Turkish Journal of Electrical Engineering and Computer Sciences

Volume 27 | Number 4 Article 33

1-1-2019

Toxicity prediction of small drug molecules of aryl hydrocarbon receptor using aproposed ensemble model

VISHAN KUMAR GUPTA

PRASHANT SINGH RANA

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 2833 – 2849

doi:10.3906/elk-1809-9

Toxicity prediction of small drug molecules of aryl hydrocarbon receptor using a proposed ensemble model

Vishan Kumar GUPTA*, **Prashant Singh RANA**

Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Received: 02.09.2018 • **Accepted/Published Online:** 26.03.2019 • **Final Version:** 26.07.2019

Abstract: Quantitative structure–activity relationships and quantitative structure–property relationships have proved their usefulness for predicting toxicities of drug molecules regarding their biological activities. In silico toxicity prediction techniques are essential for reducing testing on rodents (in vivo) and for a less time-consuming and more cost-efficient alternative for the identification of toxic effects at an early stage of drug development. The authors aim to build a prediction model for better assessment of toxicity to quickly and efficiently test whether certain chemical compounds have the potential to disrupt the processes in the human body that may adversely affect human health. Here, we have proposed a computational method (in silico) for the toxicity prediction of small drug molecules using their various physicochemical properties (molecular descriptors) that can bind to the aryl hydrocarbon receptor. Pharmaceutical data exploration laboratory software is used for extracting the features of drug molecules. The dataset of the aryl hydrocarbon receptor contains 9008 drug molecules, where 1063 are active and 7945 are inactive, and each drug molecule contains 1444 features. It is a novel prediction model based on ensemble learning that can efficiently classify active (binding) and inactive (nonbinding) compounds of the dataset. In our proposed ensemble model, we primarily performed feature selection using the Boruta library in R, after which we resolved the class imbalance problem itself by ensemble learning where we divided the dataset into seven data frames, which have approximately equal numbers of active and inactive drug molecules. An ensemble model based upon the votes of seven random forest models is proposed, which gives an accuracy of 93.76%. K-fold cross-validation is conducted to measure the consistency of the model. Finally, the validity of the proposed ensemble model for some drug molecules of acquired immune deficiency syndrome therapy and androgen receptor has been proved.

Key words: Aryl hydrocarbon receptor, molecular descriptor, feature selection, class imbalance, toxicity, ensemble model

1. Introduction

Most drugs are small molecules that are invented to interact with, bind, and regulate the activity of specific biological receptors. Receptors are a group of proteins present in the cell that interact and bind with other molecules to perform the various tasks necessary for the maintenance of life. Receptors include a vast array of cell-surface receptors (hormone receptors, neurotransmitter receptors, cell-signaling receptors, etc.), enzymes, and other functional proteins. Owing to physiological stressors and genetic abnormalities, the function of specific enzymes and receptors may alter to the point that our well-being is diminished. These alterations seem to cause

*Correspondence: vishangupta@gmail.com

minor physical symptoms, such as a running nose due to allergies, or life-threatening and debilitating events like depression or sepsis [1].

Typically, drugs are small organic molecules that accomplish their desired activity by binding with a target site on a receptor. The initial phase in the discovery of a new drug is usually to identify and separate the receptor to which it should bind, followed by testing many small molecules for their ability to bind to the target site [2]. Researchers must distinguish the active (binding) compounds from the inactive (nonbinding) compounds. This can lead to the design of new compounds that will not only bind but will also have all other properties needed for a drug. These properties are solubility, oral absorption, appropriate duration of action, toxicity, lack of side effects, and so on. [1]

The data challenge of Tox21 with the collaboration of the National Center for Biotechnology Information (NCBI) is held to help researchers understand the chemical and compound toxicology that can disrupt biological pathways in a manner that may result in toxic effects. It is an open challenge where researchers must predict about compounds' interventions in biochemical pathways by using only physicochemical structure data. Active drug molecules are those molecules that can bind to one or more biochemical pathway assays and create some toxic effects in our bodies. These toxic effects are stress response (SR) effects and nuclear receptor (NR) effects. Both SR and NR effects are highly relevant to human health because the activation of nuclear receptors can disrupt endocrine system function, and the activation of stress response pathways can lead to liver injury or cancer [3]. We can build computational models to predict the activity of the drug molecules in one or more of the 12 pathway assays of NR or SR based on their physicochemical properties. In this paper, we are analyzing the toxic effects only on the aryl hydrocarbon receptor. Table 1 shows the 12 biological pathway assays that can give distinct adverse health effects on its activation.

Table 1. Nuclear receptor signaling and stress response pathways.

Nuclear receptor panel	AR: androgen receptor, full AR-LBD: androgen receptor, LBD ER: estrogen receptor alpha, full ER-LBD: estrogen receptor alpha, LBD AhR: aryl hydrocarbon receptor PPAR-gamma: peroxisome proliferator-activated receptor gamma aromatase
Stress response panel	Nrf2/ARE: nuclear factor-like 2/antioxidant responsive element HSE: heat shock factor response element (HSE) ATAD5: genotoxicity indicated by ATAD5 MMP: mitochondrial membrane potential p53

Generally, the in silico approach is a predictive science utilized for defining discovery and safety efforts in therapeutics [4]. The primary purpose of toxicity prediction with the use of computational methods is to reduce the testing on living cells or tissues. Therefore, it is an alternative to the bioassay. The concept of Russell and Burch regarding the growing popularity of the 3Rs (replacement, reduction, refinement) focuses on

[1]Kaggle (2018). Drug Activity Prediction [online]. Website https://www.kaggle.com/c/DrugActivityPrediction [accessed 02 July 2018].

the limited use of animals and the unlimited use of computational techniques for toxicity testing [5]. In silico models also can predict ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties in chemical space, which can reduce the dependency of chemical laboratory synthesis (in vitro) [6].

The aryl hydrocarbon receptor (AhR) is a protein, and it is a member of the family of basic helix-loop-helix transcription factors. AhR adopts the responses of environmental pollutants such as aromatic hydrocarbons through the induction of phase I and phase II enzymes. These adaptive responses are toxic responses with various side effects, whereas the elicitation of metabolizing enzymes results in the production of toxic metabolites. AhR is also called the dioxin receptor because it is a ligand-activated transcriptional regulator that binds dioxin and other exogenous contaminants and is responsible for their toxic effects. Dioxin and dioxin-like compounds (DLCs) are highly toxic environmentally persistent organic pollutants (POPs), which can cause developmental problems and immunological disorders by interfering with hormones. DLCs can also create disorders in the nervous system, endocrine system, and reproductive functions and can even cause cancer. Exposure to high levels of dioxin in humans may result in skin lesions, such as patchy and chloracne darkening of the skin, impairment of the immune system, and modified liver function [7].

In this paper, we have proposed a novel ensemble-based binary classification model for forecasting the activity of AhR drug molecules whether a given specific compound is active (1) or inactive (0). In our proposed ensemble model, initially, we have performed feature selection using the Boruta library in R, and then the class imbalance problem is resolved through an ensemble learning method where we divide the dataset into seven data frames, which have approximately equal numbers of active or inactive drug molecules. Subsequent to this, each data frame is trained and tested at 70% and 30%, respectively. An ensemble model based on the votes of seven random forest models is created, which is our proposed ensemble model that has also resolved the issue of class imbalance. K-fold cross-validation is performed to measure the robustness of the proposed ensemble model. Finally, we have proved the validity of this model for some new drug molecules that are neither part of the training dataset nor part of the testing dataset. Therefore, we applied our proposed ensemble model to some drug molecules of AIDS therapy and some drug molecules of androgen receptors for validation, where our model has given the best accuracy. The significant contributions of this paper are as follows:

1. To develop better toxicity assessment features, methods, and algorithms for drug molecules of AhR.
2. To develop a machine learning-based model for quick and efficient testing of certain chemical compounds that have probable chance of disrupting the processes in the human body.
3. To develop a stand-alone application for helping researchers to predict the toxicity of newly discovered chemical compounds and environmental chemicals.
4. To develop a computational method (in silico) for checking the toxicity of drug molecules of AhR rather than inside the living organism (in vivo) or within glass (in vitro).

Figure 1 shows the general diagram of the prediction model, where the various physicochemical properties of any small drug molecule are taken, and its activity is predicted through our prediction model. The research community of the state of art Tox21 data challenge did not consider the problem of feature dimensionality or the class imbalance problem during the model formation, but we have built the proposed ensemble model considering these issues and tuned the parameter for the betterment of prediction accuracy [14].

The paper is composed as follows: Section 2 contains related work regarding toxicity prediction using various methods. Section 3 introduces a quick overview of the dataset, feature extraction using PaDEL, feature

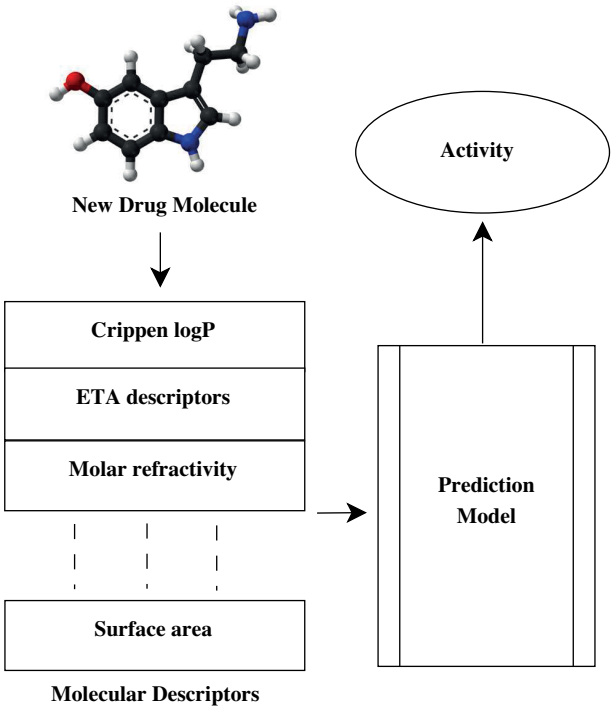

Figure 1. Prediction method.

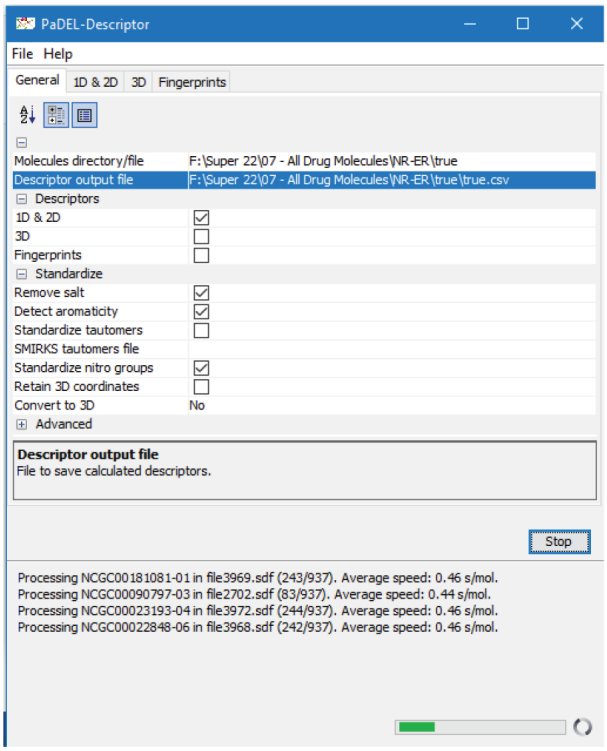

Figure 2. PaDEL-Descriptor GUI.

selection, and the class imbalance problem. Section 4 clarifies the procedure of the proposed ensemble model. Section 5 presents the description of the random forest model, which is used as a base classifier for ensemble learning. Section 6 presents the different performance evaluation parameters of the model for classification. Section 7 describes the investigated, compared, and validated results, followed by the conclusion in Section 8.

2. Related work

Basak et al. proposed a hierarchical QSAR approach, which used topological indices to predict aryl hydrocarbon receptor binding potency on a set of 34 chlorinated dibenzofurans [8].

Kola and Landis proposed that toxicity is also the central issue for the development of a new drug. According to reported clinical trials, more than 30% of drug candidates fail because of undetected toxic effect [9].

Piparo et al. proposed a computational model for predicting aryl hydrocarbon receptor binding, where the authors decided to use QSAR models for the binding prediction virtual screening due to the unavailability of the AhR X-ray crystal structure. They used a training set of 84 AhR ligands [10].

Cassano et al. developed the CAESAR QSAR model to minimize false negatives to make them more usable for the European REACH legislation (Registration, Evaluation, Authorization, and Restriction of Chemical Substances). The CAESAR online application ensures that both industry and regulators can easily access and use the developmental toxicity model. The CAESAR platform is a freely available tool for the study of human toxicity [11].

Drwal et al. proposed molecular similarity-based and naive Bayes classification for the prediction of the toxicity of the nuclear receptor and stress response pathway, which was screened from the Tox21 data challenge of 2014. It was implemented in KNIME software [12].

Stefaniak proposed a machine learning model to predict the activity of drug molecules in the nuclear receptor panel and stress response panel using low-dimensional molecular descriptors and machine learning algorithms. The models were built using the rotation forest and ADTree classifier, and the performance of the model was measured using area under the receiver operating characteristic curve metrics [13].

Capuzzi et al. built QSAR models for 12 stress response and nuclear receptor signaling pathway toxicity assays as part of the 2014 Tox21 challenge. These models were built using random forest, deep neural networks, and various combinations of descriptors, where deep neural networks performed better. The drawback of this methodology is the high demand for computational resources [14].

3. Materials and methods

3.1. Pharmaceutical Data Exploration Laboratory (PaDEL)

Pharmaceutical Data Exploration Laboratory (PaDEL) software is used to compute the molecular descriptors and fingerprints. The input of PaDEL-Descriptor is the structure-data files (SDFs) of AhR drug molecules, and its output is a comma-separated values (CSV) file. The CSV file contains a total of 9008 drug molecules, and each drug molecule has 1444 features. The PaDEL-Descriptor is a Java-based free and open source software, which is similar to Dragon, MOE, and MARVIN Beans and supports more than 90 different molecular file formats including PDB, SDF, and SMILE. The molecular descriptors are extracted using the Chemistry Development Kit (CDK) library of Java, which is related to the chemoinformatics and bioinformatics that are used internally in PaDEL. The PaDEL software can calculate 1876 molecular descriptors (1444 1D and 2D descriptors, and 431 3D descriptors) and 12 kinds of fingerprints. We have used only 1444 1D and 2D descriptors in our dataset for activity prediction of drug molecules. Figure 2 shows the graphical user interface (GUI) of PaDEL-Descriptor [15], and Figure 3 shows the format of a structure-data file for an active drug molecule of AhR.

3.2. Dataset

The AhR signaling pathway data is taken from PubChem (https://pubchem.ncbi.nlm.nih.gov/bioassay/743122), where 743122 is the PubChem identification number for AhR. PubChem is maintained by the National Center for Biotechnology Information and provides access to biomedical and genomic information from its website. In this study, our dataset consists of a total of 9008 AhR drug molecules, of which 1063 are active molecules and the remaining 7945 are inactive molecules. All the drug molecules have 1444 features, which are also known as physicochemical properties or molecular descriptors, which are extracted by PaDEL-Descriptor. The most common molecular descriptors are the partition coefficient (AlogP), molar refractivity (AMR), volume, elements count, ETA descriptors, autocorrelation, nBase, nRing, apol, number of hydrogen atoms (nH), and number of carbon atoms (nC). Table 2 lists some essential physicochemical properties of the drug molecules of AhR and their descriptions.

Table 3 shows an overview of the dataset that contains various AhR drug molecules, such as NCGC00257625-01, NCGC00259354-01, and NCGC00255335-01. The columns of Table 3 show the various molecular descriptors/features, such as ATS0m, AATSC8m, and Mlogp. These features are extracted from the structured-data file using PaDEL-Descriptor. Here, activity is a target class, which shows whether a drug molecule is active or inactive.

```
NCGC00015959-03
  Marvin  07111412562D

 25 30  0  0  0  0            999 V2000
    3.4098   -1.3130    0.0000 N   0  3  0  0  0  0  0  0  0  0  0  0
    4.8329   -1.3130    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.4098   -2.1380    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.1248   -2.5436    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.6948   -2.5436    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.8329   -2.1380    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.1248   -0.8937    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.5547   -0.8937    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0

  1  3  1  0  0  0  0
  1  7  2  0  0  0  0
  1 25  1  0  0  0  0
  2  7  1  0  0  0  0
  2  6  2  0  0  0  0
  2  8  1  0  0  0  0
  3  4  2  0  0  0  0
  3  5  1  0  0  0  0
  4 13  1  0  0  0  0
  4  6  1  0  0  0  0
  5  9  1  0  0  0  0

M  CHG  1   1   1
M  END
>  <Formula>
C20H14NO4

>  <FW>
332.3289

>  <DSSTox_CID>
25204

>  <Active>
1
```

Figure 3. Structure-data file (SDF) format for a single drug molecule of the aryl hydrocarbon receptor.

3.3. Feature selection using the Boruta algorithm

During the process of model building, the feature selection is used to filter the highly correlated variables, descriptors with too many zero values, several missing values, and unwanted noise from the dataset. Our dataset has 1444 features, which are very high in quantity; therefore, it will increase the time and space complexity during model building. Feature selection is a process of selecting the important features that may improve the performance of the model and remove those attributes that have redundant and irrelevant information. Here, the process of feature selection is carried out using the Boruta() function under the **Boruta** library in R. It is a wrapper algorithm that finds relevant features on the basis of the values of meanImp, medianImp, minImp, maxImp, and normHits [16].

The input parameters of Boruta function are the dataset of 1444 features and target variable (activity). After the execution of this algorithm, only 150 attributes are confirmed as important, whereas 1294 attributes are confirmed as unimportant. Now only the confirmed attributes are used for model building. Table 4 shows all the important features of AhR dataset.

3.4. Class imbalance

Class imbalance is a problem in machine learning where the main class of interest is rare, which triggers bias of the classifier. Here, our dataset for the prediction of activity contains two classes; one is active and the

Table 2. Physicochemical properties of aryl hydrocarbon receptor's drug molecules.

S. no.	Name	Description
1	Crippen logP	Atom-based calculation of logP using Crippen's method, also called the octanol/water partition coefficient.
2	Eccentric connectivity index	It is a distance-based atomic descriptor used for numerical modeling of biological activities, which are of varied nature.
3	Fragment complexity	It reduces the "interaction" complexity and correlates with the increased probability of binding to a target.
4	Kappa shape indices	The kappa shape records are the premise of a technique for molecular structure quantization in which the characteristics of molecular shape are encoded into three indices (kappa values).
5	Molecular linear free energy relation	These descriptors are intended to reflect the crucial molecular properties, which are critical in solvation-related procedures, specifically polarity, size, and hydrogen bonding.
6	Weighted path	Weighted path numbers used to describe the molecular descriptors for structure–property–activity studies.
7	Charged partial surface area	These descriptors were initially designed for studies of structure–physical relationships. They capture information about different features of molecules that are responsible for polar intermolecular interactions.
8	Molecular refractivity (AMR)	Molecular refractivity is a measure of the aggregate polarizability of a mole of a substance. It is dependent on the pressure, temperature, and index of refraction.
9	Extended topochemical atom (ETA)	Index for modeling drug-induced and chemical toxicities.
10	Autocorrelation (ATS_0, ATS_1)	Index that measures the degree of linear relationship between a given time series and a lagged version of itself over successive time intervals.

Table 3. Dataset of aryl hydrocarbon receptors.

Name	Activity	ATS0m	AATSC8m	GATS5m	Mi	MLogP	VE3_D	Apol
NCGC00257625-01	1	3829.964592	–0.459127243	0.813611134	7.618377096	3.44	–6.693111535	55.24
NCGC00259354-01	1	2901.539444	2.391427338	1.207252358	7.36495489	3.66	–7.011135521	30.40
NCGC00255335-01	1	3909.376441	2.721650422	0.887052279	7.561662379	3.55	–8.791765353	19.62
NCGC00181290-01	0	3820.09244	–0.038428669	0.79059275	7.645415724	3.22	–7.4769931	50.46
NCGC00181294-01	0	8146.012676	–1.213070372	1.475781891	7.685123769	2.56	–132.8711591	31.53
NCGC00181300-01	0	8149.887015	–1.651536457	1.059976061	7.660974482	3.55	–8.190848602	41.56

other is inactive. This dataset is extremely imbalanced, as the total number of active drug molecules is 1063 (minority class) and the total number of inactive drug molecules is 7945 (majority class). Therefore, active drug molecules are far fewer than inactive drug molecules. The main function of class balancing is to balance the class symmetry of instances. There are several conventional approaches to handle the class imbalance problem, which are undersampling, oversampling, and the synthetic minority oversampling technique (SMOTE) [17, 18]. Here, the class imbalance problem is resolved by the ensemble learning method, as ensemble learning is more effective

Table 4. Important features of aryl hydrocarbon receptor.

S. no.	Features	S. no.	Features	S. no.	Features	S. no.	Features	S. no.	Features
1	AMR	31	SpMax1_Bhm	61	SHsNH2	91	maxdsN	121	MLFER_S
2	naAromAtom	32	SpMin1_Bhm	62	SHdsCH	92	maxdS	122	MLFER_E
3	nAromBond	33	SpMax1_Bhv	63	SHaaCH	93	gmin	123	MLFER_L
4	nN	34	SpMax1_Bhe	64	SHother	94	MAXDP	124	MPC5
5	ATS3m	35	SpMin1_Bhe	65	SdsCH	95	DELS	125	piPC2
6	AATS0m	36	SpMax1_Bhp	66	SaaCH	96	MAXDP2	126	piPC3
7	AATS1m	37	SpMax1_Bhi	67	SsssCH	97	DELS2	127	piPC4
8	AATS2m	38	SpMin1_Bhi	68	SdssC	98	ETA_dEpsilon_B	128	piPC5
9	AATS4m	39	C2SP2	69	SaasC	99	ETA_Beta	129	piPC6
10	AATS0v	40	SCH.6	70	SsNH2	100	ETA_BetaP	130	piPC7
11	AATS4v	41	SCH.7	71	SssNH	101	ETA_Beta_ns	131	TpiPC
12	AATS0p	42	VCH.6	72	SdsN	102	ETA_BetaP_ns	132	R_TpiPCTPC
13	AATSC1p	43	SP.3	73	SdS	103	ETA_dBeta	133	PetitjeanNumber
14	AATSC1i	44	SP.5	74	SsCl	104	ETA_dBetaP	134	n6Ring
15	MATS1v	45	Mv	75	minHdsCH	105	ETA_Beta_ns_d	135	nT6Ring
16	ATSC2s	46	Mpe	76	minHaaCH	106	ETA_BetaP_ns_d	136	topoRadius
17	AATSC1v	47	Mp	77	minHother	107	ETA_Eta	137	topoDiameter
18	AATSC1p	48	ECCEN	78	mindsCH	108	ETA_EtaP	138	topoShape
19	AATSC1i	49	nwHBa	79	minaaCH	109	ETA_Eta_R	139	GGI4
20	MATS1v	50	nHsNH2	80	mindssC	110	ETA_Eta_F	140	SpMax_D
21	MATS1p	51	nHdsCH	81	minaasC	111	ETA_Eta_F_L	141	SpDiam_D
22	MATS1i	52	nHaaCH	82	mindsN	112	FMF	142	SpAD_D
23	GATS1m	53	ndsCH	83	mindS	113	nHBDon_Lipinski	143	SpMAD_D
24	GATS1v	54	naaCH	84	maxwHBa	114	HybRatio	144	EE_D
25	GATS1p	55	ndssC	85	maxHdsCH	115	MIC4	145	VE1_D
26	GATS1i	56	naasC	86	maxHaaCH	116	MIC5	146	TopoPSA
27	nBondsS3	57	nsNH2	87	maxdsCH	117	nAtomP	147	AMW
28	nBondsD	58	ndsN	88	maxaaCH	118	MDEC.33	148	WTPT.3
29	nBondsD2	59	ndS	89	maxdssC	119	MDEN.11	149	WTPT.5
30	nBondsM	60	SwHBa	90	maxaasC	120	MDEN.12	150	WPATH

than data sampling techniques to enhance the classification performance of imbalanced data. It is performed by the creation of seven data frames by dividing the dataset. These data frames all have approximately equal numbers of active and inactive drug molecules [19] (see Section 4.3 for more details).

3.5. Target class

Activity is the target class that contains two instances, which are active (1) and inactive (0). Active compounds have the capability to bind with AhR and produce toxic effects by modulating its activity, and inactive compounds are nontoxic and do not bind with AhR. The intensity of the toxic effects of an active drug molecule can be analyzed by its activity score. The active drug molecules are harmful, which can disrupt the processes in the human body. Therefore, we can remove these kinds of molecules in the early stage of drug development (preclinical trials) to save the lives of animals as well as money and time. Figure 4 shows the flowchart where our proposed ensemble-based classification model performs the categorization of a new drug molecule in active and inactive categories.

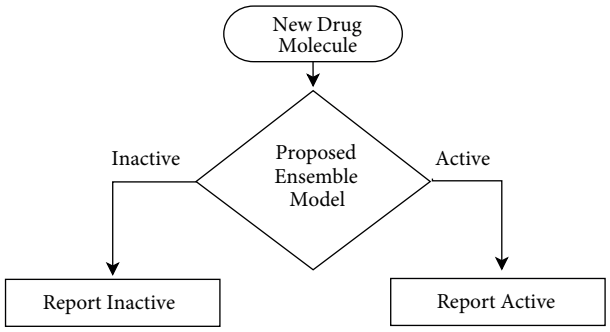

Figure 4. Flow chart.

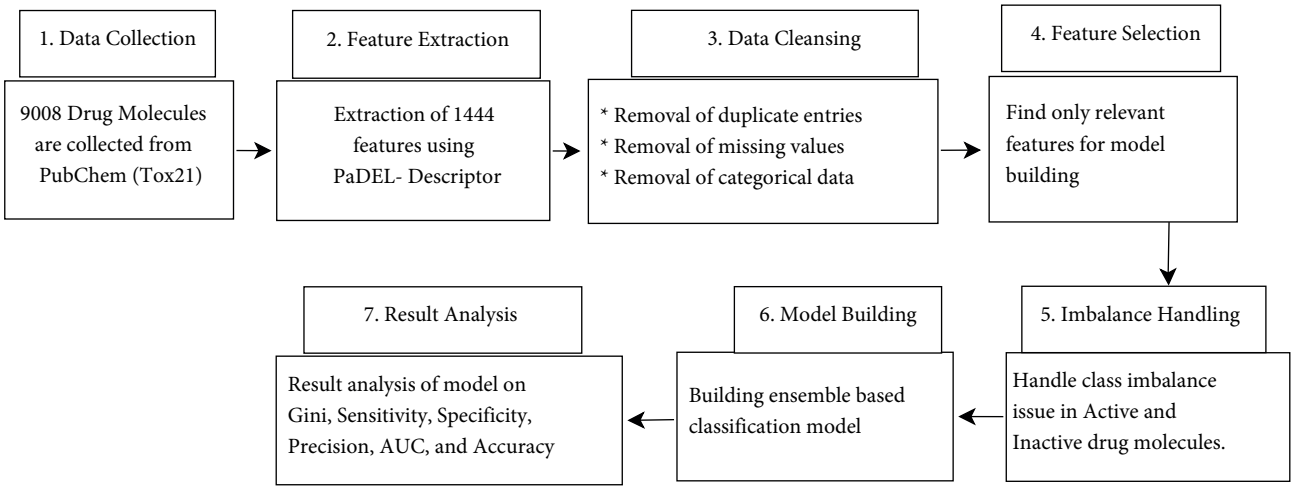

Figure 5. Methodology used.

4. Proposed ensemble-based prediction model

Ensemble learning is a technique to improve classification accuracy by combining the series of base classifiers. All the base classifiers vote for any new data tuple; based on these votes, a class label prediction is returned. The ensemble classification model can build using the same base classifiers on different splits of the same training dataset or different base classifiers on the same training dataset. Here, we used the first technique, where we created different splits of the same training dataset, and the random forest model is applied as a base classifier on all these split datasets. This approach is designed to improve our classification accuracy as well as to solve the issue of the class imbalance problem [19, 20]. Here, the random forest model is taken as a base classifier because the performance of this model is better than other models. Figure 5 shows the methodology of the proposed ensemble-based prediction model, and Figure 6 shows only the approach of ensemble learning applied in the proposed ensemble model. The following five phases show the methodology of our proposed ensemble model.

4.1. Phase 1: Dataset generation

The unprocessed dataset of AhR is obtained from the PubChem website, which is in the structure-data file (.SDF extension) format. This dataset is grouped into two directories; one of them contains only active drug molecules and the other contains only inactive drug molecules of AhR. These two directories are given as input

to the PaDEL-Descriptor software individually, which generates two .CSV extension files, one for active drug molecules and the other for inactive drug molecules. These two datasets are combined to form a resultant dataset, whose target variable is a binary class activity (see Section 3.2 for more details).

4.2. Phase 2: Data cleansing and feature selection

Data preprocessing is a technique of improving the quality of data because high-quality input data will provide highly accurate and consistent knowledge [21]. Data preprocessing can be performed by using data cleansing and feature selection. In order to remove various discrepancies, we have cleaned the dataset before model formation. Initially, we found a few corrupted and missing entries in our dataset. First, we corrected these corrupted entries and then analyzed those attributes that had missing values in their cells. We filled these missing values by the average value of that particular column. Since our dataset has 1444 features that are very high in dimensionality and we applied the feature dimensionality reduction method, which reduces the features as well as the execution time of the classifiers in machine learning [22]. Here, the Boruta algorithm is applied to the dataset, which returns only 150 features out of 1444 features (see Section 3.3 for more details).

4.3. Phase 3: Class imbalance handling

The dataset found from PubChem is highly imbalanced, as it has a total of 9008 drug molecules of which 1063 are active and 7945 are inactive. To resolve this problem, we primarily segregate the active and inactive drug molecules of the dataset, where we find that the number of inactive drug molecules is almost seven times higher than the active drug molecules. Therefore, we divide the dataset of inactive drug molecules into seven data frames. Subsequently, the copy of all active drug molecules is added in all seven data frames so that all the data frames have approximately equal numbers of active and inactive drug molecules. Now these seven data frames are different and balanced datasets available for model building by using ensemble learning.

4.4. Phase 4: Classification model building using ensemble learning

Now we have seven balanced and different datasets. We have trained each dataset at 70% data using the base classifier of random forest and combined all these classifiers using the ensemble learning approach.

4.5. Phase 5: Voting system

An ensemble model based on the votes of seven random forest models is created, which is the proposed ensemble model. Subsequently, we prepared a single testing dataset, which is the combination of 30% tuples of each dataset frame. The performance of the proposed ensemble model is evaluated on this testing dataset (see Figure 6). Now this model can be used to predict the activity of any new drug molecule of AhR.

5. Random forest model

The prediction of activity of any drug molecule is important while deciding its toxic effects on human health. The results of our proposed ensemble model using the random forest model are better in comparison to other existing models of classification. Each model has its various parameters where some parameters have their constant values, while others can take different values. We can improve the performance of models by manipulating these values, and this process is called the tuning of parameters. Table 5 shows various models with their corresponding packages, methods, and tuned parameters. Random forest with its tuned parameters has been

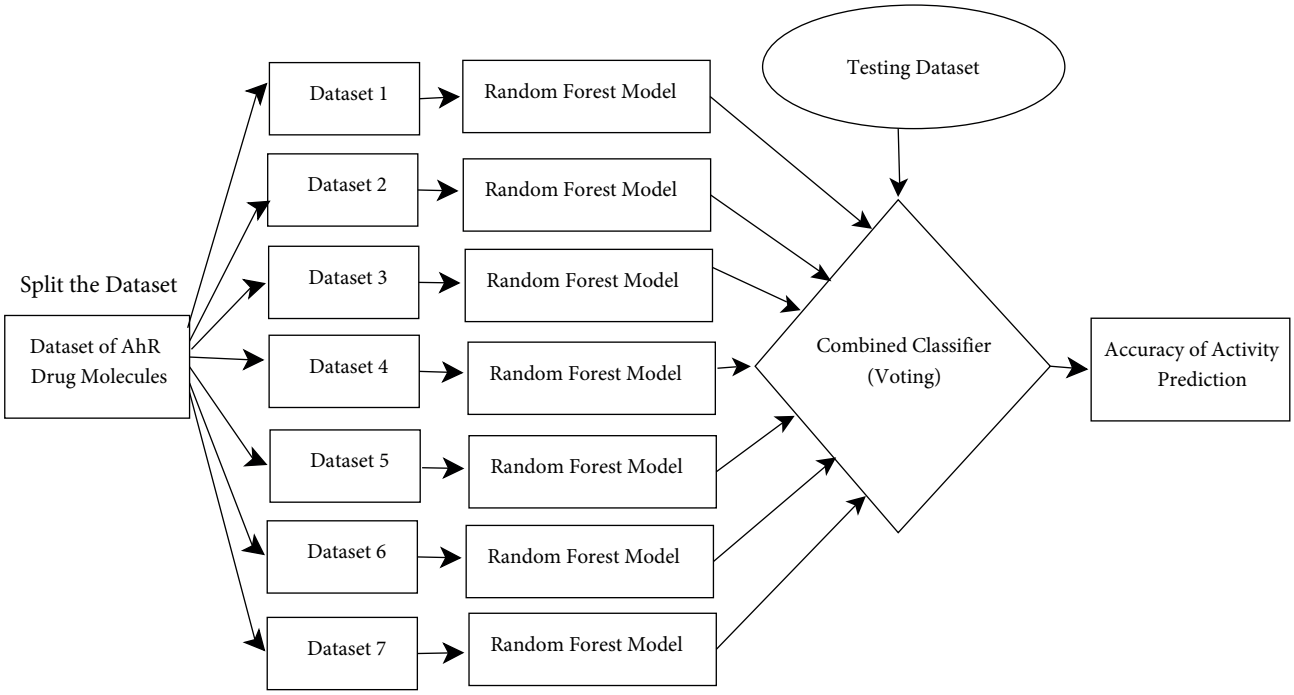

Figure 6. Ensemble method for activity prediction.

used in the proposed ensemble model, and other models with tuned parameters have been used for comparison with the proposed ensemble model. All the models are implemented in R, under the GNU general public license. The random forest method has "mtry" and "ntree" parameters for tuning, where mtry shows the number of variables randomly sampled as candidates at each split and ntree is the number of trees to grow. The value of ntree should be large enough to ensure that every input row gets predicted at least a few times. Therefore, we optimized the performance of the random forest model by setting the values of mtry = 2 and ntree = 500. We used the random forest method as randomForest(formula, trainDataset, ntree=500, mtry=2), where its formula comprised 150 important features and target class as shown below:

$$Activity \sim f(AMR + naAromAtom + nAromBond + \cdots + WTPT.5 + WPATH). \quad (1)$$

Table 4 shows all the features used in this formula. Random forest is an aggregate classifier, which is the collection of several decision trees. Random forest is itself an ensemble-based model, where each tree votes and the most popular class is returned during classification [23, 24]. If n is the number of records and d is the depth of the tree, then the time complexity of the random forest algorithm is O(ntree *mtry *d *n) and the space complexity of random forest algorithm is O(n*d). Therefore, we can say that the random forest model depends on the depth and size of the decision tree [17].

6. Binary classification-based performance evaluation parameters

Performance comparison for the various binary classification model is generally performed on some specific parameters, which are the Gini coefficient, sensitivity, specificity, precision, AUC, and accuracy. These parameters are explained below.

Table 5. Machine learning models used and their tuning parameters.

Model	Required package	Method	Tuning parameters
Random forest (RF)	randomForest	randomForest	mtry=2, ntree=500
Decision tree (DT)	rpart	rpart	usesurrogate=0, maxsurrogate=0
Support vector machine(SVM)	kernlab	ksvm	kernel=rbfdot, type=C -svc
Neural network (NN)	nnet	nnet	size=10
Linear model (LM)	none	lm	method = "qr"

6.1. Gini coefficient

The Gini coefficient is used to measure the distribution inequality of data [25]. Gini values range between 0 and 1. The 0 and 1 values of the Gini coefficient indicate perfect equality of data and perfect inequality of data, respectively. Assuming that a model M has a Gini coefficient of 0.6 and model D has a Gini coefficient of 0.45, then model M is considered a productive model in contrast to model D.

6.2. Sensitivity

Sensitivity (Sens) is also known as the true positive rate (recognition) or recall [23]. It is the ratio of actual positives that are correctly identified as positives by the classifier. It is computed as:

$$Sensitivity = \frac{TP}{TP + FN}. \tag{2}$$

6.3. Specificity

Specificity (Spec) is also known as the true negative rate [23]. It is the ratio of actual negatives that are correctly identified as negatives by the classifier. It is computed as:

$$Specificity = \frac{TN}{TN + FP}. \tag{3}$$

6.4. Precision

Precision can be thought of as a measure of exactness, i.e. what percentage of tuples labeled as positive are actually such [23]. It is computed as:

$$Precision = \frac{TP}{TP + FP}. \tag{4}$$

6.5. AUC

The area under the curve (AUC) measures the quality of the classifier. The receiver operating characteristic (ROC) curve is a curve drawn between the true positive rate (TPR) and false positive rate (FPR). We can find these parameters with the confusion matrix. The area under the ROC is called the AUC. The AUC value ranges between 0 and 1. The quality of a model is outstanding if it has AUC close to 1. A model with a high AUC value as compared to another model is considered an efficient model [25].

6.6. Accuracy

Accuracy is the most important criterion for measuring the exactness of any classifier [24]. Accuracy can be computed as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100, \tag{5}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

7. Result analysis, comparison, and validation

The Gini coefficient, sensitivity, specificity, precision, AUC, and accuracy are the model performance evaluation parameters for any binary classification model, which are described in Section 6. These parameters evaluate activity prediction for our proposed ensemble model as well as for some existing models. The comparative performance of our proposed ensemble model with some existing classification models is shown in Table 6. The Gini coefficient, specificity, sensitivity, precision, AUC, and accuracy of our proposed ensemble model is 0.932, 0.967, 0.936, 0.964, 0.966, and 93.76%, respectively. The results show that our proposed ensemble model has outperformed the other models for the 30% testing dataset of AhR. The random forest, decision tree, support vector machine, neural network, and linear model [18] are the existing models used for comparison.

Table 6. Performance comparison of proposed ensemble model with existing classification models.

Decision method	Gini coefficient	Sensitivity	Specificity	Precision	AUC	Accuracy(%)
Proposed ensemble model	0.932	0.967	0.936	0.964	0.966	**93.76**
Random forest	0.916	0.953	0.979	0.978	0.953	91.45
Decision tree	0.901	0.915	0.931	0.942	0.911	90.21
Support vector machine	0.896	0.805	0.839	0.919	0.893	82.40
Neural network	0.813	0.895	0.884	0.834	0.837	82.34
Linear model	0.817	0.531	0.513	0.723	0.545	77.26

7.1. K-fold cross-validation

The k-fold cross-validation approach partitions the dataset into k equally sized subsets or "folds". During each execution, one of the partitions is chosen for testing, while the rest of the segments are used for training. This procedure is repeated k times so that each partition is used for testing exactly once. In each fold, random data are provided for training and testing to measure the robustness of the model [26]. Here, we have used a 7-fold cross-validation method for activity prediction, where the result of cross-validation (Figure 7) shows the consistent performance for all the evaluation parameters of the proposed ensemble model [14]. The value of k has been selected in such a manner that each training and testing partition of the broader dataset are large enough to represent it statistically, but there is no formal rule to choose the value of k. In this case, the dataset is divided into seven data frames, which have equal numbers of drug molecules. At the value of k = 7, we can take collectively any six data frames for the training of the model and the remaining one data frame for the testing of the model. Table 7 describes the accuracy of the proposed ensemble model by applying 7-fold cross validation one time.

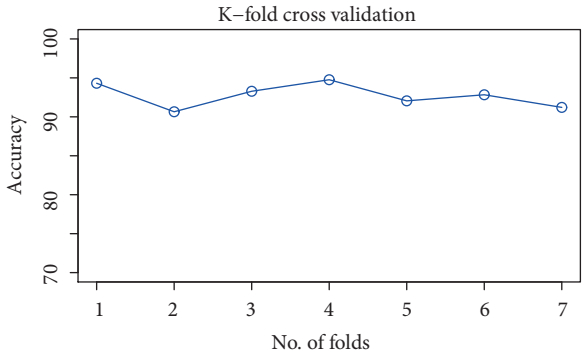

Figure 7. K-fold cross-validation for activity prediction.

Table 7. Accuracy in 7-fold cross-validation of proposed ensemble model.

Folds	Accuracy
1	94.31
2	90.65
3	93.28
4	94.76
5	92.05
6	92.83
7	91.21

7.2. Validation of the proposed ensemble model

Validation of the proposed ensemble model means that we are testing this model on some new drug molecules, which are neither part of the training dataset nor part of the testing dataset. If the prediction accuracy of this model for these new drug molecules is similar to our testing dataset to some extent, then we can say that our proposed ensemble model has been validated. Here, we have validated our proposed ensemble model on some AIDS therapy drug molecules and some androgen receptor drug molecules, which are not the part of the actual dataset. Nonnucleoside and nucleoside reverse transcriptase inhibitors (NNRTIs) are the first types of drug available to treat HIV that block HIV enzymes, and corresponding to nevirapine (NVP), delavirdine (DLV), efavirenz (EFV), and rilpivirine (RPV), which are the essential drugs for AIDS therapy [27]. These drugs show potent anti-HIV-1 activity and modest toxicity. Nevirapine is linked with hepatic toxicity, and it causes liver injury during therapy, which is also followed by fever, oral lesions, blistering, conjunctivitis, swelling, and muscle or joint aches. The major toxicity of delavirdine is skin rashes. Efavirenz has fatal severe side effects on the liver and the central nervous system. Rilpivirine also has some side effects, which are sores in the mouth and redness or swelling of the eyes, face, lips, mouth, tongue, or throat. Now we have two-dimensional structures of all nine drug molecules, which are downloaded as structure-data files from the PubChem database, and their molecular descriptors have been extracted with the help of PaDEL-Descriptor. Now we apply the proposed ensemble model for activity prediction of NVP, DLV, EFV, and RPV and five drug molecules of the androgen receptor. The output of the proposed ensemble model is summarized in Table 8. The results of the table show that the drug molecules predicted to be active are actually found to be active and the drug molecules predicted inactive are actually found to be inactive. These correct predictions of all the drug molecules show the validity

of the proposed ensemble model. Figure 8 shows the validation process of the proposed ensemble model on the four drug molecules of AIDS therapy.

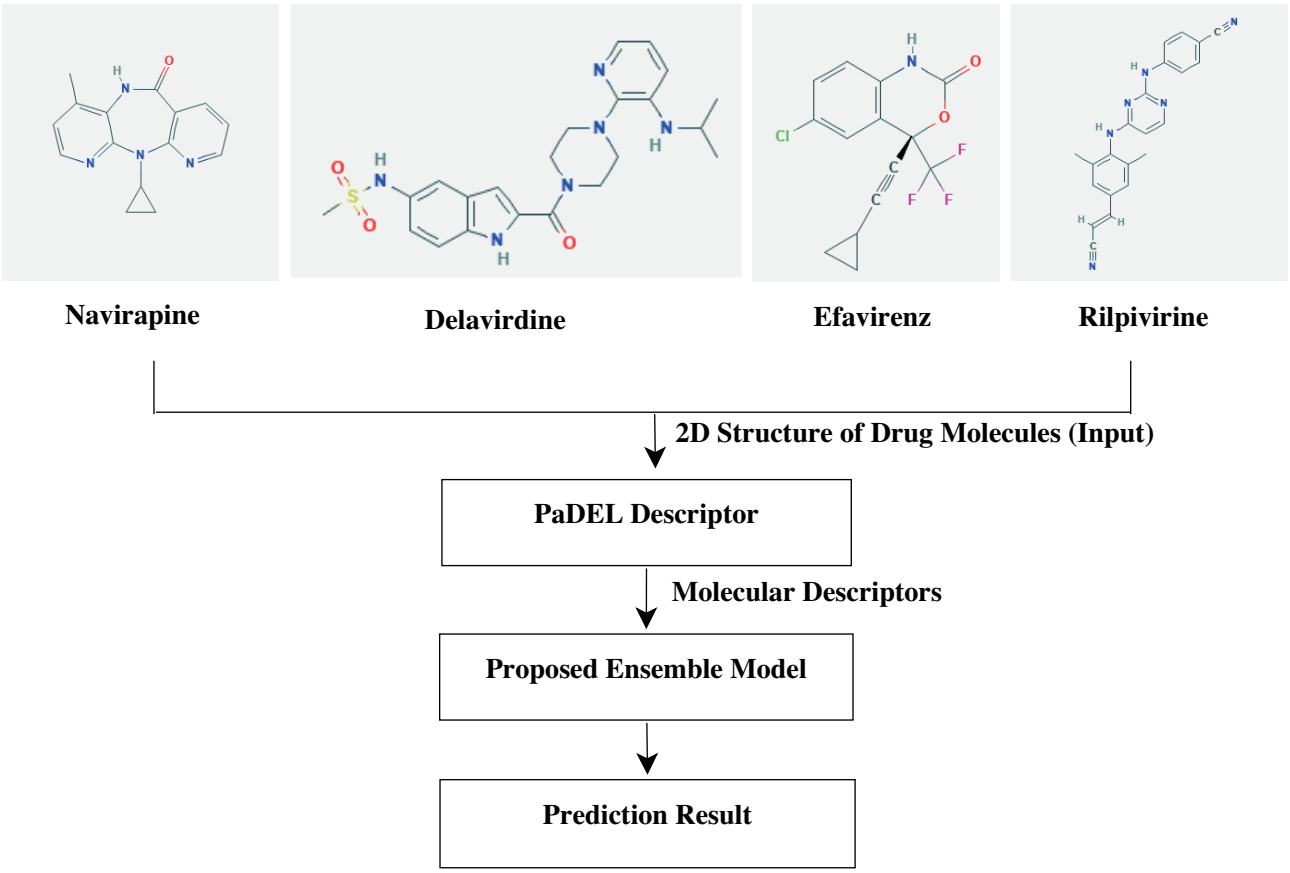

Figure 8. Activity prediction of AIDS therapy drug molecules using proposed ensemble model.

Table 8. Validation of proposed ensemble model on some AIDS therapy and androgen receptor drug molecules.

Target drug molecule	Actual class	Predicted class	Accuracy(%)
Nevirapine (NVP)	1	1	100%
Delavirdine (DLV)	1	1	100%
Efavirenz (EFV)	1	1	100%
Rilpivirine (RPV)	1	1	100%
NCGC00261776-01	1	1	100%
NCGC00261900-01	0	0	100%
NCGC00260869-01	0	0	100%
NCGC00261842-01	0	0	100%
NCGC00261926-01	0	0	100%

8. Conclusion

In this paper, we have proposed an ensemble-based efficient computational method, which has solved the problem of toxicity prediction of drug molecules that activate the aryl hydrocarbon receptor signaling pathway. It is a decision support system to predict the toxicity of unknown drug molecules that act on AhR, where we can get the results of toxicity prediction by uploading the SDF of any single drug molecule. The target class for the toxicity prediction is activity. The dataset used in this study is very high in features and extremely imbalanced. Initially, we have performed feature selection by the Boruta method and balanced the dataset by using an ensemble learning approach. Here, the ensemble method is used for dual purposes. First, it resolves the problem of class imbalance, and second, it is used for classification. The proposed ensemble model has been evaluated with various performance parameters, i.e. the Gini coefficient, sensitivity, specificity, precision, AUC, and accuracy, for the activity prediction. Through intensive experiments, it is found that our proposed ensemble model, in spite of having a highly imbalanced dataset, has given better accuracy than other existing models, which are random forest, decision tree, support vector machine, neural network, and linear model, and its performance is nearly linear in k-fold cross-validation. Finally, to prove the validity of the proposed ensemble model, we have tested it on AIDS therapy drug molecules and some drug molecules of the androgen receptor, where we found 100% accuracy. The limitation of this proposed model is that it can predict the activity of only those kinds of drug molecules on which it has been trained. This model cannot recognize different types of drug molecules' activity, because these drug molecules can have different physicochemical properties or features.

Acknowledgments

This research was funded by DST-SERB (Science and Engineering Research Board, Government of India) under the "Early Career Research Scheme" with File No. ECR/2015/000150/LS. We gratefully acknowledge the support of the NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research. We are extremely thankful to the anonymous reviewers for their various insightful comments and suggestions.

References

[1] Rastogi SC, Mendiratta N, Rastogi P. Bioinformatics Methods and Applications, Genomics, Proteomics and Drug Discovery. New Delhi, India: Prentice-Hall, 2008.

[2] Chen YZ, Ung CY. Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. Journal of Molecular Graphics and Modelling 2001; 20 (3): 199-218. doi: 10.1016/S1093-3263(01)00109-7

[3] Mayr A, Gunter K, Unterthiner T, Hochreiter S. Deep tox: Toxicity prediction using deep learning. Frontiers in Environmental Science 2016; 3: 80. doi: 10.3389/fenvs.2015.00080

[4] Toropov AA, Toropova AP, Raska JI, Leszczynska D, Leszczynski J. Comprehension of drug toxicity: software and databases. Computers in Biology and Medicine 2014; 45: 20-25. doi: 10.1016/j.compbiomed.2013.11.013

[5] Tannenbaum J, Bennett BT. Russell and Burch's 3Rs then and now: the need for clarity in definition and purpose. Journal of the American Association for Laboratory Animal Science 2015; 54 (2): 120-132.

[6] Khan MTH. Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. Current Drug Metabolism 2010; 11 (4): 285-295. doi: 10.2174/138920010791514306

[7] Stockinger B. Beyond toxicity: aryl hydrocarbon receptor-mediated functions in the immune system. Journal of Biology 2009; 8 (7): 61. doi: 10.1186/jbiol170

[8] Basak SC, Mills D, Mumtaz MM, Balasubramaniun K. Use of topological indices in predicting aryl hydrocarbon receptor binding potency of dibenzofurans: a hierarchical QSAR approach. Indian Journal of Chemistry 2003; 42A: 1385-1391.

[9] Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nature Reviews Drug Discovery 2004; 3: 711–715. doi: 10.1038/nrd1470

[10] Pipero EL, Koehler K, Chana A, Benfenati E. Virtual screening for aryl hydrocarbon receptor binding prediction. Journal of Medicinal Chemistry 2006; 49 (19): 5702–5709. doi: 10.1021/jm060526f

[11] Cassano A, Manganaro A, Martin T, Young D, Piclin N et al. CAESAR models for developmental toxicity. Chemistry Central Journal 2010; 4 (S1): S4. doi: 10.1186/1752-153X-4-S1-S4

[12] Drwal M, Siramshetty VB, Banerjee P, Goede A, Preissner R et al. Molecular similarity-based predictions of the Tox21 screening outcome. Frontiers in Environmental Science 2015; 3: 54. doi: 10.3389/fenvs.2015.00054

[13] Stefaniak F. Prediction of compounds activity in nuclear receptor signaling and stress pathway assays using machine learning algorithms and low-dimensional molecular descriptors. Frontiers in Environmental Science 2015; 3: 77. doi: 10.3389/fenvs.2015.00077

[14] Capuzzi SJ, Politi R, Isayev O, Farag S, Tropsha A. QSAR modeling of Tox21 challenge stress response and nuclear receptor signaling toxicity assays. Frontiers in Environmental Science 2016; 4: 3. doi: 10.3389/fenvs.2016.00003

[15] Yap CW. PaDEL-Descriptor: an open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry 2010; 32 (4): 1466-1474. doi: 10.1002/jcc.21707

[16] Kursa MB, Rudnicki WR. Feature selection with the boruta package. Journal of Statistical Software 2010; 36 (11): 1-13. doi: 10.18637/jss.v036.i11

[17] Hooda N, Baba S, Rana PS. B2FSE framework for high dimensional imbalanced data: a case study for drug toxicity prediction. Neurocomputing 2018; 276: 31-41. doi: 10.1016/j.neucom.2017.04.081

[18] Feng W, Huang W, Ren J. Class imbalance ensemble learning based on the margin theory. Applied Sciences 2018; 8 (815): 1-28. doi: 10.3390/app8050815

[19] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 2002; 16: 321–357. doi: 10.1613/jair.953

[20] Jiang J, Wang N, Chen P, Zhang J, Wang W. DrugECs: An ensemble system with feature subspaces for accurate drug-target interaction prediction. BioMed Research International 2017; 2017: 6340316. doi: 10.1155/2017/6340316

[21] Arumugam A. A predictive modeling approach for improving paddy crop productivity using data mining techniques. Turkish Journal of Electrical Engineering & Computer Sciences 2017; 25: 4777-4787. doi: 10.3906/elk-1612-361

[22] Takci H. Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences 2018; 26: 1-10. doi: 10.3906/elk-1611-235

[23] Han J, Kamber M, Pei J. Data Mining Concepts and Techniques. Waltham, MA, USA: Morgan Kaufmann Elsevier, 2012.

[24] Tan PN, Kumar V, Steinbach M. Introduction to Data Mining. Manesar, INDIA: Pearson Education, 2016.

[25] Khanna D, Rana PS. Multilevel ensemble model for prediction of IgA and IgG antibodies. Immunology Letters 2017; 184: 51–60. doi: 10.1016/j.imlet.2017.01.017

[26] Rana PS, Sharma H, Bhattacharya M, Shukla A. Quality assessment of modeled protein structure using physicochemical properties. Journal of Bioinformatics and Computational Biology 2015; 13 (2): 1550005. doi: 10.1142/S0219720015500055

[27] Usach I, Melis V, Peris JE. Non-nucleoside reverse transcriptase inhibitors: a review on pharmacokinetics, pharmacodynamics, safety and tolerability. Journal of the International AIDS Society 2013; 16 (1): 18567. doi: 10.7448/IAS.16.1.18567