c outcomes where p_c is the chance of an instance of S belongs to outcomes c. Another metric is the Information Gain, which measures how powerful an attribute can sort data as in equation (5)

$$InformationGain(S, F) = Entropy(S) - \sum_{f\in F} \frac{|S_f|}{|S|} Entropy\left(S_f\right) \tag{5}$$

Given a collection S having set of features S_f and count of elements in S with feature F having value f.

F. RANDOM FOREST (RF)

Random Forest (RF) is proposed by Breiman [35]. A collection of bagged decision trees based on the idea of ensemble learning where combining several machine algorithms to form a big generalized machine learning algorithm [36]. Several trees are built using bootstrap aggregating algorithms by extracting a random subset of data built by the trees [37]. Finally, based on certain splitting criteria such as Gini [38] trees are built. Trees classify the existing features and nominate the tree class based on voting then the forest selects the most voted classification path of all other trees.

The RF algorithm can be implemented as follows:

step 1: Select attributes Y from total attributes X where Y < X.

step 2: Calculate node N from random attributes Y by building a split.
step 3: Calculate the next node O using the best split.
step 4: Repeat the previous steps until only one single node is reached.
step 5: Build N trees by repeating step 1 to step 4.
step 6: Prediction data P is obtained from the N trained trees using classification voting.
step 7: Build the final model based on the highest voted predicted attributes.

G. SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE)

During our first experiment in this study, we found that the accuracy and the specificity of the used classifiers are high and the sensitivity (true positive rate) is very low due to the dataset was imbalanced as the number of drugs interacting with adenosine molecule is relatively small compared to the non-interacting drugs. This problem affected our classification performance results. We used an oversampling technique called Synthetic Minority Oversampling Technique (SMOTE) as proposed by Chawla *et al.* [39] and used in many fields such as bio-informatics [40]. In this study, we used SMOTE which made the number of interacting drugs and non-interacting drugs with adenosine receptors nearly equal which balanced our dataset. To balance the dataset SMOTE uses the following equation

$$D_{syn} = D_i + (D_{Knn} - D_i) \times r \tag{6}$$

where D_{syn} is the synthetic data, D_i are minority samples, D_{Knn} a sample of k-nearest neighbor from minority samples and r is a random number between 0 and 1

SMOTE algorithm can be implemented as follows:

step 1: Determine both D_i (feature vector) and D_{Knn} (k-nearest neighbor from minority samples).
step 2: Output the difference between the feature vector and the k-nearest neighbor from minority samples.
step 3: Multiply output by r (a random number between 0 and 1).
step 4: Add the output to the feature vector D_i to select a new point on the line segment between feature vectors.
step 5: Repeat steps from 1 to 4 to identify new feature vectors.

H. PERFORMANCE MEASURE

The proposed framework was assessed using Accuracy, Sensitivity, Specificity, Postive Predicted Value (PPV) and Negative Predicted Value (NPV) as shown below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

$$PPV = \frac{TP}{TP + FP} \tag{10}$$

$$NPV = \frac{TN}{TN + FN} \tag{11}$$

In this study, TP means true positive (sign of interaction with adenosine receptors), TN means true negative (drug not interacting with adenosine receptors), FN (predicted positive drug-adenosine receptor pairs to be not interacting) and FP (predicted negative drug-adenosine receptor pairs to be interacting) where positive means there is an interaction between drug and the receptor while negative there is no interaction between them.

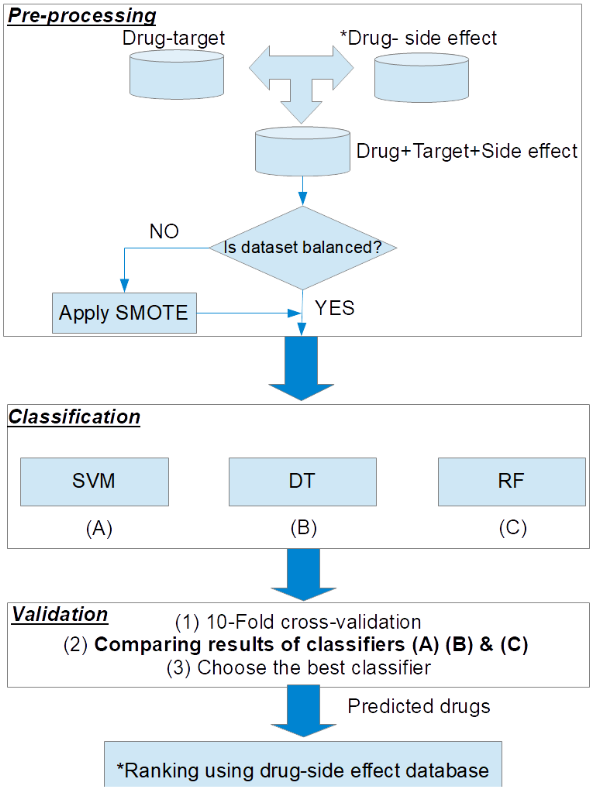


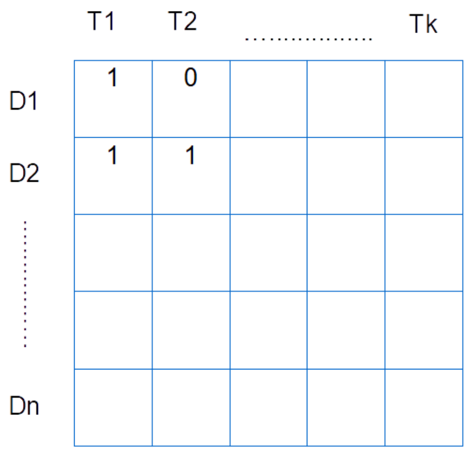
FIGURE 2. **Proposed drug-adenosine receptors interaction framework.**

IV. PROPOSED MODEL

The aim of our framework is to predict drug-target interactions by applying machine-learning techniques and select the best classifier between these different techniques. We generated a drug fingerprint for each drug using drug features enclosed in Structure Data File (SDF) by calling the ChemmineR library in R. We resolved the bias in the dataset by using the SMOTE technique as shown in Figure 2. Afterward, we used the machine learning classification techniques to classify drugs based on the drug fingerprint similarity and the known drug-target interactions to train our model which will help to predict new interactions. Finally, we ranked the newly predicted drugs based on the drug side effect feature to eliminate drugs having dangerous side effects.

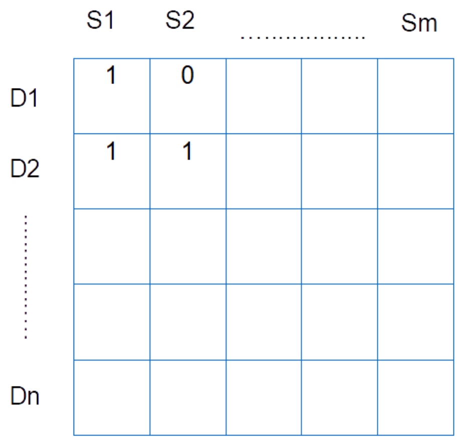
A. PREPROCESSING PHASE

In the processing phase, we combined both datasets and generated a new dataset as mentioned earlier in TABLE 2. After that we labeled (represented by a random number) each drug, target and drug side effect then each label was encoded. Lastly, we started building matrices, the first matrix represented drug-target pairs where a drug (D1) interacts with one target or more (Tk) integrating to this matrix the drug fingerprint as shown in Figure 3. The second matrix represented drug-side effect pairs where a drug (D1) had several side effects (Sm) as shown in Figure 4.



	T1	T2	………………		Tk
D1	1	0			
D2	1	1			
⋮					
Dn					

FIGURE 3. Drug-target matrix.



	S1	S2	………………		Sm
D1	1	0			
D2	1	1			
⋮					
Dn					

FIGURE 4. Drug-side effect matrix.

We extracted our part of interest in which the drugs interacting with adenosine receptors (A3, A1 and A2a) and their corresponding side effects to form a new matrix then we applied the SMOTE technique to balance our dataset.

B. CLASSIFICATION PHASE

In the classification phase we trained our model using SVM, DT and RF with hyper-parameters Sigmoid kernel and Gini criterion respectively and the data of the adenosine receptors A3, A1 and A2a as an input to the classifiers.

C. VALIDATION AND TESTING PHASE

We split the data into 70% training and 30% testing then we applied 10-fold cross-validation technique to test and validate our data by splitting the training set into 10 folds where k equals 10 then we trained our models on 9 folds and we tested it on the one remaining fold, then we took an average of different 10 accuracies of the model evaluation which helps in concise analysis. The final step is to compare the results of the different classifiers and choose the best classifier.

V. EXPERIMENT RESULTS

Drug discovery undergoes many phases before a certain drug can be approved to be taken by patients to treat a certain disease. Prediction of these drugs must be highly accurate because predicting the wrong drugs can affect the patient causing unpleasant side effects that could lead to death. Machine-learning (classification) techniques were used to predict if there is an interaction between drugs and adenosine receptors. Before, carrying the experiment on the whole data we held 5% of the real data to ensure the data is synthesized correctly before applying our model on the whole data using SMOTE. Three different experiments were conducted on three different types of adenosine receptors (A3, A1 and A2a) and the results of the three classifiers were compared after using SMOTE technique.

A. EXPERIMENT ON PART OF THE DATA

Before carrying the main experiment (whole data), We extracted 5% of the data to validate using the SMOTE technique on the whole data. The highest accuracy was obtained by RF with an average accuracy of 70% and an average sensitivity of 76% while the lowest accuracy was obtained by SVM with an average accuracy of 60% and an average sensitivity of 59% as shown in TABLE 4, TABLE 5, TABLE 6, TABLE 7, TABLE 8 and TABLE 9.

TABLE 4. Adenosine A3 receptor before using SMOTE on part of the data.

(%)	Before SMOTE on part of the data (A3 receptor)		
	SVM	DT	RF
Accuracy	50	33	66
Sensitivity	50	25	50
Specificity	50	50	75
PPV	66	50	50
NPV	33	25	75

B. EXPERIMENT ON THE WHOLE DATA

In this section, we carried the experiment on the whole data and showed the results after using SMOTE on the adenosine receptors (A3, A1 and A2a).

1) ADENOSINE A3 RECEPTOR USING SMOTE ON THE WHOLE DATA

After applying the SMOTE technique on the whole data, the dataset is balanced with 205 interacting drugs and 377 non-interacting drugs with adenosine receptors.

TABLE 5. Adenosine A3 receptor using SMOTE on part of the data.

(%)	Using SMOTE on part of the data (A3 receptor)		
	SVM	DT	RF
Accuracy	60	50	70
Sensitivity	50	50	80
Specificity	75	50	60
PPV	75	60	66
NPV	50	40	75

TABLE 6. Adenosine A1 receptor before using SMOTE on part of the data.

(%)	Before SMOTE on part of the data (A1 receptor)		
	SVM	DT	RF
Accuracy	33	50	50
Sensitivity	50	100	100
Specificity	25	25	0
PPV	25	40	25
NPV	50	100	100

TABLE 7. Adenosine A1 receptor using SMOTE on part of the data.

(%)	Using SMOTE on part of the data (A1 receptor)		
	SVM	DT	RF
Accuracy	70	70	70
Sensitivity	71	66	75
Specificity	66	75	66
PPV	83	80	60
NPV	50	60	80

TABLE 8. Adenosine A2a receptor before using SMOTE on part of the data.

(%)	Before SMOTE on part of the data (A2a receptor)		
	SVM	DT	RF
Accuracy	33	50	66
Sensitivity	0	66	75
Specificity	33	33	50
PPV	0	50	75
NPV	100	50	50

TABLE 9. Adenosine A2a receptor using SMOTE on part of the data.

(%)	Using SMOTE on part of the data (A2a receptor)		
	SVM	DT	RF
Accuracy	50	60	70
Sensitivity	57	66	75
Specificity	33	50	50
PPV	66	66	85
NPV	25	50	33

SVM, DT and RF achieved an accuracy of 70.53%, 70.26% and 73.68% respectively and sensitivity of 76.84%, 71.58% and 76.84% added to it a specificity of 64.21%, 68.95% and 70.53% for SVM, DT and RF respectively. Also a PPV (Positive Predictive Value) of 68.22%, 69.74% and 72.28% and NPV (Negative Predictive Value) of 73.49%, 70.81% and 75.28% in case of SVM, DT and RF respectively as shown in TABLE 10.

TABLE 10. Adenosine A3 receptor using SMOTE on the whole data.

(%)	Using SMOTE on the whole data (A3 receptor)		
	SVM	DT	RF
Accuracy	70.53	70.26	73.68
Sensitivity	76.84	71.58	76.84
Specificity	64.21	68.95	70.53
PPV	68.22	69.74	72.28
NPV	73.49	70.81	75.28

2) ADENOSINE A1 RECEPTOR USING SMOTE ON THE WHOLE DATA

After applying the SMOTE technique on the whole data, the dataset is balanced with 278 interacting drugs and 390 non-interacting drugs with adenosine receptors. SVM, DT and RF achieved an accuracy of 61.90%, 66.48% and 66.30% respectively and sensitivity of 56.41%, 60.07% and 59.71% added to it a specificity of 67.40%, 72.89% and 72.89% for SVM, DT and RF respectively. Also a PPV (positive predictive value) of 63.37%, 68.91% and 68.78% and NPV (Negative Predictive Value) of 60.73%, 64.61% and 64.40% in case of SVM, DT and RF respectively as shown in TABLE 11. The SVM, DT and RF accuracy ratio decreased by 8.63%, 3.78% and 7.38% respectively compared to the accuracy in A3 receptor experiment. While there was a slight increase in specificity by 3.19%, 3.94% and 2.36%.

TABLE 11. Adenosine A1 receptor using SMOTE on the whole data.

(%)	Using SMOTE on the whole data (A1 receptor)		
	SVM	DT	RF
Accuracy	61.90	66.48	66.30
Sensitivity	56.41	60.07	59.71
Specificity	67.40	72.89	72.89
PPV	63.37	68.91	68.78
NPV	60.73	64.61	64.40

3) ADENOSINE A2A RECEPTOR USING SMOTE ON THE WHOLE DATA

After applying the SMOTE technique on the whole data, the dataset is balanced with 278 interacting drugs and 390 non-interacting drugs with adenosine receptors. SVM, DT and RF achieved an accuracy of 69.78%, 74.36% and 75.09% respectively and a sensitivity ratio of 75.82%, 77.29% and 79.49% added to it a specificity of 63.74%, 71.43% and 70.70% for SVM, DT and RF respectively. Also a PPV (positive predictive value) of 67.65%,73.01% and 73.06% and NPV (Negative Predictive Value) of 72.50%, 75.88% and 77.51% in case of SVM, DT and RF respectively as shown in TABLE 12.

C. RANKING THE INTERACTING DRUGS WITH ADENOSINE TARGETS

Instead of drugs that help cure patients from certain diseases, it causes side effects symptoms that can lead to death. Therefore we choose 5 random interacting drugs with each

TABLE 12. Adenosine A2a receptor using SMOTE on the whole data.

(%)	Using SMOTE on the whole data (A2a receptor)		
	SVM	DT	RF
Accuracy	69.78	74.36	75.09
Sensitivity	75.82	77.29	79.49
Specificity	63.74	71.43	70.70
PPV	67.65	73.01	73.06
NPV	72.50	75.88	77.51

TABLE 13. Adenosine A3 ranked drugs.

Drug	Side effects
Adenosine	87
Baclofen	100
Atenolol	112
Caffeine	133
Amiodarone	241

TABLE 14. Adenosine A1 ranked drugs.

Drug	Side effects
Lovastatin	34
Clotrimazole	239
Gabapentin	265
Mefloquine	403
Cladribine	557

TABLE 15. Adenosine A2a ranked drugs.

Drug	Side effects
Tamoxifen	7
Raloxifene	29
Miconazole	106
nifedipine	131
sildenafil	172

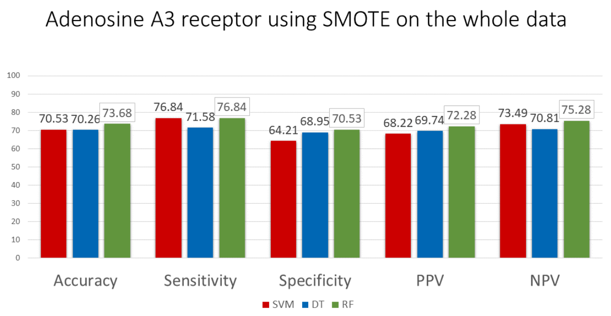


FIGURE 5. Adenosine A3 receptor using SMOTE on the whole data.

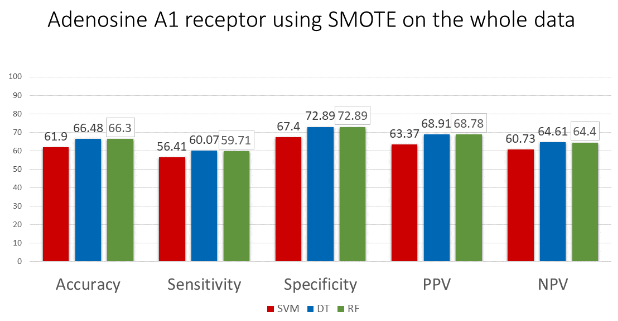


FIGURE 6. Adenosine A1 receptor using SMOTE on the whole data.

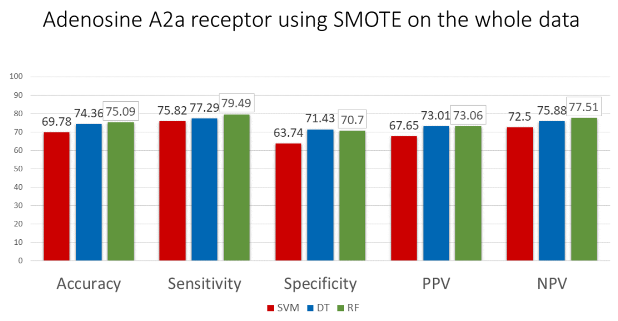


FIGURE 7. Adenosine A2a receptor using SMOTE on the whole data.

adenosine receptors and ranked it from least to most based on side effects. For adenosine receptor A3 the interacting drugs were Adenosine, Amiodarone, Atenolol, Baclofen and Caffeine and have side effects such as agitation, high blood pressure and bronchospasm. Also, Cladribine, Clotrimazole, Gabapentin, Lovastatin and Mefloquine interacted with adenosine A1 receptors. While Miconazole, Nifedipine, Raloxifene, Sildenafil and Tamoxifen interacted with adenosine A2a receptor as shown in TABLE 13, TABLE 14 and TABLE 15.

VI. RESULTS DISCUSSION

The experiments showed that RF and DT got the highest accuracy in classifying drugs interacting with adenosine receptor A2a 75.09% and 74.36% respectively. The incorrect classification affected the three classifiers across the three adenosine receptors since the interacting drug instances with these receptors are too small compared to the non-interacting ones. So we used the SMOTE technique in our experiments to create synthetic data to solve the imbalanced dataset problem. RF had the highest accuracy among the three classifiers across the three target receptors with an average accuracy of 71.69%, highest sensitivity with an average sensitivity of 72.01% and highest PPV with an average PPV of 71.37%. While the lowest specificity was scored by SVM with an average specificity of 65.11% and the lowest NPV with an average NPV of 68.90% as shown in Figure 5, Figure 6 and Figure 7.

VII. CONCLUSION

Cancer is considered one of the most dangerous diseases affecting humans. High-cost lab experiments and researches are applied to find a cure for cancer. Enhancing the drug discovery process highly depends on analyzing and processing drug features to develop new drugs that will interact with targets in the human body to cure the diseases. We proposed a machine learning model to help in predicting drugs interacting with targets based on drug fingerprints. In this study, we focused on a special kind of targets called adenosine receptors. We suffered a problem of the unbalanced dataset which was a misleading factor in the classification performance and accuracy. We used SMOTE to solve the problem of the unbalanced dataset using three different classifiers across three different adenosine targets. RF achieved the best classification performance with an accuracy of 75.09%. Finally, we ranked the output drugs interacting with adenosine receptors based on the drug side effect. Adenosine was the least interacting drug with adenosine A3 receptor with 87 side effects while lovastatin was the least interacting drug with adenosine A1 receptor with 34 side effects and finally tamoxifen as the least interacting drug with adenosine A2a receptor with 7 side effects. In the future, we will apply another classification technique to enhance the accuracy of the prediction also, increase the drug and target instances. We will also, consider weighting drug side effects based on medical experiences to determine the degree of severity of the predicted drug.

REFERENCES

[1] *Types of Cancer 2018*, Cancer Res., London, U.K., 2018.

[2] *Skin Cancer 2018*, Amer. Cancer Soc., Atlanta, GA, USA, 2016.

[3] *Cancer Statistics*, Nat. Cancer Inst., Bethesda, MD, USA, 2018.

[4] S. Gessi, S. Merighi, P. A. Borea, S. Cohen, and P. Fishman, "Adenosine receptors and current opportunities to treat cancer," in *The Adenosine Receptors*. Cham, Switzerland: Humana Press, 2018, pp. 543–555. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-90808-3_23

[5] P. Fishman, S. Bar-Yehuda, F. Barer, L. Madi, A. S. Multani, and S. Pathak, "The A3 adenosine receptor as a new target for cancer therapy and chemoprotection," *Exp. Cell Res.*, vol. 269, no. 2, pp. 230–236, Oct. 2001.

[6] Z. Li, P. Han, Z.-H. You, X. Li, Y. Zhang, H. Yu, R. Nie, and X. Chen, "In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences," *Sci. Rep.*, vol. 7, no. 1, Sep. 2017, Art. no. 11174.

[7] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, pp. i246–i254, Jun. 2010.

[8] X. Chen, C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, "Drug–target interaction prediction: Databases, Web servers and computational models," *Briefings Bioinf.*, vol. 17, no. 4, pp. 696–712, Aug. 2015.

[9] P. Fishman, S. Bar-Yehuda, M. Synowitz, J. D. Powell, K. N. Klotz, S. Gessi, and P. A. Borea, "Adenosine receptors and cancer," in *Adenosine Receptors in Health and Disease* (Handbook of Experimental Pharmacology). Berlin, Germany: Springer, 2009, pp. 399–441. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-89615-9_14

[10] D. Allard, M. Turcotte, and J. Stagg, "Targeting A2 adenosine receptors in cancer," *Immunol. Cell Biol.*, vol. 95, no. 4, pp. 333–339, Feb. 2017.

[11] *Drug Statistics*, Can. Inst. Health Res., Edmonton, AB, Canada, 2018. [Online]. Available: https://www.drugbank.ca/

[12] E. Coelho, J. Oliveira, and J. Arrais, "Ensemble-based methodology for the prediction of drug-target interactions," in *Proc. IEEE 29th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2016, pp. 36–41.

[13] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity," *Science*, vol. 321, no. 5886, pp. 263–266, Jul. 2008.

[14] S. Gunther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork, and R. Preissner, "SuperTarget and matador: Resources for exploring drug-target relationships," *Nucleic Acids Res.*, vol. 36, pp. D919–D922, Dec. 2007.

[15] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "DrugBank: A comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Res.*, vol. 34, no. 1, pp. D668–D672, Jan. 2006.

[16] B. L. Roth, E. Lopez, S. Patel, and W. K. Kroeze, "The multiplicity of serotonin receptors: Uselessly diverse molecules or an embarrassment of riches?" *Neuroscientist*, vol. 6, no. 4, pp. 252–262, Aug. 2000.

[17] D. Galeano and A. Paccanaro, "Drug targets prediction using chemical similarity," in *Proc. 42th Latin Amer. Comput. Conf. (CLEI)*, Oct. 2016, pp. 1–7.

[18] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: A general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. 1, pp. D535–D539, Jan. 2006.

[19] A. Sinha, P. Singh, A. Prakash, D. Pal, A. Dube, and A. Kumar, "Putative drug and vaccine target identification in leishmania donovani membrane proteins using Naïve Bayes probabilistic classifier," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 1, pp. 204–211, Jan. 2017.

[20] A. Kumar, P. Misra, B. Sisodia, A. Shasany, S. Sundar, and A. Dube, "Proteomic analyses of membrane enriched proteins of Leishmania donovani Indian clinical isolate by mass spectrometry," *Parasitol. Int.*, vol. 64, no. 4, pp. 36–42, Aug. 2015.

[21] M. Hao, S. H. Bryant, and Y. Wang, "Predicting drug-target interactions by dual-network integrated logistic matrix factorization," *Sci. Rep.*, vol. 7, no. 1, Jan. 2017, Art. no. 40376.

[22] M. Wen, Z. Zhang, S. Niu, H. Sha, R. Yang, Y. Yun, and H. Lu, "Deep-learning-based drug–target interaction prediction," *J. Proteome Res.*, vol. 16, no. 4, pp. 1401–1409, Mar. 2017.

[23] S. David Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: A knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Res.*, vol. 36, pp. D901–D906, Nov. 2007.

[24] H. E. Manoochehri and M. Nourani, "Predicting drug-target interaction using deep matrix factorization," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2018, pp. 1–4.

[25] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug–target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, Jun. 2008.

[26] A. Saad, F. A. Maghraby, and Y. M. Omar, "Predicting drug target interaction by integrating drug fingerprint and drug side effect using machine learning," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl.*, Mar. 2019, pp. 281–290.

[27] O. Ursu, J. Holmes, C. G. Bologa, J. J. Yang, S. L. Mathias, V. Stathias, D.-T. Nguyen, S. Schürer, and T. Oprea, "DrugCentral 2018: An update," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D963–D970, Oct. 2018.

[28] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, Oct. 2015.

[29] D.-S. Cao, Q.-N. Hu, Q.-S. Xu, Y.-N. Yang, J.-C. Zhao, H.-M. Lu, L.-X. Zhang, and Y.-Z. Liang, "In silico classification of human maximum recommended daily dose based on modified random forest and substructure fingerprint," *Anal. Chim. Acta*, vol. 692, nos. 1–2, pp. 50–56, Apr. 2011.

[30] K. M. Giacomini, R. M. Krauss, D. M. Roden, M. Eichelbaum, M. R. Hayden, and Y. Nakamura, "When good drugs go bad," *Nature*, vol. 446, no. 7139, pp. 975–977, Apr. 2007.

[31] E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Côté, B. K. Shoichet, and L. Urban, "Large-scale prediction and testing of drug activity on side-effect targets," *Nature*, vol. 486, no. 7403, pp. 361–367, Jun. 2012.

[32] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.

[33] A. Pradhan, "Support vector machine—A survey," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 8, pp. 82–85, Aug. 2012.

[34] G. Bombara, C.-I. Vasile, F. Penedo, H. Yasuoka, and C. Belta, "A decision tree approach to data classification using signal temporal logic," in *Proc. 19th Int. Conf. Hybrid Syst., Comput. Control (HSCC)*, 2016, pp. 1–10.

[35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[36] S. F. Abdoh, M. A. Rizka, and F. A. Maghraby, "Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques," *IEEE Access*, vol. 6, pp. 59475–59485, 2018.

[37] Y. Wu, H. Wang, and F. Wu, "Automatic classification of pulmonary tuberculosis and sarcoidosis based on random forest," in *Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Inform. (CISP-BMEI)*, Oct. 2017, pp. 1–5.

[38] P. P. Graczyk, "Gini coefficient: A new way to express selectivity of kinase inhibitors against a family of kinases," *J. Med. Chem.*, vol. 50, no. 23, pp. 5773–5779, Oct. 2007.

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002.

[40] T. Deepa and M. Punithavalli, "An E-SMOTE technique for feature selection in high-dimensional imbalanced dataset," in *Proc. 3rd Int. Conf. Electron. Comput. Technol.*, Apr. 2011, pp. 322–324.

ABDELRAHMAN I. SAAD was born in Jeddah, Saudi Arabia, in 1992. He received the bachelor's degree in information systems from the Arab Academy for Science Technology and Maritime Transport (AASTMT), Cairo, Egypt, in 2015, where he is currently pursuing the master's degree in information systems. From 2016 to 2017, he was a Soldier with the Egyptian Army. Since 2018, he has been a Graduate Teaching Assistant. His current research interests include bioinformatics, machine learning, and big data.

YASSER M. K. OMAR received the Ph.D. degree in biomedical engineering from Cairo University, Cairo, Egypt. He has been an Assistant Professor with the Department of Computer Science, Faculty of Computing and Information Technology, Arab Academy for Science Technology and Maritime Transport (AASTMT). His current research interests include bioinformatics, medical imaging, data visualization, machine learning, and computing algorithms.

FAHIMA A. MAGHRABY received the B.S., M.S., and Ph.D. degrees from Ain Shams University, Cairo, Egypt, in 2003, 2008, and 2014, respectively, all in computer science. From 2004 to 2014, she was a Lecturer Assistant with the Institute of Computer Science, Shorouk Academy, Cairo. Since 2014, she has been a Lecturer with the Faculty of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport (AASTMT), Cairo. Her current research interests include bioinformatics, imaging processing, artificial intelligence, and cloud computing.