Toxicity Prediction of Pre-Clinical Trial Drugs using Physicochemical Properties and Computational Intelligence Approaches

A Thesis

submitted for the award of the degree of

Doctor of Philosophy

in

Computer Science and Engineering Department

Submitted by

Vishan Kumar Gupta
(Reg no: 951503003)

Under the Guidance of

Dr. Prashant Singh Rana
Assistant Professor
Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala,
Punjab - 147004, India
March 2020

Certificate

I, Vishan Kumar Gupta, Registration No. 951503003, hereby declare that the work which is being presented in this thesis entitled, "Toxicity Prediction of Pre-Clinical Trial Drugs using Physicochemical Properties and Computational Intelligence Approaches" in partial fulfillment of the requirement for the award of "Doctor of Philosophy" submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, Punjab is an authentic record of my own work carried out under the supervision of Dr. Prashant Singh Rana, refers other research works which have been duly listed in the reference section. The matter presented in this thesis has not been submitted for the award of any other degree of this or any other university.

(Vishan Kumar Gupta)
Registration No. 951503003

This is to certify that the above statements made by the candidate is correct and true to the best of my knowledge.

Verified by:

(Dr. Prashant Singh Rana)
Supervisor
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology, Patiala, Punjab, India

.....dedicated to my grandparent Late. Sh. Siyaram Gupta and Late. Smt. Shanti Devi

Acknowledgements

First, I would like to express my deep gratitude to my supervisor **Dr. Prashant Singh Rana** for his invaluable advice and encouragement at every step of my Ph.D. program. Without his unfailing support and belief in me, this thesis would not be possible. His contribution to this thesis goes well beyond his role as an academic supervisor and includes constant support on a personal level without that this journey may not be completed. I am truly grateful for this guidance. He is a great mentor for my life also.

I would like to express my gratitude to our director **Prof. Prakash Gopalan** for providing a research environment which helps me to achieve my stated objectives. My sincere thanks to Head of the Computer Science and Engineering Department Prof. Maninder Singh and my research committee members Dr. Rajesh Sharma, Dr. Vijay Kumar Chahar, and Dr. M.D. Singh for their constant guidance and motivation.

I would like to give special thanks to IMS Engineering College, Ghaziabad, who supports me and give some time for the completion of this research work.

I would like to thanks to Dr. Avdhesh Gupta, Dr. Narendra Rathore, and Mr. Priyaranjan Kumar for their support in various aspect of this research work. I would also like to thank my colleagues Mr. Dinesh Kumar, Mr. Ankit Kumar, and Mr. Mukesh Kumar Singh for helping and encouraging me.

Finally, I would like to express my sincere and deep gratitude to my parents (Mr. Ramsewak Gupta and Mrs. Radha Gupta), my sister Arti Garg and my brother Dhiraj Gupta for their love, encouragement, care, and support. Finally, I would like to thanks to my wife Himani Gupta for having faith in me and supporting me at every step, without her support, I could not complete my Ph.D. program and finally a lot of love to my son Sarthak and daughter Paakhi.

Abstract

Development of quantitative structure activity relationships (QSARs), quantitative structure property relationships (QSPRs), and quantitative structure toxicity relationships (QSTRs) have been practiced for the prediction of various toxicities of drug molecules in terms of their activity, activity score, potency, and efficacy. These predictions are based on the in silico toxicity prediction techniques, which are essential for reducing animal testing (in vivo), less time-consuming and 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 their 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 various nuclear receptor (NR) signalling pathways like androgen receptor (AR), estrogen receptor (ER), and aryl hydrocarbon receptor (AhR), and various stress response (SR) signalling pathways like antioxidant response elements (ARE).

The pharmaceutical data exploration laboratory (PaDEL) software is used for extracting the features of drug molecules. Aryl hydrocarbon receptor contains 9008 drug molecules where 1063 are active, and 7945 are inactive, the estrogen receptor dataset has 8481 drug molecules where 1084 are active, and 7397 are inactive, the androgen receptor dataset has 10273 drug molecules where 461 are active, and 9812 are inactive, and the antioxidant response elements dataset has total 7439 drug molecules, of which 1147 are active and 6292 are inactive. Initially, the class imbalance is resolved using SMOTE algorithms for the ER dataset, and we have divided the dataset into equal size of data frames which have an equal number of active and inactive drug molecules for the dataset of AR, AhR, and ARE. Feature selection is performed by Boruta algorithm, CFS algorithm, Gini importance, and Random forest importance algorithm. It is found that the extended topochemical atom (ETA) descriptors, electro-topological state descriptors, Crippen's logP, and Molar refractivity (MR) are quite rich in chemical information to encode the structural features that contribute to the toxicities and these indices may be used in combination with other topological and physicochemical descriptors for the development of predictive QSAR model.

Initially, five classification methods are trained on the dataset of ER for activity, activity score, potency, and efficacy prediction and it is found that random forest is having the

best accuracy in comparison of other models. Similarity, a multilevel ensemble model is proposed for the dataset of AR, where our proposed multilevel ensemble model is outperformed in comparison to other models. An ensemble model based on the votes of random forest is proposed for the prediction of toxicity of AhR drug molecules, where our proposed ensemble model is performed better instead of other models. An ensemble model based on the votes of AdaBoost, random forest, decision tree and support vector machine is proposed for the prediction of toxicity of the ARE signaling pathway dataset, where our proposed ensemble model outperformed other models. The K-fold cross-validation is performed to measure the consistency of all proposed models for all the target classes. Finally, we have proved the validity of all the proposed models on some AIDS Therapy's, general food additives, cosmetics, detergents, preservatives, luciferase-tagged ATAD5, and some other similar kinds of drug molecules.

Keywords: Androgen Receptor; Machine Learning; Molecular Descriptor; Random Forest; Activity; Activity Score; Potency; Efficacy; Toxicity; Feature Selection; Ensemble Learning; PaDEL; CFS; Boruta; Information Gain; Nuclear Receptor; Stress Response; Aryl Hydrocarbon Receptor; Estrogen Receptor; Antioxidant Response Element; Toxicity; Class Imbalance.

Table of Contests

List of Figures

List of Tables

List of Abbreviations

AdaBoost	Adaptive Boosting
AhR	Aryl Hydrocarbon Receptor
AR	Androgen Receptor, full
AR-LBD	Androgen Receptor, Ligand-Binding Domain
ARE	Antioxidant Responsive Element
AUC	Area Under the Curve
CFS	Correlation-based Feature Selection
CSV	Comma-Separated Values
DCPA	Dimethyltetrachloroterephthalate Acid
DLV	Delavirdine
DT	Decision Tree
EDTA	Ethylenediaminetetraacetic Acid
EFV	Efavirenz
EML	Ensemble Machine Learning
ER	Estrogen Receptor Alpha, full
ER-LBD	Estrogen Receptor Alpha, Ligand-Binding Domain
ETR	Etravirine
GBM	Generalized Boosted Regression Modeling
GLM	Generalized Linear Model
HSE	Heat Shock Factor Response Element
HTS	High Throughput Screening
LC	Lethal Concentration
LD	Lethal Dose
LM	Linear Model
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MMP	Mitochondrial Membrane Potential
MOL	Molecular data file
MSG	Monosodium Glutamate
NCBI	National Center for Biotechnology Information
NFL	No Free Lunch
NIH	National Institute of Health
NN	Neural Network
NNRTIs	Non-nucleoside and Nucleoside Reverse Transcriptase Inhibitors

Nrf2	Nuclear factor erythroid 2-related factor 2
NVP	Nevirapine
PaDEL	Pharmaceutical Data Exploration Laboratory
PDB	Protein Data Bank
PPAR	Peroxisome Proliferator-Activated Receptor Gamma
QSAR	Quantitative Structure Activity Relationships
QSPR	Quantitative Structure Property Relationships
QSTR	Quantitative Structure Toxicity Relationships
RCSB	Research Collaboratory for Structural Bioinformatics
RF	Random Forest
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristics
RPV	Rilpivirine
$\mathbf{R^2}$	Coefficient of Determination
SDF	Structure Data File
Sens	Sensitivity
Spec	Specificity
SMOTE	Synthetic Minority Over sampling Technique
SMILES	Simplified Molecular Input Line Entry System
SVM	Support Vector Machine
Tox21	Toxicology in the 21st Century

Chapter 1

Introduction

This chapter presented the introduction of the research work. This chapter is started with the drug design process. In this process, toxicity is a concerning issue; therefore, we described toxicity and its causes. In this research, we mainly predicted toxicity of pre-clinical trial drugs using physicochemical properties and supervised learning. So, different physicochemical properties, various machine learning approaches, and performance evaluation parameter are described. We used the datasets of nuclear receptors and stress response pathways, and target classes are activity, activity score, potency, and efficacy. The brief descriptions about nuclear receptors and stress response pathways are also discussed in this chapter, and the target classes are described in subsequent chapters. This chapter is ended with the thesis organization.

1.1 Drug Design

Drugs are the human-made endogenous molecule that inhibits or activates the function of a biological molecule, such as protein, which in turn results in a therapeutic benefit to the patient. The drug is a combination of chemicals that prevent disease or assist in restoring the health of the human being. One of the primary means of conserving human health is maintaining by administrating the small molecule, which is the chemical construction of the drug. Drug designing is an inventive process; a new medication is found based on the knowledge of biological target [3]. The more exact term for designing a molecule is ligand design, which binds precisely to its target. Before a ligand design, we need to optimize many properties of drugs such a drug side effects, metabolic half-life, etc., for the safety and effectiveness of it. Organically drug is a small molecule (a lower molecular weight of < 900 Daltons).

Successful drug molecules bind to specific components (usually proteins) of the target cells. After it, these cells activate or deactivate those components, which lead to modifying cells or destroying them. Introducing a novel drug is not an easy job. The drug development industries have been spending enormous efforts and time not only to discover new drugs but also to improve existing medicines. The process of launching a new

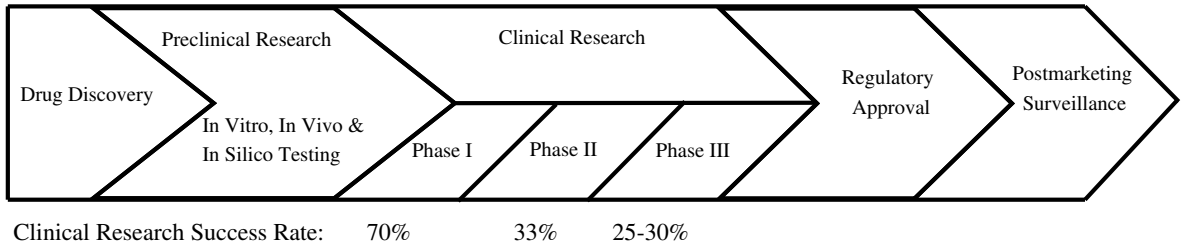


Figure 1.1: Overview of drug development process

drug in the market is very time-consuming, complicated, expensive, and has a high rate of failure. Averagely, the discovery of new drug development takes 10-15 years and costs 400-800 million US dollars [4].

In contemporary drug discovery, a new therapy must undergo extensive testing and meet strict regulatory requirements before it can be approved and made available to patients [5]. Figure 1.1 shows the major milestones in drug development.

Step 1: Discovery and development

The drug development process typically begins in the laboratory, where researchers study the specific methods and pathways involved in the development of the disease. Insights from this study enable researchers to identify or develop targets and associated therapies that may be able to stop or slow the progress of a disease [5].

Step 2: Pre-Clinical research

Before testing a new therapy in humans, researchers must perform pre-clinical studies to evaluate its safety and potential toxicity. The results of pre-clinical research decide whether a therapy should be tested in humans or not [5]. The following methods can perform pre-clinical testing:

- **In vivo**: "In the living organism," experiments that are done in the body of a living organism.

- **In vitro**: "In glass" or "in a test tube," experiments that are performed in a plastic or glass vessel in the laboratory.

- **In silico**: "Performed on computers" or "via computer simulations," The in silico drug designing techniques identify the drug target molecules by employing the bioinformatics tools.

Step 3: Clinical research

Clinical research is performed in people, which are performed in a succession of four phases:

Phase I: Clinical trials are performed in small numbers (20-100) of healthy people or volunteers with the disease over a short period (i.e., several months). The primary purpose is to evaluate the safety and amount of dosage that can be given before side effects become intolerable or dangerous.

Phase II: Clinical trials are performed in more significant numbers of people with the disease (100-300 volunteers) over a more extended period (several months to 2 years). The primary purpose is to evaluate efficacy and side effects.

Phase III: Clinical trials are performed in larger numbers of people with the disease (300-3000 volunteers) over several years. The purpose of this study is to determine whether the new therapy offers a clinical benefit in a population of people, which is intended to treat. Phase III trials compare the effectiveness and safety of the new therapy against the current standard treatment.

Step 4: FDA drug review

If a drug developer has evidence from its early tests and pre-clinical and clinical research that a drug is safe and effective for its intended use, the company can apply to market the medicine. The FDA is a review team that thoroughly examines all submitted data on the drug and makes a decision to approve or not to approve it.

Step 5: FDA post-market safety monitoring

Post-marketing surveillance trials are performed, where the FDA monitors all the drugs and device safety once products are available for public use. The purpose of these studies is to monitor the safety of new therapies up to several thousand peoples over the disease for a longer period [5].

Figure 1.2 shows the more general flow of drug discovery and development from the identification of disease to FDA approval of the medicine. After a pre-clinical trial, there is a need to file IND (Investigational New Drug Process), where developers must include animal study data, toxicity (side effects that cause significant harm) data, and manufacturing information. Then the drug moves to a clinical trial, after clinical trial, filing an NDA (New Drug Application) is required, which tells the full story of a drug. The purpose of NDA is to demonstrate that a drug is safe and effective for its intended use in the population studied. Once the FDA receives an NDA, the review team decides whether it is complete or not. In the case where it is not complete, the review team can refuse to file the NDA. If it is complete then review team has 2 to 3 years to decide on

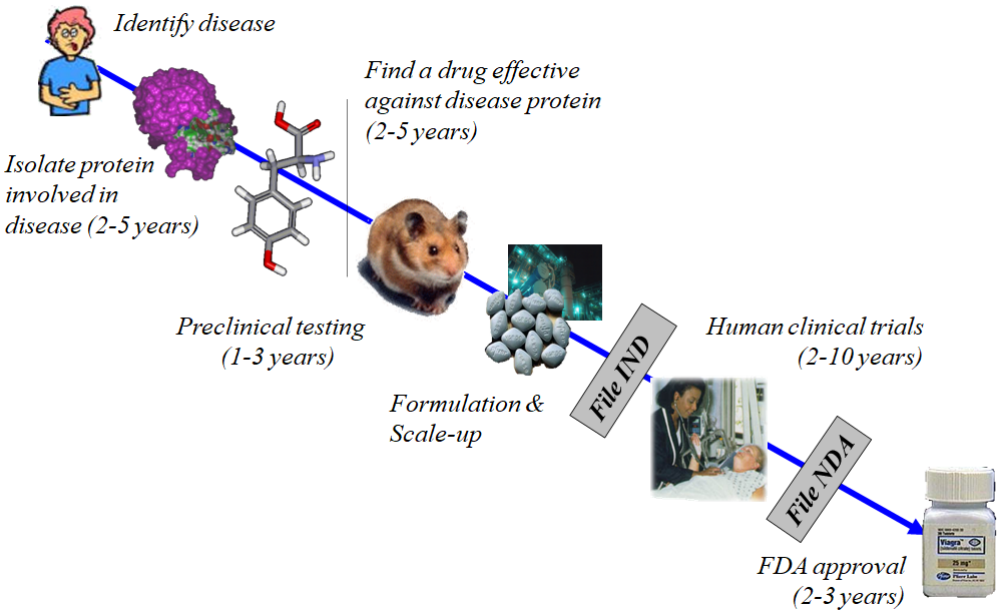


Figure 1.2: In vivo drug discovery process [1]

the approval of the drug [5].

To decide whether a particular drug will become a drug like a candidate, practically, we should use fast and cheap methods that determine whether the drug candidate is or is not suitable for market sale. The performance of computational methods (in silico) for toxicity testing is accessed to judge their potential to reduce in vitro experiments and in vivo experiments.

To reduce the use of in vivo and in vitro experiments during the pre-clinical trial, the goal of our research is to propose a computational method (in silico) for toxicity testing using machine learning models and physicochemical properties of drug molecules. Along with the individual machine learning model, the ensemble framework is more beneficial to create decision-making algorithms for the toxicity assessment of various drug molecules. We have evaluated the performance of our models on different performance metrics like sensitivity, specificity, ROC curve, accuracy, correlation, coefficient of determination, RMSE, etc.

1.2 Physicochemical Properties and Molecular Descriptors

In general "physicochemical" is the study of the physical-chemical properties of distinct drug molecules. There are the following properties [6]:

- **General:** Boiling point, melting point, vapour pressure, nernst partition coefficient (P), heat of reaction, activation energy, etc.
- **Quantum chemical:** Bond energy, atomic charge, electron density, molecular reactivity, resonance energy, etc.
- **Steric:** Molecular volume, vander waals volume, vander waals radius, shape and surface area, etc.
- **Structural:** Number of hydrogen bonds, extent of branching, number of rings, etc.
- **Electric:** Dielectric constant, ionization potential, charge ratio, dipole moment, etc.

Physicochemical properties are usually taken as measurement testing, and these are depending on the state of drugs. For example, the drugs which can be absorbed in the fat are different from those drugs that can be absorbed in the water. These are the implicit properties of the drugs, which depends on its molecular structure.

For each toxicant, it is necessary to select a set of descriptors related to a particular mechanism of activity. However, from the toxicokinetic point of view, two parameters are of general importance for all toxicants:

The Nernst partition coefficient (P) establishes the solubility of toxicant molecules in the two-phase octanol-water system. This parameter affects the distribution and accumulation of toxicant molecules in the organism [6].

The dissociation constant (pKa) defines the degree of ionization of molecules of a toxicant into charged cations and anions at a particular pH. This constant represents the pH at which 50% ionization is achieved [6].

There are more than 8000 physicochemical properties, so the scope of physicochemical properties in pharmacology has much projection in drug-like chemical space. According to Todeschini and Consonni, the molecular descriptor is the final result of the logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number [7]. There are a variety of

commercial and open-source tools to get out molecular descriptors based on physicochemical properties. We are using the pharmaceutical data exploration laboratory (PaDEL) software to calculate molecular descriptors. Table 1.1 shows various molecular descriptor types, their quantity, and its descriptor names, which are calculated by PaDEL-Descriptor. These are all 2D molecular descriptors.

1.3 Drug Toxicity and High Throughput Screening

Toxicityis the degree to which a substance can damage an organism, or it can give adverse health effects to our body. Toxicology is a study of these adverse effects (symptoms, mechanism, treatment, detection of poisoning, especially poisoning in the people) [6].

More than two thousand data form medicines research centre shows that toxicity is the leading cause of failure of compounds in drug development. The challenges in drug discovery are to discover the efficacy against toxicities or adverse events. Although the most attrition of drugs is related to safety (70%), which comes over candidate selection pre-clinically, and motivate to develop a better predictive model of toxicity [8].

High throughput screening (HTS) is a technologically enabled field of applied science that creates an interface between chemical libraries and biological assays to explore and identify chemical bioactivity rapidly [9].

It is the process of finding a new drug against a chosen target for a particular disease, wherein large libraries of chemicals are tested for their ability to modify the target. Suppose, if the target is a novel G-protein coupled receptor, then compounds will be screened for their ability to inhibit or stimulate that receptor. If the target is a protein kinase, the chemicals will be tested for their ability to inhibit that kinase. This process will require several iterative screening runs, where the properties of the new molecular entities will improve, and allow the favoured compounds to go forward to in vitro and in vivo testing for activity in the disease model of choice.

In the company, massive data of vast numbers of molecules and chemical compounds exhibit types of toxicity in parallel which can be investigated by HTS experiments. Hence traditional HTS experiments are the time consuming and intensive process. There are more than 40 million commercially compounds available on PubChem or Zinc database for the virtual screening using the computational method.

Table 1.1: Calculated molecular descriptors of drugs

Descriptor type	**Number**	**Descriptor name**
Acidic group count	1	nAcid
ALOGP	3	ALogP, ALogp2, AMR
APol	1	apol
Aromatic atoms count	1	naAromAtom
Aromatic bonds count	1	nAromBond
Atom count	14	nAtom, nHeavyAtom, nH, nB, nC, nN, nO, nS, nP, nF, nCl, nBr, nI, nX
Autocorrelation	346	ATS0m, ATS1m, ATS2m, ATS3m, ATS4m, etc
Barysz matrix	91	SpAbs_DzZ, SpMax_DzZ, SpDiam_DzZ, SpAD_DzZ, SpMAD_DzZ, etc
Basic group count	1	nBase
BCUT	6	BCUTw-1l, BCUTw-1h, BCUTc-1l, BCUTc-1h, BCUTp-1l, BCUTp-1h
Bond count	10	nBonds, nBonds2, nBondsS, nBondsS2, nBondsS3, nBondsD, nBondsD2, nBondsT, nBondsQ, nBondsM
BPol	1	bpol
Burden modified eigenvalues	96	SpMax1_Bhm, SpMax2_Bhm, SpMax3_Bhm, SpMax4_Bhm, etc
Carbon types	9	C1SP1, C2SP1, C1SP2, C2SP2, C3SP2, C1SP3, C2SP3, C3SP3, C4SP3
Chi chain	10	SCH-3, SCH-4, SCH-5, SCH-6, SCH-7, VCH-3, VCH-4, VCH-5, VCH-6, VCH-7
Chi cluster	8	SC-3, SC-4, SC-5, SC-6, VC-3, VC-4, VC-5, VC-6
Chi path cluster	6	SPC-4, SPC-5, SPC-6, VPC-4, VPC-5, VPC-6
Chi path	32	SP-0, SP-1, SP-2, SP-3, SP-4, SP-5, SP-6, SP-7, ASP-0, ASP-1, etc
Constitutional	12	Sv, Sse, Spe, Sare, Sp, Si, Mv, Mse, Mpe, Mare, Mp, Mi
Crippen logP and MR	2	CrippenLogP, CrippenMR
Detour matrix	11	SpMax_Dt, SpDiam_Dt, SpAD_Dt, SpMAD_Dt, EE_Dt, VE1_Dt, VE2_Dt, VE3_Dt, VR1_Dt, VR2_Dt, VR3_Dt
Eccentric connectivity index	1	ECCEN
Atom type electrotopological state	489	nHBd, nwHBd, nHBa, nwHBa, nHBint2, nHBint3, nHBint4, nHBint5, etc
Extended topochemical atom	43	ETA_Alpha, ETA_AlphaP, ETA_dAlpha_A, ETA_dAlpha_B, ETA_Epsilon_1, etc
FMFDescriptor	1	FMF
Fragment complexity	1	fragC
Hbond acceptor count	4	nHBAcc, nHBAcc2, nHBAcc3, nHBAcc_Lipinski
Hbond donor count	2	nHBDon, nHBDon_Lipinski
Hybridization ratio	1	HybRatio
Information content	42	IC0, IC1, IC2, IC3, IC4, IC5, TIC0, TIC1, TIC2, etc
Kappa shape indices	3	Kier1, Kier2, Kier3
Largest chain	1	nAtomLC
Largest Pi system	1	nAtomP
Longest aliphatic chain	1	nAtomLAC
Mannhold LogP	1	MLogP
McGowan volume	1	McGowan_Volume
Molecular distance edge	19	MDEC-11, MDEC-12, MDEC-13, MDEC-14, MDEC-22, etc
Molecular linear free energy relation	6	MLFER_A, MLFER_BH, MLFER_BO, MLFER_S, MLFER_E, MLFER_L
Path counts	22	MPC2, MPC3, MPC4, MPC5, MPC6, MPC7, etc
Petitjean number	1	PetitjeanNumber, etc
Ring count	68	nRing, n3Ring, n4Ring, n5Ring, n6Ring, etc
Rotatable bonds count	4	nRotB, RotBFrac, nRotBt, RotBtFrac
Rule of five	1	LipinskiFailures
Topological	3	topoRadius, topoDiameter, topoShape
Topological charge	21	GGI1, GGI2, GGI3, GGI4, GGI5, GGI6, GGI7, etc
Topological distance matrix	11	SpMax_D, SpDiam_D, SpAD_D, SpMAD_D, EE_D, etc
Topological polar surface area	1	TopoPSA
Van der Waals volume	1	VABC
Vertex adjacency information (magnitude)	1	vAdjMat
Walk counts	20	MWC2, MWC3, MWC4, MWC5, MWC6, etc
Weight	2	MW, AMW
Weighted path	5	WTPT-1, WTPT-2, WTPT-3, WTPT-4, WTPT-5
Wiener numbers	2	WPATH, WPOL
XLogP	1	XLogP
Zagreb index	1	Zagreb

1.3.1 Causes of toxicity

Most people are exposed to many different chemicals during their lifetimes through various sources, including food, household cleaning products, and medicines. In some cases, these chemicals can be toxic. In fact, more than 30% of promising pharmaceuticals are failed in human clinical trials because they are determined to be toxic despite promising of pre-clinical studies in animal models [10]. Creating new methods for assessing chemical toxicity of drug molecules is essential for a scientist to save the life of animal, time and money. The goal is to quickly and efficiently test whether certain chemical compounds have the potential to disrupt processes in the human body that may lead to adverse health effects. Physicochemical properties of chemical compounds always guide to determine its activity (active or inactive); therefore, it has been rigorously used to classify as drug and non-drug.

In this work, we try to explore the machine learning methods (such as random forest, support vector machine, neural network, linear model, decision tree, and AdaBoost, etc.) with physicochemical properties to predict the toxicity of the chemical compounds. There are more than 8000 molecular descriptors (physicochemical properties) are identified for chemical compounds that will help in train and test the predictive model.

Active drug molecules are those molecules, which can bind to one or more biochemical pathway assays and create some toxic effects into our body. These toxic effects are nuclear receptor effects (NR) and stress response effects (SR). Both the NR and SR are highly affect 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 [11]. The role of a drug is to correct the functioning of these nuclear receptor signalling pathways or stress response signalling pathways [12]. We can build computational models to predict the activity of the drug molecules in the androgen receptor, estrogen receptor, aryl hydrocarbon receptor, and antioxidant response element based on their physicochemical properties [13]. Following are the 12 biological pathway assays, where seven are the nuclear receptors, and five are the stress response pathways, these bioassays can give distinct adverse health effects on their activation.

1. Predict compound activity in all nuclear receptor signaling pathways.
(a) AR : Androgen receptor
(b) AhR: Aryl hydrocarbon receptor
(c) AR-LBD: Androgen receptor with LBD
(d) ER: Estrogen receptor alpha chemical compound
(e) ER-LBD: Estrogen receptor alpha with LBD

(f) Aromatase
(g) PPAR-gamma: Peroxisome proliferator-activated receptor gamma

2. Predict compound activity in all stress response pathways.
(a) NRF2/ARE: Nuclear related factor 2/Antioxidant responsive element
(b) ATAD5
(c) HSE: Heat shock factor response element
(d) MMP: Mitochondrial membrane potential
(e) P53

Androgen receptor is a type of nuclear receptor that is activated by binding any of the androgenic hormones, including testosterone and dihydrotestosterone. Estrogen receptor (ER) is a nuclear hormone receptor which is activated by the estrogen hormones. The aryl hydrocarbon receptor is a transcription factor that regulates gene expression. Antioxidant response element are the central part of the signal transduction pathway in eukaryotic cells that respond to oxidative stress [14].

We have proposed various models to predict the activity (active or inactive) of drug molecules using the information of above discussed 12 well-studied biochemical pathway assays. The study will be based on the chemical structure information of the drug molecules.

1.4 Computational Intelligence Approaches

Computational intelligence (CI) is featured by computational systems, and it is similar to human intelligence which is featured by the brain. CI is an umbrella term under which several methodologies are grouped. These are:

- Machine learning
- Ensemble learning
- Deep learning
- Classification
- Regression
- Clustering

1.4.1 Machine learning

Machine learning is a branch of artificial intelligence, and it is about the construct and study of a system which learns from data. Figure 1.3 shows the categorization of machine learning approaches. It is divided into two categories, i.e. supervised learning and unsupervised learning. Supervised learning generates a function that maps inputs to desired outputs (also called labels; human experts provide it for labelling the training examples). Unsupervised learning models is a set of inputs, like clustering. Here, labels are not known during training.

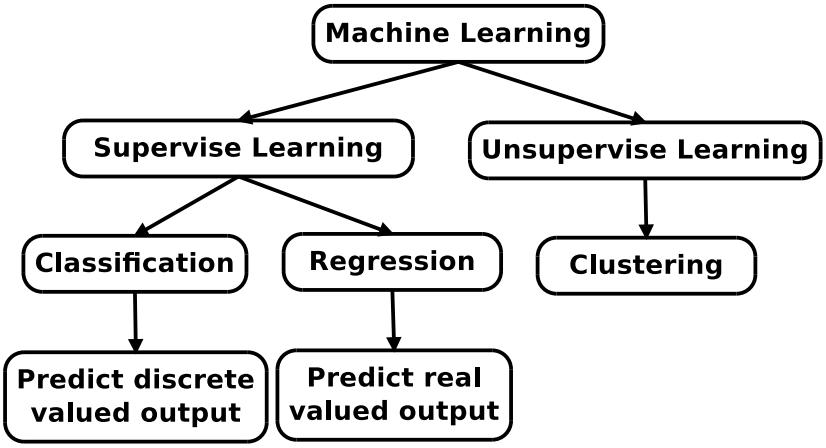


Figure 1.3: Machine learning categorization.

Supervise learning is further classified into two categories, which are:
(i) Classification: Predict the discrete-valued output.
(ii) Regression: Predict the real-valued output.

A classification technique is a systematic approach to building classification models from an input dataset. Decision tree, neural networks, random forest, AdaBoost, and support vector machines are the classifiers. Figure 1.4 shows the general approach of the classification of data.

In our case, a machine learning system could be trained on different drug molecules to learn to distinguish between active or inactive massages. After learning, it can then be used to classify new drug molecules into active and inactive categories.

The Table 1.2 shows the recently developed classification and regression models. All the models are available in R open-source software, which is licensed under GNU GPL.

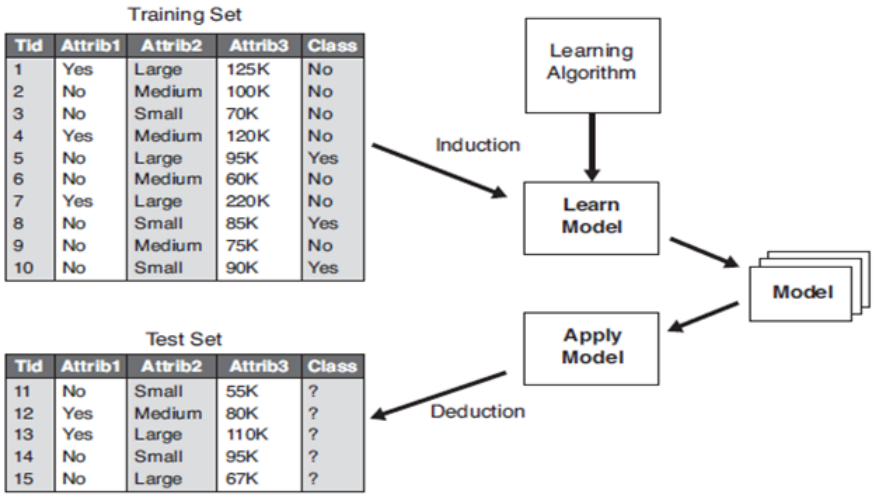


Figure 1.4: General approach to building classification model

Table 1.2: Machine learning models using R

SN	Model	Model type	Method	Package	Tuning parameter
1	ada	Classification	ada	ada	maxdepth, iter, nu
2	bag	Dual Use	bag	caret	vars
3	Boruta	Dual Use	Boruta	Boruta	mtry
4	svmLinear	Dual Use	svmLinear	kernlab	C
5	glm	Dual Use	glm	stats	None
6	J48	Classification	J48	RWeka	C
7	knn	Dual Use	knn	caret	k
8	lda	Classification	lda	MASS	None
9	lm	Regression	lm	stats	None
10	neuralnet	Regression	neuralnet	neuralnet	layer1, layer2, layer3
11	nnet	Dual Use	nnet	nnet	size, decay
12	rf	Dual Use	rf	randomForest	mtry
13	rpart	Dual Use	rpart	rpart	cp
14	RRF	Dual Use	RRF	RRF	mtry,coefReg,coefImp

1.4.2 Ensemble learning

Ensemble learning is a process of combining more than one model to solve a given computational intelligence problem. Generally, it is used to enhance the predictability as well as to improve the robustness of a model. The combination of learners create an ensemble model; these learners are known as base learners. The ensemble model is generally much stronger than these base learners. The ensemble approach is used because it is capable of boosting weak learners [15]. The significant improvement in the prediction by the use of an ensemble learning approach encourages the researchers to solve the problems of different fields. There are many more benefits of ensemble approach which include effective prediction results, selection of relevant features, a combination of data, incremental learning, class imbalance handling, and correction of errors.

The ensemble approach uses divide and conquer method in which a complex problem is divided into multiple chunks that are easy to analyze and solve. This approach has the

advantage that is the ensemble model can adapt any diversity in the data more correctly as compared to single model [16]. It suggests that the ensemble approach is more efficient than the single model [17]. The progress of the ensemble approach depends upon the diversity in the individual model corresponding to misclassified instances [18].

Polikar stated that there are four ways to attain this diversity. Firstly, train the individual model with different data chunk. Secondly, use different training parameters. Thirdly, use different properties to train the model and finally, combine different types of models [19]. According to Dietterich [20], there are three reasons which conclude that the ensemble model is efficient than the single model. The first is that the training dataset does not always facilitate the required information to select one correct hypothesis. The second is that the weak models are not properly trained. The third is that the hypothesis space being searched might not get the proper target function while an ensemble model can produce a good approximation.

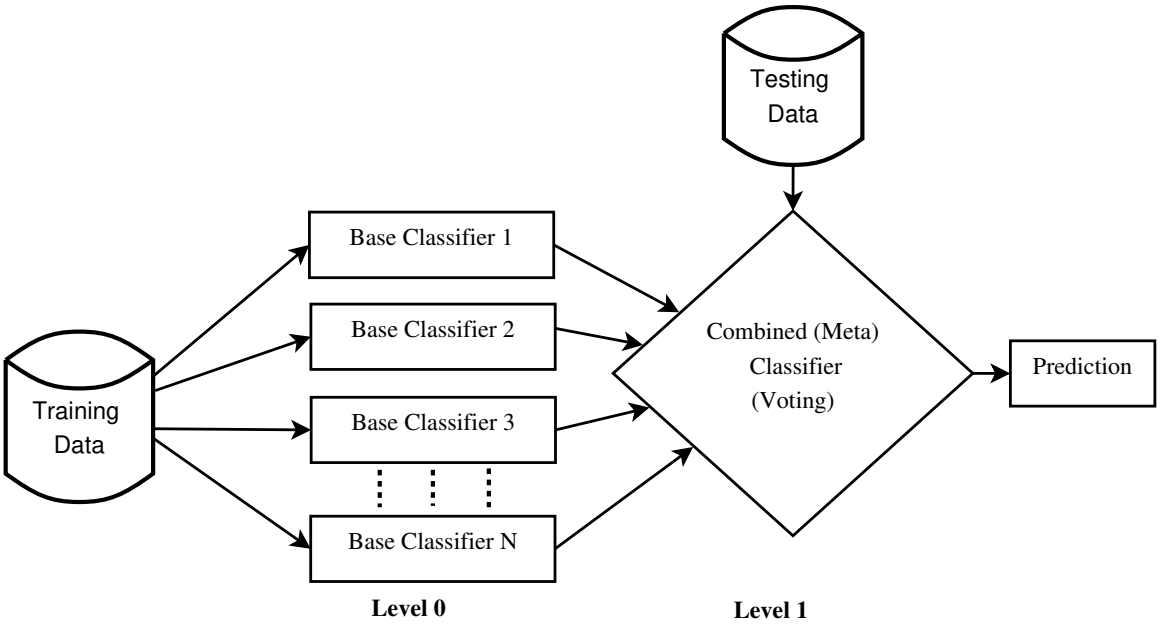


Figure 1.5: Ensemble learning with bagging

1.4.2.1 Need of ensemble learning

The process of ensemble learning is an effective approach to achieve a highly accurate model by combining less accurate ones [17]. There is no one best machine learning model to solve all the cases of a given problem [21]. Various techniques are used to improve the performance of machine learning models which include efficient preprocessing of the data, collect a large number of features and perform feature selection task to get relevant

ones, explore different machine learning models. If results are not desirable, then we can combine the less accurate models. In an ensemble model, more than one opinions are there for a single instance. Thus, if one model fails to predict the correct output, there is a chance that the other models predict it correctly [22].

There are two errors in the trained model, including bias error and variance [15]. Bias error quantifies that on an average, the predicted values differ from the targeted values. High bias represents that the model is underperforming, which means it has missed some essential trends. On the other hand, the variance is used to quantify that the prediction produced on the same observation differ from each other. High variance means the model is overfitted and will produce adverse predictions on instances except for training dataset. To deal with these errors, an ensemble approach is an efficient way [15]. There are two methods to combine the models which include bagging and boosting. The ensemble model is a meta-algorithm, which is a combination of different models and in order to minimize the variance bagging ensemble approach is used, and in order to minimize the bias boosting ensemble approach is used.

1.4.2.2 Techniques of ensemble learning

Ensemble approach is beneficial to enhance the models performance. There are two ways to combine the different models and are explained below:

1. **Bagging:** Bagging means bootstrap aggregation, which is a simple and successful method to ensemble the models. It is used to improve the unstable classification problems. For instance, weak models like decision tree can fluctuate when any training point changes its position and may become a different tree. This ensemble method can be applied to other models as well. Bagging method is beneficial for the vast and high dimensional datasets. It is introduced by Leo Breiman [23] to minimize the variance of the model. In bagging, outputs of n models are aggregated which are generated by using N bootstrap sets, as shown in Figure 1.5. These sets are generated by using complete dataset via feature selection and random method with replacement. The parallel training of each model is possible because the training of each model is independent. In the end, voting or averaging of outputs is performed where each bootstrap set produces outputs.

2. **Boosting:** Boosting is introduced by Schapire [24], which is an ensemble technique to boost the performance of weak models and then group into a robust model. It facilitates the sequential training of the models. The first model is trained on the complete dataset while other models get trained by using training sets. These

sets are based upon the output of the previous ones. The incorrect instances are extracted to increase their weights. Therefore, these instances have a high chance of appearing in the training dataset, which is used by the next model. By using this approach, different models are well trained on different sets of the data, which helps the ensemble model to produce enhanced results [25].

1.5 Dimensionality Reduction

In theory, the information provided by additional features should help to improve the models accuracy, however, in reality, these additional features increase the risk of overfitting, i.e., memorizing noise in the data rather than its underlying structure. For a given sample size, there is a maximum number of features above which the classifiers performance degrades rather than improves. This problem is called the curse of dimensionality, and the techniques for reducing of high-dimensional data intuitively into low-dimensional data fall into dimension reduction. One of the main aspects of the curse of dimensionality is the large size of the dataset. In fact, processing high-dimensional data is already a tough task in current scientific research. Feature selection is a technique where we try to map the high-dimensional data space into lower dimensional space with minor loss of information.

1.5.1 Feature selection methods

Feature selection is a technique where we select some important set of features from a large collection of features. It means, we are interested in finding k of the d dimensions that give us the most significant information and we discard the other (d - k) dimensions. Subset selection is also a feature selection method, where the best subset contains the least number of dimensions which mostly contribute to providing better accuracy. Here, we discard the remaining, unimportant features using its two approaches, one is called forward selection and the second one is called backward selection [26][27].

Here, we are taking drug molecules of ER, AhR, AR, and ARE, where all kind of drug molecules having 1444 features that are very high in dimensional. Therefore, we applied the correlation-based feature selection (CFS) algorithm for the selection of important features of ER drug molecules during activity prediction. Similarly, we applied the information gain algorithm for the selection of important features of ER Drug molecules and AR drug molecules during their activity score, potency, and efficacy prediction. Boruta

algorithm is another important feature selection algorithm, which is used to select important features of AhR drug molecules and AR drug molecules for the prediction of activity. Similarly, the random forest importance algorithm is used to select the important features for the prediction of the activity of ARE drug molecules.

1.5.2 Advantages of dimensionality reduction

Many learning algorithms perform poorly in a high dimensional space. Often some features in the data set are just noise and thus do not contribute to (sometimes even degrade) the learning process. The model finds difficulty in analyzing datasets with many features is known as the curse of dimensionality. Dimensionality reduction can circumvent this problem by reducing the number of features in the dataset before going to the training process. It can also reduce the computation time, and the resulting classifiers take less space to store. Models with a small number of variables are often easier to interpret by domain experts. Dimensionality reduction is also useful as a visualization tool, where the high dimensional data set is transformed into two or three dimensions for display purposes [28].

1.5.3 Disadvantages of dimensionality reduction

The main drawback of dimensionality reduction is the possibility of information loss when it has done poorly because the dimensionality reduction can discard useful information instead of irrelevant information [28].

1.6 Class Imbalance

Class imbalance is the problem in machine learning where the total number of a class of data (positive) is far less than the total number of another class of data (negative). This problem is extremely common in practice and can be observed in various real-world applications where the class distributions of data are highly imbalanced. These include fraud detection, anomaly detection, medical diagnosis, oil spillage detection, facial recognition, etc. For the two-class case, one assumes that the minority class and other is the majority class. Often the minority class is very infrequent, such as 1% of the total dataset. If one applies most traditional (cost-insensitive) classifiers on this dataset, then the classifiers are likely to predict everything as negative (the majority class). Class

imbalance issue can be resolved by various techniques like oversampling, undersampling, synthetic minority oversampling technique (SMOTE), and ensemble learning.

1.6.1 Synthetic minority oversampling technique

SMOTE technique is used to avoid over-fitting, which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example, and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset. The new dataset is used as a sample to train the classification models [29].

Advantages

- Mitigates the problem of over-fitting caused by random oversampling as synthetic examples are generated rather than a replication of instances.
- No loss of useful information.

Disadvantages

- While generating synthetic examples, SMOTE does not take into consideration of neighbouring examples from other classes. This process increases the overlapping of classes and can introduce additional noise.
- SMOTE is not very effective for high dimensional data.

1.7 Machine Learning Models used

The random forest (RF) model is used in prediction of activity, activity score, potency, and efficacy for the prediction of small drug molecules of the estrogen receptor. We used an ensemble approach for the prediction of activity, activity score, potency, and efficacy of other nuclear receptors like androgen receptor, aryl hydrocarbon receptor, and stress response pathway like antioxidant response pathway. The brief details of all the models, which are used directly or as the base classifiers in our proposed ensemble models are given below:

1.7.1 Decision tree

Decision tree (DT) is one of the supervised learning algorithm, which is used to solve the classification problems. It can be used for discrete and continuous input and output variables. While constructing the decision tree top-down approach is considered. Each branch node in decision tree shows a preferred attribute from the given attributes and each leaf node shows a decision or output [30]. For selecting a particular node, entropy and information gain are calculated, which are discussed below:

- **Entropy**: It shows the degree of disorganization in the data. In other words, it measures the randomness in data. If the number of positive and negative instances are equal, then entropy is 1; otherwise, it is between 0 and 1. [30]. Eq. 1.1 is used to calculate the entropy:

$$Entropy = -\sum_{i=1}^{m} p_i log(p_i) \tag{1.1}$$

  p_i is the probability of class i, which calculate the proportion of class i in the set.

- **Information gain**: It measures the relative modification in the entropy concerning the input or independent attributes. In other words, it measures the expected decrease in entropy. It is used to determine the decision node from the given attributes. The best option to decrease the depth of the decision tree is to delete those attributes which have repeated decrease in entropy. For finding the root node from the given attributes, the information gain of each attribute is computed. The highest information gain attribute is used as the root node [30]. Eq. 1.2 is used to calculate the information gain of each attribute:

$$\text{Information Gain} = \text{Entropy(Entire Set)} - [\text{Average(Entropy(Each Split))}] \tag{1.2}$$

1.7.2 Random forest

RF is an ensemble-based classifier built using bagging where each of the classifiers in the ensemble is a decision tree classifier so that the collection of classifiers is a forest. It can be used for classification as well as for regression problems. It has almost same hyper-parameters as a decision tree. The individual decision trees are generated using a random selection of attributes at each node to determine the split. During classification,

each tree votes and the most popular class is returned, and during regression, it takes an average of all the generated decision trees predictions [30]. RF can also be used to select important features.

1.7.3 Support vector machine

Support vector machine (SVM) generates the hyperplane which divides the whole data into classes. It is an algorithm that receives data as input and produces output as a line, which separates these classes. The points close to the hyperplane from both the classes are known as support vectors. After this, the distance between this hyperplane and support vectors are computed and is called a margin. The main objective of the SVM is to maximize this margin. The hyperplane is the optimal hyperplane by which margin is maximized. Therefore, SVM creates a decision boundary in such a way which separates the two classes as far as possible [26].

1.7.4 Neural network

Neural network (NN) works on the underlying architecture and behaviour of the human brain. There are neurons in the human brain which process and transmit the information to each other. Dendrites are there to receive the inputs and based on these inputs; an output is produced. This output is transmitted to other neurons with the help of axon. In NN, the network contains artificial neurons known as nodes which process the information and does operations. NN contains three layers which include an input layer, hidden layer, and output layer. Input layer takes a massive amount of input data such as text, audio, image pixels, numbers, etc. Hidden layer performs pattern analysis, mathematical operations, feature extraction, etc. on the input data. The hidden layers may be more than one in the NN. The output is generated by the output layer. NN has many parameters and hyper-parameters, which produces the output. These parameters include biases, number of neurons, weights, learning rate, etc. Every node in a network has weights with it, and transfer function calculates the weighted sum of the inputs and also adds bias into it. These results act as an input to the activation function, which will further decide the nodes to get fired. The selection of the type of activation function depends upon the required output [26]. Activation functions are used in the hidden layer or the output layer or both layers.

1.7.5 Linear model

Linear model (LM) describes a continuous response variable as a function of one or more predictor variables. They can help you understand and predict the behaviour of complex systems or analyze experimental, financial, and biological data. Linear regression is a statistical method used to create a linear model. The model describes the relationship between a dependent variable y (also called the response) as a function of one or more independent variables Xi (named the predictors) [31]. The general equation for a linear model is:

$$y = \beta_0 + \sum \beta_i X_i + \epsilon_i \tag{1.3}$$

where β represents linear parameter estimates to be computed and ϵ represents the error terms.

1.7.6 Adaptive boosting

Adaptive boosting (AdaBoost) is the first original boosting technique, which creates a highly accurate prediction rule by combining many weak and inaccurate rules. Each classifier is serially trained with the goal of correctly classifying examples in every round that were incorrectly classified in the previous round [29]. For a learned classifier to make reliable predictions, it should follow the following three conditions:

- The rules should be simple.
- Classifier should have been trained on sufficient number of training examples.
- The classifier should have low training error for the training instances.

1.8 Performance evaluation parameters

Performance evaluation parameters are used to analyse the performance of the proposed ensemble models and each model. These evaluation parameters generate a score, which ranks each model according to this score. In this study, models are evaluated based on the following parameters:

1.8.1 Classification parameters

For checking the performance of proposed and existing binary classification models, some specific evaluation parameters like Gini coefficient, specificity, sensitivity, AUC, and accuracy are used. All these performance metrics are essential for any binary classifiers and calculated with the help of confusion/error matrix [30], which is shown in Table 1.3.

Table 1.3: Confusion matrix for activity prediction

		Predicted class	
		Active	Inactive
Actual class	Active	TP	FN
	Inactive	FP	TN

- TP: True Positive.
- TN: True Negative.
- FP: False Positive.
- FN: False Negative.

1.8.1.1 Area under the curve (AUC)

To check the quality of the model, AUC is calculated. AUC is the complete 2D area under the ROC curve which is from (0,0) to (1,1). High AUC value depicts the good quality of the model. Its value lies between 0 and 1. The model has AUC value near to 1 means its quality is excellent [30].

1.8.1.2 Gini coefficient

Gini coefficient is measured to calculate inequality in the distribution. It can be derived from AUC receiver operating characteristics (ROC) value. Gini coefficient is a ratio between the area in between the ROC curve and diagonal line and the area of a complete triangle. ROC curve plot represents two evaluation parameters that are sensitivity (true positive rate) and 1-specificity (false positive rate). Thus, the Gini coefficient shows the inequality between these two evaluation parameters. If the number of negatives is large, then there is an issue in the efficiency of the model. Its value lies between 0 and 1. Value 1 means inequality and value 0 means equality [32]. For example, if a model scores Gini

value 60%, then it is considered as a good model. The Gini of the model is calculated using AUC as follows:

$$Gini = 2 * AUC - 1 \tag{1.4}$$

1.8.1.3 Accuracy

Accuracy is a metric to evaluate the machine learning models. It is used to determine which model is correctly learning the patterns and relationships between features based upon the training dataset [26]. In simple words, It measures the correct predictability of the model. A high accuracy model doesnt mean that the model is predicting all the instances correctly because it may be results of misleading. For example, in imbalanced class, a model may predict all cases of majority class with high accuracy. Therefore, other parameters should also be considered to check the model performance. The accuracy of the model is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100 \tag{1.5}$$

1.8.1.4 Sensitivity

Sensitivity (Sens) or recall has been calculated to check the true positive rate of any model as well as proposed ensemble models [30]. It is the proportion of actual positives which are correctly identified as positives by the model and is computed as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{1.6}$$

1.8.1.5 Specificity

Specificity (Spec) is the ability of the model to identify the results of negative samples [30]. It has been calculated to check the true negative rate of any model as well as proposed ensemble models and is computed as follows:

$$Specificity = \frac{TN}{TN + FP} \tag{1.7}$$

Some other binary classification parameters like precision, F-score, Cohens kappa, and Matthews correlation coefficient (MCC) are also used in our work, which are described in corresponding chapters.

1.8.2 Regression parameters

To check the performance of proposed and existing regression models, we have evaluated these models on some specific parameters like root mean square error (RMSE), correlation (r), coefficient of determination (R^2), and accuracy for the continuous classes. These performance parameters are described fully in Chapter 3 and Chapter 5.

1.9 Thesis Organization

The thesis is organized into seven chapters. A brief outline of these chapters is given below:

Chapter 1: This chapter introduces drug design and development process, physicochemical properties and molecular descriptors, drug toxicity and high throughput screening, causes of toxicity, computational intelligence approaches like machine learning and ensemble learning, feature selection, and class imbalance problem. In the last, we have described various machine learning models which are used in this work, and different performance evaluation parameters.

Chapter 2: This chapter presents a literature survey, which is related to our work. We have also discussed the various existing model for toxicity prediction. Further, prominent research gaps are identified, and on the basis of these research gaps, research objectives are defined.

Chapter 3: This chapter describes the prediction of the activity of small drug molecules of estrogen receptor using the multilevel prediction model. This work focuses on the classification and regression with various physicochemical properties to predict the activity, activity score, potency, and efficacy, while the dataset is highly imbalanced in class and having a vast number of features.

Chapter 4: This chapter describes the prediction of the toxicity of small drug molecules of aryl hydrocarbon receptor using ensemble-based classification models. The random forest model is used as a base classifier. This work focuses on the classification of active or inactive drug molecules using machine learning models

and physicochemical properties to predict the toxicity of drug molecules. Here, activity is the target class of toxicity prediction. AhR dataset is highly imbalanced in class and having a very large number of features.

Chapter 5: This chapter describes the prediction of the activity of small drug molecules of androgen receptor using multilevel ensemble model. We have created an ensemble-based classification model for the prediction of activity, and we have created the ensemble-based regression model for the prediction of activity score, potency, and efficacy. This work focused on the classification and regression with various physicochemical properties to predict the activity, activity score, potency, and efficacy, while the dataset is highly imbalanced in class and having a huge number of features.

Chapter 6: This chapter describes the prediction of the toxicity of small drug molecules of the antioxidant response element using an ensemble-based classification model. Here, random forest, decision tree, AdaBoost, and support vector machine are used as the base classifiers. Here, activity is the target class for toxicity prediction. This work focused on the classification of active or inactive drug molecules using ensemble-based classification model and physicochemical properties. ARE dataset is also highly imbalanced in class and having a very large number of features.

Chapter 7: This chapter summarizes the key findings and main contributions of the thesis and lists the possible future research directions.

Chapter 2

Literature Review

This chapter reviews the research work of various researchers, which is related to toxicity prediction or activity prediction using physicochemical properties and various computational intelligence approaches. Various important research contributions related to toxicity prediction using machine learning models are presented in detail. Here, we have also mentioned various existing models of toxicity prediction, research gaps, and research objectives.

2.1 Related Work

Toxicology has become an essential element in health sciences because the many government and non-government organizations utilize information from toxicology to evaluate and regulate hazards in the truancy of broad human exposures. Toxicology methods are specially used by industry in drug development, which provide useful information for the design of drug molecules [6].

Every day we are exposed to various chemicals via food additives, cleaning and cosmetic products, and medicines; some of them might be toxic. However, testing the toxicity of all existing compounds by biological experiments is neither financially nor logistically feasible. Therefore, in 2008, the U. S. National Institutes of Health (NIH) and the U. S. Environmental Protection Agency (EPA) was agreed on collaborating on future toxicity testing activities (Committee on Toxicity Testing and Assessment of Environmental Agents, National Research Council, 2007). Their efforts were later joined by the U. S. Food and Drug Administration (FDA) under the umbrella of the Tox21 program. The program's stated goals are to develop better toxicity assessment methods, as current practices are not likely to scale with the increased demand for effective toxicity testing [33]. National Center for Advancing Translational Sciences (NCATS) had provided assay activity data and chemical structures on the Tox21 collection of 10,000 compounds (Tox21 10K). Tox21 program is proposed by the USA in 2011 under the department of health and human services committee on toxicity testing and assessment of environmental agents. The motivated goal of the Tox21 program was to develop better toxicity

assessment methods [34]. NIH Center for Translational Therapeutics (NCTT) has made a chemical inventory browser, which is free to deliver research about the toxicity of the chemicals.

In 2014, the government agencies like NIH, EPA, and FDA launched the Tox21 data challenge under the "Toxicology in the 21st Century" (Tox21) initiative. The goal of this challenge was to assess the performance of computational methods in prediction of the toxicity of chemical compounds. State of the art toxicity prediction methods are developed specifically over chemical descriptors. Machine learning and deep learning methods are new in this field, and these were applied in this data challenge to predict the toxicity of drug molecules, where these methods outperformed all other participated methods. In this challenge, we observed that deep neural networks automatically learn features resembling well-established toxicophores, and the deep learning approach won both the panel's challenges (nuclear receptor panel and stress response panel) as well as the overall grand challenge; therefore, it sets a new standard of toxicity prediction [34].

To predict the toxicity of chemical compounds for the welfare of human and environmental health. Critical examinations of toxic drug activities are required to achieve 100% accuracy. Hence, it is necessary to utilize computational intelligence (CI) approaches where machine learning and deep learning are too essential for toxicity prediction because these methods provide the following benefits [35][27]:
(1) Automatic tools to search for hypotheses explaining data.
(2) Much more accurate than manual testing techniques in pharmacology.
(3) Don't need a programmer or human expert.
(4) Flexible in dealing with any learning task.

Many chemical compounds by which people get exposed in their lives are full of hazardous and harmful. Kola and Landis proposed that toxicity is also the central issue for the development of the new drug. According to a report in clinical trial, more than 30% drug candidates fail because of undetected toxic effect [10].

Investigations recently present a general assumption, which shows that the physicochemical properties which participate in increasing the toxicity in the drug strongly depend on microscopic features and structure of the molecule. Behind this assumption, the physicochemical property is modeled as the response variable, which enables us to set up QSPR (quantitative structure-property relationship), QSTR (quantitative structure-toxicity relationship), and QSAR (quantitative structure-activity relationship). All these models have been considered for the activity/toxicity prediction of various drug molecules on the basis of the chemical structure of the molecule [36][37].

Current methods for testing the toxicity of a high number of chemicals rely on High-Throughput Screening (HTS). HTS experiments can investigate whether a chemical compound at a given concentration level exhibits a specific type of toxicity for several different compounds in parallel. These experiments are repeated with varying concentrations of the chemical compound, which determines dose-response curves. From these curves, one can reliably determine whether a chemical compound can activate a given receptor or pathway, inhibited it, or did not interact at all.

HTS allows researchers to conduct millions of pharmacological, genetic and chemical tests. Conducting these HTS experiments is a time and cost-intensive process. Typically, a compound has to be tested for several types of toxicity at different concentration levels. Thus, the whole procedure has to be rerun for many times for each compound [38]. Even an unprecedented expensive effort, the Tox21 project, could test only a few thousands of compounds for only 12 toxic effects [33]. Therefore, computational intelligence methods are in demand for the prediction of precise toxic effects.

In 2012, Merck molecular activity challenge has won using the neural network and the winner team has shown the use of multi-task learning to predict the biological activities on a single protein. Applications of a multi-task neural network for QSAR predictions successfully give us a hint of a deep learning approach for target prediction and drugs' toxicity prediction [39].

Russell and Burchs proposed the concept of 3Rs, where they focused on the limited use of animals (in vivo) in the prediction of toxicity. Non-animal methods like in vitro and in silico are gaining popularity under the 3Rs concept. In recent years tremendous progress has been made in computational biology and mathematical modeling.

Jerrold Tannenbaum et al. (2015) [40] extended the work of Russell and Burchs '3Rs Then and Now,' These scientists focused on the 3Rs are:

- **Replacement:** Methods which avoid or replace the use of animals in research.
- **Reduction:** Use of methods that enable researchers to obtain more information from the same number of animals.
- **Refinement:** Use of methods that minimize potential pain, distress, and enhance animal welfare for the animals used.

In silico toxicology or predictive toxicology has emerged as a fast and cost-effective alternative technique for the early assessment of various toxic chemicals. In silico toxicity prediction techniques include kinetic modeling (where biochemical pathways relevant to

toxicology), expert systems (where information about chemicals and biological systems are given as inputs and algorithms developed to predict the toxicities), and data-driven systems include SAR and QSAR (which adopt data mining techniques such as K-nearest neighbours, neural networks, bayesian statistics, and support vector machines) [41].

Basak et al. proposed a hierarchical QSAR approach, which used topological indices to predict aryl hydrocarbon receptor binding potency on the set of 34 chlorinated dibenzofurans [42] .

Piparo et al. proposed a computational model for predicting aryl hydrocarbon receptor binding, where authors had decided to use QSARs 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 [43].

Cassano et al. developed the CAESAR QSAR model to minimize false negatives to make them more usable for European legislation REACH (Registration, Evaluation, Authorization, and Restriction of Chemical Substances). The CAESAR on-line application ensures that both industry and regulators can easily access and use the developmental toxicity model. The CAESAR platform is freely available tools for human toxicity [44].

QSAR and in vitro derived metabolic parameters are merged by physiologically based biokinetic (PBBK) model, and their assays of toxic organs or tissues can essentially decrease the use of animals for risk assessment regarding toxicity testing [45]. The freely available applications are OECD, QSAR, CAESAR, and Toxmatch, or the commercial expert system like MCASE and TOPKAT are useful tools in toxicity prediction [46].

Drwal et al. proposed molecular similarity-based and Naive Bayes classification for the prediction of the toxicity of nuclear receptors and stress response pathways, which was screened from the Tox21 data challenge 2014. It was implemented in KNIME software [47].

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 Rotation forest and ADTree classifier, and the performance of the model was measured using area under the receiver operating characteristic (AUROC) curve metrics [48].

Capuzzi et al. built QSAR models for 12 stress response and nuclear receptor signaling pathways toxicity assays as part of the 2014 Tox21 data challenge. These models were built using the random forest, deep neural networks and various combinations of descriptors, where deep neural networks had performed better. The drawback of this

methodology is the high demand for computational resources [45].

Fang Bai et al. proposed a prediction model for the response of antioxidant response elements compound by deep learning. In this approach, a series of predictive models by applying multiple deep learning algorithms including deep neural networks (DNN), convolution neural networks (CNN), recurrent neural networks (RNN), and highway networks (HN) were constructed and validated based on Tox21 challenge dataset and applied to predict whether the compounds are the activators or in-activators of AREs [49].

ADMET properties of candidate drug molecules have been predicted, which are used in QSAR/QSPR modelling approaches. These predictions were done using SVM and artificial neural network (ANN) [50].

By using the ligand protein inverse docking approach, scientists have predicted the potential toxicity and side effects of small drug molecules which bind to protein targets [51].

In the case of drug design recently several QSAR analyses with different ANNs have been used where the microscopic features and molecule structure have been characterized by molecular descriptors [35].

Till now, several works have been published, which have the role of AI tools, and have been applied in toxicity prediction and QSAR models, but the majority of them have been done by ANN. For example, Adamczak and Duch [52] discussed the use of ANN and applied it on analyzing of QSAR series and compared its results with the following other three related approaches:

(1) Inductive logic programming (ILP)

(2) Standard linear regression

(3) Decision tree.

As per the literature study, most of the data mining and machine learning techniques have been conducted for the prediction and identification of the toxicity and potency of the drug molecules [53]. Machine learning models like random forest, linear curve fitting and ensemble-based approaches are used, and its graphical evidence is presented to prove the validity of the models [54]. Table 2.1 also lists some essential work related to toxicity prediction using various machine learning models and molecular descriptors.

Table 2.1: Related work

Year	Authors	Method	Work done
2019	Limeng Pu1 et al. [55]	Machine Learning	eToxPred is a new approach to reliably estimate the toxicity and synthetic accessibility of small organic compounds. eToxPred employs machine learning algorithms trained on molecular fingerprints to evaluate drug candidates. The performance is assessed against multiple datasets containing known drugs, potentially hazardous chemicals, natural products, and synthetic bio-active compounds. Encouragingly, eToxPred predicts the synthetic accessibility with the mean square error and the toxicity with the accuracy.
2018	Yunyi Wu et al. [56]	Machine Learning and Deep Learning	In this article, authors have reviewed machine learning methods that have been applied to toxicity prediction. Authors have also discussed the input parameter to the machine learning algorithm, especially its shift from chemical structural description only to that combined with human transcriptome data analysis, which can greatly enhance prediction accuracy.
2017	Ashok K. Sharma et al. [57]	Machine learning based classification and regression models	In this work, by integrating machine-leaning and chemoinformatics approaches, we have developed a computational method ToxiM for the prediction of toxicity of molecules using fingerprints and descriptors as input features.
2019	Fang Bai et al. [49]	Deep Learning	This is the classification of anti-oxidant response elements using deep learning approaches. Dragon software is used for feature extraction and data is collected from tripod (Tox21 challenge data).
2015	Andreas Mayr et al. [11]	Deep Learning	This research is under the Tox21 data challenge, Applying Deep Neural Network to toxicity prediction.
2014	George E. Dahl et al. [58]	Neural Network	Application of multi-task neural networks for QSAR predictions, successfully give us a hint to approach deep learning for target prediction and prediction of toxic drugs.
2009	Devid Hecht et al. [59]	Computational Intelligence Approaches	Engaging computational intelligence in improving ADMET models in order to enhance drug design in its two stages, discovery and development.
2009	Mark T. D. Cronin et al. [46]	Category Formation and Read-Across	Reviewing the most important techniques of in silico and exposing prediction strategies of human health effects. In the results, we found that the transparency of local models based on protein reactions and chemical similarity for toxicity prediction are more accurate than global models.
2005	Helma C. [41]	In silico methods	In silico predictive toxicology techniques are a fast and cost-efficient alternative or supplement to bioassays for the identification of toxic effects at an early stage of drug development.
2002	Dan Neagu et al. [60]	Neural/ Neuro-Fuzzy Networks	Developing models based on Neuro-Fuzzy Structures for knowledge representation to deal with large data of organic compounds. Combining different methods for the same problem increase the accuracy of toxicity predictions which has shown 10% more than classical approaches.

2.2 Toxicity Prediction Models

Many methods have been developed in the last 20 years of research in toxicity prediction using physicochemical properties and machine learning approaches. Most of the methods are broadly classified into the following categories:

1. **Tox 21**: Toxicology in the 21st Century (Tox21) is a unique collaboration between several federal agencies to develop new ways to test whether substances adversely affect human health rapidly. Substances in Tox21 includes a diverse range of products such as commercial chemicals, pesticides, food additives/contaminants, and medical compounds. Tox21 used deep learning and structure of chemicals in the form of the molecular descriptors to the prediction of drug's toxicity.

2. **QSAR**: Quantitative structure-activity relationship models are classification or regression models used in the chemical sciences, biological sciences, and engineering. Like other classification models, QSAR classification models relate the predictor variables to a categorical value of the response variable. But regression QSAR models relate the predictor variables to the potency of the response variable.
   In QSAR modeling, the predictors consist of physicochemical properties or theoretical molecular descriptors of chemicals; the QSAR response variable could be a biological activity of the substances. QSAR models first summarize a supposed relationship between chemical structures and biological activity in a dataset of chemicals. Second, QSAR models predict the activities of new chemicals [36]. A QSAR has the following form of a mathematical model:

$$Activity = f(P) + error \tag{2.1}$$

   Here, P is the physicochemical properties or structural properties of any drug, and the error includes a model error (bias) and observational variability, i.e., the variability in observations even in a correct model.
3. **QSPR**: Similar to QSAR, quantitative structure-property relationships (QSPR) is used when a chemical property is modeled as the response variable. Different properties or behaviours of chemical molecules have been investigated in the field of QSPR [37].
4. **QSTR**: Quantitative structure-toxicity relationships (QSTRs) models are typical examples for the prediction of toxicity, which relates variations in the molecular structures to toxicity. There are many applied modelling techniques in QSTR, such as partial least squares, artificial neural networks, and principal component regression (PCR) [61].
5. **The CAESAR models**: The primary goal of CAESAR projects was to develop QSAR models and make them easily accessible and usable by anyone (regulators, manufacturers, etc.). The first part of this section describes QSAR models for the developmental of toxicity, and the second part discusses the platform developed to make the models accessible [44]. CAESAR was a European country funded project, which was explicitly dedicated to develope QSAR models for the REACH legislation. CAESAR models have been assessed according to the OECD principles for the validation of QSAR.
6. **TOPKAT**: Kurt Enslein's company developed toxicity prediction by Komputer-Assisted Technology (TOPKAT). Health designs use electrotopological descriptor rather than chemical structures to predict mutagenic reactivity with DNA and are an extension of classical QSAR analysis. TOPKAT intelligence is derived solely

from bacterial mutagenicity data and produces a probability that a submitted compound could present genotoxicity issues [62].

7. **MCASE**: Multiple computer-aided structure evaluation (MCASE) is another artificial intelligence tool, uses a different approach to evaluating a submitted molecular structure. It disassociates a test molecule into 2 to 10 atom fragments. It statically evaluates the strengths of associations of the fragments with mutagenicity, generating a qualitative prediction that is refined through taking into consideration of physicochemical properties and the existence of potential deactivated fragments. For the specific version of MCASE, the learning set is based solely on bacterial mutagenicity data derived from 2032 compounds, most of which are environmental toxicant, only 204 of which are pharmaceutical [62].
8. **DSSTox**: DSSTox provides a high-quality free chemistry resource for supporting improved predictive toxicology. A distinguishing feature of this effort is the accurate mapping of bioassay and physicochemical property data associated with chemical substances to their corresponding chemical structures [63].

2.3 Research Gaps and Objectives

After the exhaustive review of literature and various existing models for toxicity analysis, this section presents the multiple research gaps and to overcome these gaps, some research objectives are proposed.

2.3.1 Research gaps

1. **Fast and accurate method is required for the toxicity screening of pre-clinical trial drugs:**
   More than 30% of promising pharmaceutical have failed in human clinical trials because they are determined to be toxic despite promising pre-clinical studies in animal models [64]. To develop a fast and accurate method for the toxicity screening of pre-clinical trial drugs, resulting in saving of time, money, and animal lives [10][4].

2. **Novel data structures for efficient storage and retrieval of massive data sets:**
   The off-the-shelf techniques and technologies used to store and analyse data cannot work efficiently and satisfactorily for massive datasets. Therefore, there is a need

for new data structures and heuristics for handling bulk data of various kind of drug molecules.

3. **Speeding up the discovery of novel drug:**
   Modern drug discovery involves the identification of screening hits, medicinal chemistry and optimization of those hits to increase the affinity, selectivity (to reduce the potential of side effects), efficacy/potency, metabolic stability (to increase the half-life), and oral bioavailability. Once a compound that fulfils all of these requirements, then it will begin the process of drug development before clinical trials.

4. **Dealing with data pre-processing:**
   Data gathering is often, resulting in corrupted data, out-of-range values, missing values, etc. Unscreened information that has not been checked carefully for the predictive system can cause misleading results. The quality of data and its presentation are checked before going to start modelling or analysis [65].

5. **Redundant data:**
   When the integration of multiple databases, the redundant data (between attributes) often occurs. A redundant attribute may be demonstrated if the attribute can be derived from another or a set of attributes. When we deal with around 10,000 compound libraries, then redundant attributes can occur [66].

6. **Missing values:**
   Real life datasets sometimes exposed to incomplete data such as attributes which are related to a particular record that have no values. The missing attributes values have become a challenging issue in data pre-processing stage, especially in medical data mining. Initial reports in many clinical trials allow some attributes to be left blank [67]

7. **Noisy data:**
   The corrupted data is a synonym of noisy data, this is not meaningless data, but any data that has been received, stored, or changed in such a manner that is different from creating it. Machine learning models and analysis are adversely affected by noisy data, which unnecessarily increases the amount of required storage space and give wrong performance results [68].

8. **Feature selection of drug molecules:**
   It is essential to select those features in the data which are most relevant to the problem domain; this process is known as feature selection. Its primary purpose is to reduce the complexity of the dataset. The random forest has an in-built property to select important features. Instead of principal component analysis

(PCA), multiple techniques need to be analyzed like information gain, correlation-based feature selection (CFS), Boruta algorithm. Moreover, many papers surveyed which lack the feature importance module itself.

9. **Addressing class imbalance problem:**
   In data mining, a class imbalance is the greatest issue because of its direct effects on the classification process. This problem occurs when we have two classes such that one of the classes (e.g. positive) has fewer samples than another (e.g. negative). This problem can observe in various disciplines where most machine learning algorithms act well when the number of instances of the classes is equal. It is a serious challenge have to face in the most existing classification methods which tend to have a disproportionate performance on minor labelled examples. The Extremely imbalanced dataset impacts on the classification methods, and it is one of the important causes of missing generalization in machine learning algorithms. Our aim to optimize the overall accuracy with handling the class imbalance issue [69]. Most medical datasets are not balanced in their class labels [70][71].

10. **Accelerating training of predicative model:**
    This stage of the work is lying after data filtering or preprocessing. The model is constructed based on useful, meaningful, and related data. Around 8000, chemical compounds are considered, which are helpful in training and testing of the predictive model. Our prediction model is based on the supervised learning approach, where we are dealing with the number of techniques which have emerged and developed in the last decade. In order to select a suitable model, we need an empirical study and performance comparison of various learning algorithms. Even though No Free Lunch theorem says, among the various classification algorithms, there is none of them has priority over others. Different performance metrics (trade-offs) are measured by a classifier in every domain of predictive systems. It depends on the dataset selected for training and how well the model trained on this data [72].

11. **Exploration of new machine learning models:**
    In the literature, support vector machine, random forest, neural network, linear model, and decision tree are mostly used. To get the effective predictions, other models and their combinations are to be explored for the toxicity prediction.

2.3.2 Research objectives

The following research objectives are formulated:

1. To identify better toxicity assessment features, methods, and algorithms to handle the big data.
2. To develop new methods for assessing chemical toxicity that will predict the activity of drug molecules.
3. To develop computational techniques to quickly and efficiently test certain chemical compounds.
4. To develop a stand-alone application that helps the researchers and research community to predict the toxicity of the newly discovered chemical compound.

Chapter 3

Activity Assessment of Small Drug Molecules in Estrogen Receptor using Multilevel Prediction Model

The authors have proposed an efficient multilevel prediction model for better activity assessment to test whether certain chemical compounds can disrupt processes in the human body that may create negative health effects. Here, a computational method (in-silico) for the quality prediction of drugs in term of their activity, activity score, potency, and efficacy for estrogen receptors (ERs) by using various physicochemical properties (molecular descriptors). PaDEL-Descriptor is used for features extraction. The ER dataset has 8481 drug molecules where 1084 are active, and 7397 are inactive, and each drug molecule has 1444 features. This dataset is highly imbalanced and has a substantial number of features. Initially, a class imbalance problem is resolved through synthetic minority oversampling technique (SMOTE) algorithm, and feature selection is done using FSelector library of R. A machine learning based multilevel prediction model is developed where classification is performed on its first level and regression on its second level. By using all these strategies simultaneously, outperformed accuracy is achieved in comparison to many other computational approaches. The K-fold cross-validation is performed to measure the consistency of the model for all the target classes. Finally, the validity of the proposed method on some AIDS therapy's drug molecules is proved.

3.1 Introduction

Most of the drugs are small molecules which are invented to interact, 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 are the hormone receptors, neurotransmitter receptors, cell-signaling receptors, enzymes, and other functional proteins [12]. Estrogen receptor (ER) is a nuclear hormone receptor which is activated by the estrogen hormones. Two classes of ER exist where one of them is nuclear ERs (ERα and ERβ) and another is

the membrane estrogen receptors (mERs). ER is a nuclear endocrine receptor, plays a vital role in the development, reproduction, physiological, and metabolic state. Disruptors are those molecules which have adverse effects on the functioning of ER. Endocrine disruptors are chemical compounds that can intervene with hormone (or endocrine) system at specific doses. Hormones control in any system in the human body can be disturbed by hormone disruptors [73]. Figure 3.1 shows the crystal structure of an ER-alpha to illustrate that it is a long protein that folds in such a fashion as to attract and hold molecules of estrogens [74]. The Protein Data Bank in Europe (PDBe) id of this figure is 5dvv.

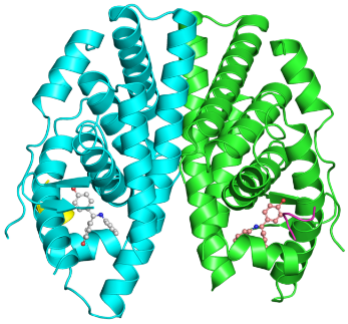


Figure 3.1: Crystal structure of ER-alpha ligand (PDBe id: 5dvv)

The interaction of endocrine disrupting chemicals with steroid receptors can disrupt traditional endocrine operation, and through this interaction, ER signaling can modify genomic and non-genomic ER activity. It is also essential to understand the impact of environmental chemicals on the ER signaling pathway. Inappropriate ER signaling can lead to increased risk of hormone-dependent cancer, abnormal foetal growth, impaired fertility, breast cancer, prostate cancer, deformations of the body, cancerous tumors, learning disabilities, sexual development problems such as feminising of males or masculinising effects on females, and altered metabolism in white adipose tissue [75].

Typically, Drugs are small organic molecules that accomplish their desired activity by binding to a target site on a receptor. The initial phase in the discovery of a new drug is usually to identify and isolate the receptor to which it should bind, followed by testing many small molecules for their ability to bind the target site. The task of authors is to separate the active (binding) compounds from the inactive (non-binding) compounds. This determination can lead to the design of the new compounds that not only binds but also have all the other properties required for a drug [76]. These properties are solubility, oral absorption, lack of side effects, appropriate duration of action, toxicity, etc. A successful drug candidate has the right attributes (physicochemical properties) to bind with molecular target. The molecules bind to specific components (usually proteins) of the

target cells and get activated or deactivated those components that lead to modify cells behavior or destruct them. We are interested in considering only those drug molecules for which target get activated.

Generally, the in silico approach is a predictive science which is utilized for defining discovery and safety efforts in therapeutics. In silico approaches are used for toxicity predictions, and it represents a large number of information-based biological and chemical programs. The toxicology-oriented computational methodology is based on building toxicity databases that give a plausibility to carry out the quantitative structure-activity relationships (QSARs) and quantitative structure-property relationships (QSPRs) analysis. The QSAR and QSPR are the well known linear regression models which have proved their usefulness for predicting the quality of drug molecules in term of their biological activities [77]. The primary purpose of activity prediction using computational methods is to reduce the pressure of animal testing, cost and time reduction in the early stages of drug discovery.

The concept of Russell and Burch (1959) about the growing popularity of the 3Rs (replacement, reduction, refinement) focuses on the limited use of the animal for activity testing (in vivo) [40]. In silico models also can predict ADMET (absorption, distribution, metabolism, excretion, and toxicity) related properties in chemical space that reduce the dependency of chemical laboratory synthesis (in vitro) [78][50].

Active drug molecules are those molecules which can bind to one or more biochemical pathway assays and activate them. Figure 3.2 shows the panels of various nuclear receptors (NRs) and stress response (SR) pathway assays. The activation of the ER (NR-ER) or androgen receptor (NR-AR) is the NR effect, and the activation of mitochondrial membrane potential (SR-MMP) or antioxidant response element (SR-ARE) is the SR effect. Both NR and SR effects are highly relevant to human health. Activation of NR can disrupt endocrine system function, and activation of SR pathways can lead to liver injury or cancer [11]. Here, we are developing the computational model to predict the activity of ER's compounds only from the set of 12 pathway assays. These activity predictions are based on the chemical structure information of molecules [11][48]. Subsequently, all the active drug molecules have three properties to decide the quality of it which are activity score, potency, and efficacy [73].

In this paper, we have proposed a novel machine learning based multilevel prediction model. It has two phases, in its first phase, we have developed classification model by using five decision methods to predict the activity of ER drug molecules, and in its second phase, we have developed regression model only for those drug molecules which are found active by classification. Since, all the active drug molecules further have three

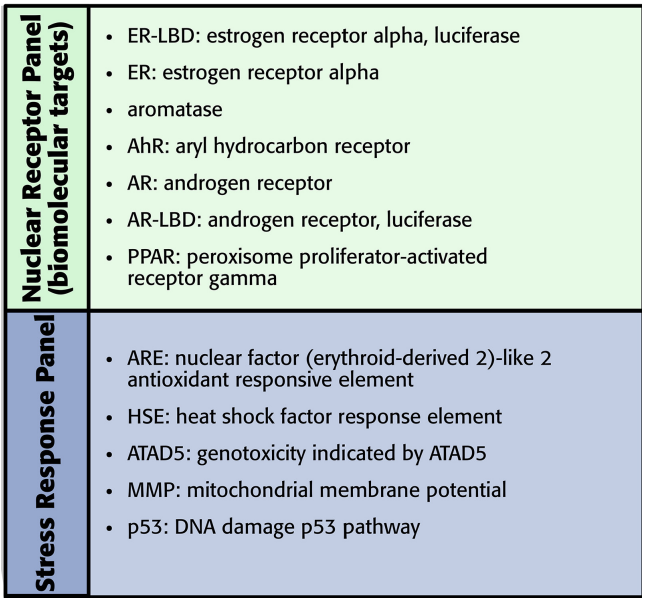


Figure 3.2: NR signaling and SR pathways

properties which are activity score, potency, and efficacy [73], therefore, these properties are considered as target classes and predicted individually by using five decision methods in a regression phase. Models used for classification and regression are same that are linear model, decision tree, random forest, support vector machine, and neural network [30]. By examining these models, our objective to build up an efficient binary classification model for activity forecast whether a given specific compound is active (1) or inactive (0). Simultaneously, our objective to build up an efficient regression model for the prediction of activity score, potency, and efficacy. Activity is the binary class and activity score, potency and efficacy are the continuous classes. The significant contributions of this work are as follows:

1. We developed better activity, activity score, potency, and efficacy assessment features, methods, and algorithms for drug molecules of ER.

2. We created a new method for assessing chemical activity which will have the potential to improve the procedure followed by scientists to evaluate environmental chemicals and develop new medicines.

3. We developed a framework to quick and efficient testing of certain chemical compounds that have probable chances to disrupt processes in the human body.

4. We developed a stand-alone application(s) that helps the researchers, and research community to predict the activity of the newly discovered chemical compound.

5. We developed a computational method (in silico) for checking the activity of drug

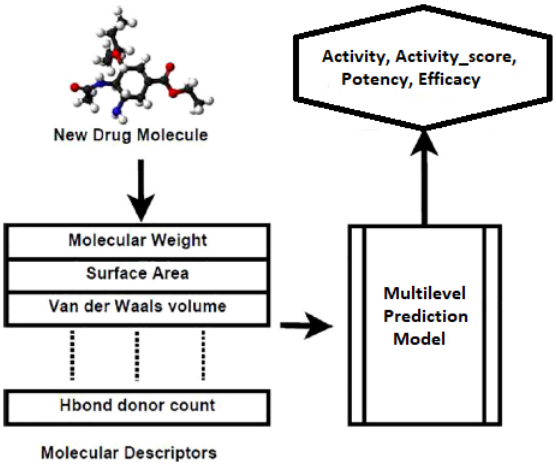


Figure 3.3: Multilevel prediction method

molecules rather than inside the living animal (in vivo) or in glass (in vitro) to save the life of animals, and money.

Figure 3.3 shows the general diagram of the prediction model where we are taking the various physicochemical properties of any small drug molecules, and our proposed multilevel prediction model is predicting its activity, activity score, potency, and efficacy.

The paper is composed as follows. A quick overview of the dataset, feature extraction using Pharmaceutical data exploration laboratory (PaDEL), feature selection, class imbalance problem, and about the target classes are introduced in Section 3.2. The procedure to generate the proposed multilevel prediction model is clarified in Section 3.3. The description of various machine learning models used in this work is presented in Section 3.4. Various model evaluation parameters for classification and regression are presented in Section 3.5. Section 3.6 describes the investigated, compared, and validated results which are followed by the discussion according to the analysis of results in Section 3.7. At last, conclusion is presented in Section 3.8.

3.2 Materials and Methods

3.2.1 Pharmaceutical data exploration laboratory (PaDEL)

PaDEL [2] software is used to compute molecular descriptors and fingerprints. The input of PaDEL-Descriptor is structure-data file (SDF) of ER drug molecules, and its output is a CSV file that contains total 8481 drug molecules and 1444 features. The PaDEL-Descriptor is a java based free and open source software similar to Dragon, MOE, Grid,

MARVIN Beans and supports more than 90 different molecular file formats like PDB, SDF, SMILE, and so on. The descriptors are computed using the Chemistry Development Kit library of java, used internally in PaDEL for features extraction related to Chemoinformatics and Bioinformatics files. PaDEL software can calculate 1876 molecular descriptors (1444 1D, 2D descriptors, and 431 3D descriptors), and 12 kinds of fingerprints. We are using only 1444 1D and 2D descriptors in our dataset for quality prediction of drug molecules. Figure 3.4 shows the graphical user interface of PaDEL-Descriptor [2], and Figure 3.5 shows the format of the structure-data file for a single drug molecule of ER. The drug molecule is active because the value of NR-ER is one in this file.

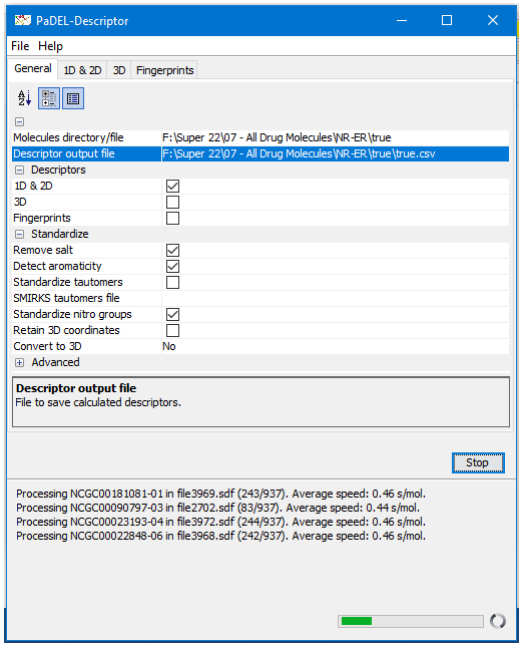


Figure 3.4: PaDEL-Descriptor GUI [2]

3.2.2 Dataset

The ER signaling pathway dataset is taken from PubChem, which is maintained by the National Center for Biotechnology Information (NCBI). It provides access to biomedical and genomic information from 'https://pubchem.ncbi.nlm.nih.gov/bioassay/743122'. In this study, our dataset consists total 8481 ER's drug molecules where 1084 are active molecules and remaining 7397 are inactive molecules [73]. All the drug molecules having 1444 features that are also called physicochemical properties or molecular descriptors which are extracted by PaDEL-Descriptor. The most common molecular descriptors are the partition coefficient (AlogP), molar refractivity (MR), surface area, volume, elements count, ETA descriptors, autocorrelation, nBase, nRing, Apol, the suitable duration of

Table 3.1: Physicochemical properties of ERs drug molecules

SN	Name	Description
1	Crippen logP	Atom-based calculation of logP using Crippen method, it is also called the octanol/water partition coefficient
2	Eccentric connectivity index	It is a distance-based atomic descriptor that is used for numerical modelling of biological activities of various nature
3	Fragment complexity	It reduces the 'interaction' complexity that correlates with the increased probability of achieving binding to a target
4	Kappa shape indices	The Kappa shape records are the premise of a technique for molecular structure quantisation in which characteristics of molecular shape are encoded into three indices (Kappa values)
5	Molecular linear free energy relation	These descriptors are intended to reflect the fundamental molecular properties, critical in solvation-related procedures, to be specific, polarity, size, and hydrogen bonding
6	Weighted path	To describe molecular descriptors for structure-property-activity studies, we use weighted-path numbers
7	Charged partial surface area	These descriptors were initially designed for studies of structure-physical relationship. It captures information about those feature of molecules which are responsible for polar intermolecular interactions
8	MR	Molecular refractivity is a measure of the aggregate polarisability of a mole of a substance. It is dependent on the pressure, temperature and the index of refraction
9	Extended topochemical atom (ETA)	Index for modelling chemical and drug-induced toxicities and some physicochemical properties relevant to such toxicities.
10	Autocorrelation (ATS_0, ATS_1)	Index that measures the linear relationship between lagged values of a time series y.

```
NCGC00161831-02
 OpenBabel06141707472D

 28 29  0  0  1  0  0  0  0  0999 V2000
    1.5474   -2.0671    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.2691   -1.6571    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    0.8372   -1.6571    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.2691   -3.3084    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.8372   -0.8372    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.9851   -2.0671    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.7068   -1.6571    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.5474   -2.8985    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.9851   -2.8985    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.2691   -4.1283    0.0000 C   0  0  1  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
  1  8  2  0  0  0  0
  2  6  2  0  0  0  0
  3  5  1  0  0  0  0
  3 19  2  0  0  0  0
  4  8  1  0  0  0  0
  4  9  2  0  0  0  0
  4 10  1  0  0  0  0
  5 13  1  0  0  0  0
M  END
>  <Formula>
C17H17ClF6N2O

>  <FW>
414.7731 (378.3122+36.4609)
>  <DSSTox_CID>
27796
>  <NR-AR>
0
>  <NR-ER>
1
$$$$
```

Figure 3.5: SDF format for single molecule of ER

action, and so on. Table 3.1 describes some essential physicochemical properties of ER's drug molecules.

Table 3.2 shows the glance of the dataset that contains various ER drug molecules and their molecular descriptors which are extracted from the SDF file using PaDEL-Descriptor. NCGC00181290-01, NCGC00181294-01, NCGC00181300-01, and so on are the name and ATS0m, AATSC8m, Mlogp, and so on are the various molecular descriptors of drug molecules. Table 3.3 shows only the different target classes which have been predicted using our proposed multilevel prediction model in its classification and regression phases.

Table 3.2: ER dataset

Name of drug molecules	Activity	ATS0m	AATSC8m	GATS5m	MLogP
NCGC00181290-01	0	3820.092	-0.0384286	0.790592	3.22
NCGC00181294-01	0	8146.0126	-1.2130702	1.4757891	2.56
NCGC00181300-01	0	8149.8870	-1.6515457	1.0599761	3.55
NCGC00257625-01	1	3829.9645	-0.4527243	0.8136134	3.44
NCGC00259354-01	1	2901.5394	2.3427338	1.2072558	3.66
NCGC00255335-01	1	3909.3741	2.7250422	0.8870279	3.55

Table 3.3: Target classes for ER dataset

Name of drug molecules	Activity score	Activity	Potency	Efficacy
NCGC00181290-01	0	-	-	-
NCGC00181294-01	0	-	-	-
NCGC00181300-01	0	-	-	-
NCGC00013034-01	1	44	4.2163	124.545
NCGC00016436-01	1	90	0.0032	389.944
NCGC00015234-11	1	49	11.8832	349.98

3.2.3 Class imbalance

Most machine learning algorithms perform well when the number of instances of each class is roughly equal. In class imbalance dataset, the number of instances in one class far exceeds the other class. It is a problem in machine learning where the main class of interest is rare then this triggers biasing of the classifier. In class imbalance, the dataset distribution reflects a significant majority of the negative class and the minority of the positive class. This issue is extreme when the extent of the minor class is around 10%. Here, our dataset for the prediction of activity contains two classes; one is active, and other is inactive. These classes are extremely imbalanced because of the number of active molecules 1084 which are far less than the number of inactive molecules 7397. The primary function of class balancing is to balance the class symmetry of instances [79]. There are several approaches to handling the class imbalance problem which is undersampling, oversampling, and synthetic minority oversampling technique (SMOTE).

Here, we applied SMOTE algorithm on the dataset where minority class is oversampled by creating 'synthetic illustration' instead of oversampling it with substitution [71]. SMOTE is used to avoid overfitting which occurs when exact copies of minority instances are added to the main dataset. A part of data is found from the minority class as an example and after that new synthetic similar instances are generated. These synthetic instances are then added to the original dataset. Now, the new dataset is utilized as a sample to train the classification models [29]. SMOTE is performed using the unbalanced package of R.

3.2.4 Feature selection using FSelector package

To remove various discrepancies from the dataset, first, we preprocess it then proceeds for the model formation. During the process of model building, feature selection is used to filter out high correlated variables, descriptors with too many zero values, missing values, categorical data and unwanted noise from the dataset. As we discussed that our dataset has 1444 features which are large in quantity, therefore, it will create a more time complexity in model execution. Feature selection is a process where we have to select essential features that may improve the performance of the model and remove those attributes that have irrelevant and redundant information.

The process of feature selection is carried out using FSelector package in R. FSelector package has various functions for selecting important attributes from a given dataset. We are using cfs() and information.gain() methods of this package which are the correlation-based and entropy-based methods, respectively. The central hypothesis of correlation-based feature selection (CFS) says that good features set contain only those features which are highly correlated with the target class but uncorrelated with other features of the dataset [80]. CFS finds an attribute subset using correlation coefficient for discrete and continuous data. A correlation coefficient is used to estimate the correlation between the subset of attributes and target class [81] (refer to Subsection 3.5.2.2 for correlation). Information gain is an attribute selection measure which is based on information theory where the attribute with the highest information gain is selected as the splitting attribute. This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least entropy (randomness or impurity) in these partitions. Info(D) is just the average amount of information needed to identify the class label of a

tuple in dataset D.

$$Info(D) = -\sum_{i=1}^{m} p_i log(p_i) \tag{3.1}$$

Here, p_i is the nonzero probability that an arbitrary tuple in D belongs to class Ci. Note that, at this point, the information we have is based only on the proportions of tuples of each class. Info(D) is also known as the entropy of D [30]. We applied CFS method and information.gain method for feature selection during classification level and regression level, respectively.

CFS has two inputs parameters one of them is the name of the dataset, and another is target class (activity). This process marks 43 features which have higher importance as compared to others. These features are:
'nS', 'nCl', 'ATS0m', 'AATS3v', 'AATS7v', 'AATS8v', 'AATS5p', 'AATs8p', 'Mi', 'AATS1i', 'AATS2i', 'AATS5i', 'ATSC1m', 'ATSC2m', 'ATSC3m', 'ATSC4m', 'ATSC5m', 'ATSC6m', 'ATSC8m', 'ATSC1p', 'ATSC5p', 'ATSC6p', 'ATSC7p', 'MLogP', 'AATSC1m', 'AATSC2m', 'AATSC3m', 'AATSC4m', 'FMF', 'AATSC4p', 'MATS2c', 'MATS5m', 'GATS2c', 'GATS4m', 'GATS2p', 'C3SP2', 'MIC2', 'VE3_D' 'CrippenLogP', 'AATSC8m', 'ZMIC2', 'ZMIC4', 'AATSC0m'.
These features are used for the activity prediction, using classification models.

Feature selection during regression phase is performed using entropy-based information.gain() method. This function also has two input parameters which are the dataset and its target class. The target class is anyone from activity score, potency, and efficacy. Table 3.4, Table 3.5, and Table 3.6 show the features and their importance in terms of weights for activity score, potency, and efficacy, respectively. During model building, only those combinations of features are selected which are giving highest accuracy for the prediction of activity score, potency, and efficacy. Finally, the first 245 features are selected for activity score class, first 287 features are selected for potency class, and first 102 features are selected for efficacy class. These all the combinations of attributes are gave the best accuracy than other combinations of attributes. The cutoff.k method of FSelector package selects k best attribute from the ranked attributes [80]. Table 3.7, Table 3.8, and Table 3.9 show the results of various evaluation parameters that are root mean square error (RMSE), correlation (r), coefficient of determination (R^2), accuracy, and total time based on the different combinations of features.

Table 3.4: Activity score features and their importance

SN	Features	Attribute Importance
1	ATS5i	0.4063393296
2	ATS3v	0.3964188662
3	ATS6i	0.3875952976
4	GGI4	0.3850211282
5	ATS4i	0.3844582355
.	.	.
.	.	.
.	.	.
453	GATS3m	0.0453890132
454	VE1_Dt	0.0438046882
455	VE2_Dt	0.0431102494
456	C2SP1	0.0429399403
457	AATSC0m	0.0409700875

Table 3.5: Potency features and their importance

SN	Features	Attribute Importance
1	MDEC.34	0.3699073829
2	MDEC.24	0.3687875474
3	ATS4i	0.3631291316
4	MDEC.14	0.3336430767
5	nTRing	0.3291762024
.	.	.
.	.	.
.	.	.
420	AATSC7i	0.0487871025
421	MIC2	0.0454323053
422	AATSC7p	0.0436585766
423	GATS3m	0.040835543
424	MATS7p	0.040702876

Table 3.6: Efficacy features and their importance

SN	Features	Attribute Importance
1	GGI2	0.2864897673
2	ATS5i	0.2644866356
3	MDEC.34	0.2607405716
4	MWC6	0.2603965535
5	MWC5	0.2602291198
.	.	.
.	.	.
.	.	.
331	GATS8v	0.0514879466
332	AATS8p	0.0509053131
333	AATS3i	0.0472519257
334	MATS1i	0.0439847613
335	ATSC8p	0.0407595757

Table 3.7: Impact of features on random forest performance for activity score

No. of features	r	R	RMSE	Accuracy	TotalTime (Seconds)
245	0.84	0.71	5.31	**81.54**	0.699
345	0.87	0.76	5.03	81.54	0.867
59	0.88	0.77	5.54	79.23	0.542
178	0.9	0.81	5.34	78.46	0.671
211	0.9	0.81	5.11	78.46	0.663
263	0.91	0.83	4.88	78.46	0.789
273	0.86	0.74	5.66	78.46	0.782
384	0.89	0.79	5.22	78.46	0.876
57	0.87	0.76	5.5	78.46	0.581
111	0.87	0.76	5.19	77.69	0.635

Table 3.8: Impact of features on random forest performance for potency

No. of features	r	R	RMSE	Accuracy	TotalTime (Seconds)
287	0.71	0.5	7.43	**88.79**	0.659
142	0.72	0.52	8.02	86.92	0.544
15	0.73	0.53	7.15	86.92	0.467
182	0.71	0.5	6.35	86.92	0.565
339	0.82	0.67	6.73	86.92	0.674
105	0.74	0.55	6.94	85.98	0.523
129	0.7	0.49	6.96	85.98	0.574
169	0.71	0.5	6.91	85.98	0.607
208	0.68	0.46	7.19	85.98	0.551
220	0.67	0.45	7.72	85.98	0.56

Table 3.9: Impact of features on random forest performance for efficacy

No. of features	r	R	RMSE	Accuracy	TotalTime (Seconds)
102	0.77	0.59	52.46	**80.27**	0.511
108	0.78	0.61	54.21	78.5	0.537
205	0.78	0.61	53.12	78.5	0.55
58	0.71	0.5	55.34	78.5	0.507
100	0.74	0.55	57.13	77.57	0.502
174	0.77	0.59	52.04	77.57	0.577
281	0.73	0.53	60.78	77.57	0.586
70	0.75	0.56	54.3	77.57	0.532
120	0.78	0.61	50.84	76.64	0.545
171	0.79	0.62	52.14	76.64	0.574

3.2.5 Target class used in classification dataset

Activity is the target class that contains its two instances which are active (1) and inactive (0). The activity reflects the quality of the drug molecules. Active compound has the ability to binds with the ER and produces various estrogenic effects by modulating the activity of ER, and the opposite of it, inactive compound cannot bind with ER [73].

3.2.6 Target Classes used in Regression Dataset

1. **Activity score:** It is a clinical index of molecule activity that combines information from its target response. Activity score of all active drug compounds range from 40 to 100, and for all inactive compounds, its value is 0. Inconclusive compounds have their activity score between 1 to 39 [73].

2. **Potency:** Amount of a drug that is needed to produce a given effect. A lowly potent drug (e.g., fentanyl, alprazolam, risperidone) provides a given response at higher concentrations, while a highly potent drug (e.g., codeine, diazepam, ziprasidone) evokes the same reaction only at lower concentrations [73].

3. **Efficacy:** In pharmacology, it is a maximum response that a particular drug is capable of producing. The effect of the drug is achieved against an applied dose. A

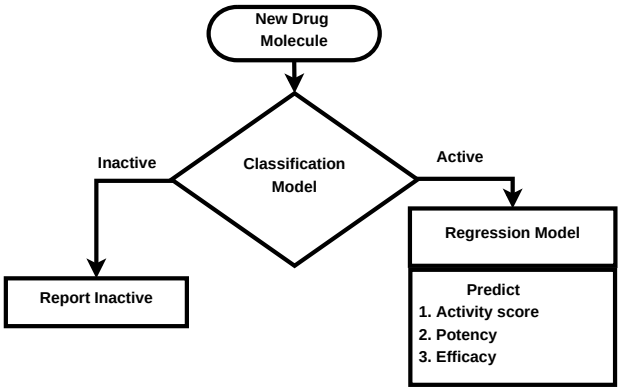


Figure 3.6: Flowchart for classification and regression models of ER's drug molecules

high efficacy means that a drug has an extreme ability to initiate a response after binding to a receptor [73].

Figure 3.6 shows the flowchart of the proposed multilevel prediction model where a new drug molecule is being classified using a classification model in active and inactive categories. Further, if a drug molecule is found active, then the regression model predicts its activity score, potency, and efficacy.

3.3 Proposed Multilevel Prediction Model

The proposed multilevel prediction model has two levels where its first level is for classification and the second level is for regression. Figure 3.7 shows the methodology of the proposed multilevel prediction model. It has seven steps, in which its initial five steps are described in Section 3.3.2, Step 6 is described in Section 3.3.3, and Step 7 is defined in Section 3.3.6.

Level 1: Classification based model building

The dataset of ER's drug molecules is found from the PubChem website of NCBI. The dataset does not have equal numbers of active and inactive drug molecules; first we have balanced the dataset using SMOTE algorithm. After overcoming this problem, we observed that our dataset is high in features; therefore, the CFS algorithm is applied (refer to Section 3.2.4 for details). After all these data preprocessing techniques on the dataset, we trained and tested the dataset at 70% and 30%, respectively by five decision methods which are decision tree, linear model, neural network, support vector machine, and random forest. The ensemble-based random forest model has outperformed the other models for various performance metrics of classification. Here, the total size of the dataset

is 8481 in which the size of the training dataset will be 5937 and size of the testing dataset will be 2544. Table 3.10 shows the size of the training and testing dataset.

Table 3.10: Training-Testing dataset for activity prediction

No. of drug molecules	Molecule name	Activity
Total training data (5937)	NCGC00255193-01	1
	NCGC00257825-01	0
	NCGC00257439-01	1
	NCGC00256532-01	0
	.	.
	.	.
	NCGC00260123-01	1
	NCGC00255292-01	0
	NCGC00254263-01	1
Total testing data (2544)	NCGC00255471-01	0
	NCGC00258791-01	0
	NCGC00255327-01	1
	.	.
	.	.
	NCGC00258791-01	0
	NCGC00255278-01	1

Level 2: Regression based model building

All the drug molecules of the dataset which are found active during the classification phase are shortlisted. Now, our new dataset is already balanced and has maximum 1084 drug molecules, and each has 1444 features. This dataset again very high in features, therefore, we performed feature selection using information.gain method (refer to Section 3.2.4 for details). Here, the target classes are activity score, potency, and efficacy for regression models. We again trained and tested the dataset at 70% and 30%, individually for all the target classes again by five decision methods that are decision tree, linear model, neural network, support vector machine, and random forest. The random forest model still surpasses among other models for various performance metrics of regression. Here, the total size of the dataset is 1084 in which the size of the training dataset will be 759 and size of the testing dataset will be 325. Table 3.11 shows the size of the training and testing dataset.

3.4 Machine Learning Models

Performance analysis such as Gini coefficient, specificity, sensitivity, precision, AUC, accuracy and total execution time are evaluated for classification models. Similarly, RMSE, correlation(r), coefficient of determination(R^2), accuracy and total execution time are evaluated for regression models. Models which are used for classification of activity and regression of activity score, potency, and efficacy are explained below. Some parameters of the models are tuned to find the better prediction outcome. Table 3.12 shows the

Table 3.11: Training-Testing dataset for activity score, potency, and efficacy prediction

No. of drug molecules	Molecule name	Activity score	Potency	Efficacy
Total training data (759)	NCGC00013034-01	44	4.2163	124.545
	NCGC00013235-01	42	4.2163	52.5472
	NCGC00016436-01	90	0.0032	389.944
	NCGC00015234-11	49	11.8832	349.98
	.	.	.	.
	.	.	.	.
	NCGC00013489-01	42	33.4915	155.466
	NCGC00013718-01	41	33.4915	72.1084
	NCGC00013725-01	41	33.4915	62.0537
Total testing data (325)	NCGC00015234-11	49	11.8832	349.98
	NCGC00015829-02	45	10.1235	182.719
	.	.	.	.
	.	.	.	.
	NCGC00015830-02	41	11.3588	29.7936

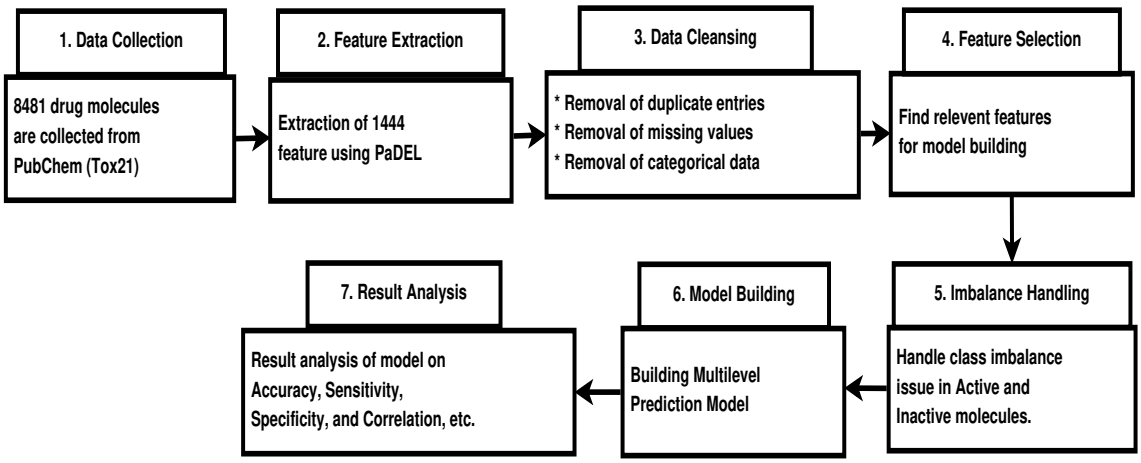


Figure 3.7: Methodology used for the proposed multilevel prediction model

various models with their required packages and tuning parameters, used in the proposed multilevel prediction model. These all models are implemented in R, under GNU general public license.

1. Decision tree (rpart): We used various algorithms called ID3, C4.5, and CART for building the decision tree which is developed by J. R. Quinlan. These algorithms are based on top-down, recursive, and divide and conquer approach [30].

2. Linear model (lm): The linear model presents the connection between a dependent variable Y as a function of at least one independent factors X by fitting the best line. This best fit line is known as the regression line and represented by a linear equation Y= a * X + b + ϵ, where ϵ is an error. Here, we are using multi-linear regression because of the number of independent variables are more than one [26].

Table 3.12: Machine learning models used and their tuning parameters

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

3. Neural networks (nnet): Here, we are using backpropagation algorithm for neural network based classification and regression. For each training tuple, the weights are updated to minimize the mean-squared error (MSE) between the systems expectation and the real target value [30].

4. Support vector machine (svm): Support vector machine can be used for classification as well as regression. It represents the input features/molecular descriptors as vectors which are projected onto higher-dimensional space. An optimal hyperplane is then constructed for separating the active and inactive drug molecules [30].

5. Random forest (randomForest): Random forest is an aggregate classifier that is the collection of several decision trees. Random forest is an ensemble based model where each tree votes and most popular class is returned during classification. The parameters of randomForest function is mtry, and ntree, where mtry is the number of features that randomly sampled as candidates at each split, and ntree is the number of trees. The time complexity of RF algorithm is: O(ntree *mtry *d *n) and space complexity of RF algorithm is: O(n*d), where n is the number of records and d is the depth of the tree. It is shown that random forest depends on the depth and size of the tree [79][30].

3.5 Model Evaluation Parameters

3.5.1 Classification model parameters

3.5.1.1 Gini coefficient

Gini coefficient is used to measure the distribution inequality of data [32]. Gini values range between 0 and 1. The 0 value of Gini expresses perfect equality of data. Assuming a model M has a Gini coefficient 0.6 and model D has a Gini coefficient 0.45, at that

point model M, is considered as a productive model in contrast to model D.

3.5.1.2 Sensitivity

Sensitivity (Sens) is also known as true positive rate (recognition) or recall [30]. It is the ratio of actual positives which are correctly identified as positives by the classifier. It is computed as

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.2}$$

3.5.1.3 Specificity

Specificity (Spec) is also known as the true negative rate [26]. It is the ratio of actual negatives which are correctly identified as negatives by the classifier. It is computed as

$$Specificity = \frac{TN}{TN + FP} \tag{3.3}$$

3.5.1.4 Precision

Precision can be thought of as a measure of exactness; it means what percentage of tuples labeled as positive are actually positive [26]. It is computed as

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

3.5.1.5 Area under the curve

An AUC measures the quality of the classifier. The receiver operating characteristics (ROC) is a curve which is drawn between true positive rate (TPR) and false positive rate (FPR). These parameters are calculated from the confusion matrix analysis. The amount of area under the ROC is called AUC. AUC value ranges between 0 and 1. The quality of a model is excellent if it has AUC value close to 1. The model which is scoring high AUC as compared to another model is considered as an efficient model [30][32].

3.5.1.6 Accuracy

Accuracy is the most important criteria to measure the exactness of any classifier [30]. The accuracy can be computed as

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100 \tag{3.5}$$

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

3.5.2 Regression model parameters

3.5.2.1 Root mean square error

RMSE is also called root mean square deviation (RMSD). It is used to compute the error in the regression model. RMSE has been used to calculate the difference between actual and predicted values [82]. Our target is to reduce RMSE as low as possible

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \tag{3.6}$$

where p is the predicted target and a is the actual target.

3.5.2.2 Correlation (r)

The extent to which actual and predicted values are related is declared by correlation. It is defined as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.7}$$

where x is the actual value and y is the predicted value. $\overline{x}$ is the mean of the all actual values and $\overline{y}$ is the mean of the all predicted values. The number of cases is n. Correlation consists of values from -1 to 1. If it's value is 1 then it is considered as good correlation [83].

3.5.2.3 Coefficient of determination (R^2)

The R^2 is utilized in statistical analysis that assesses how well a model explains and predicts the future outcomes. It described the proportion of the variance of the dependent variable explained by the regression model [83]. If the regression model is perfect then R^2 is 1 and in the case of total failure, its value will be 0

$$R^2 = r * r \tag{3.8}$$

3.5.2.4 Accuracy

It is computed as percentage deviation of the predicted target to the actual target with some acceptable error [32].

$$\begin{aligned} Accuracy &= \tfrac{100}{n} \textstyle\sum_{i=1}^{n} q_i \\ q_i &= \begin{cases} 1 & \text{if } abs(p_i - a_i) \leq err \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{3.9}$$

where a is actual target, p is predicted target, err is the some acceptable error, and n is the total number of instances.

3.5.3 K-fold cross-validation

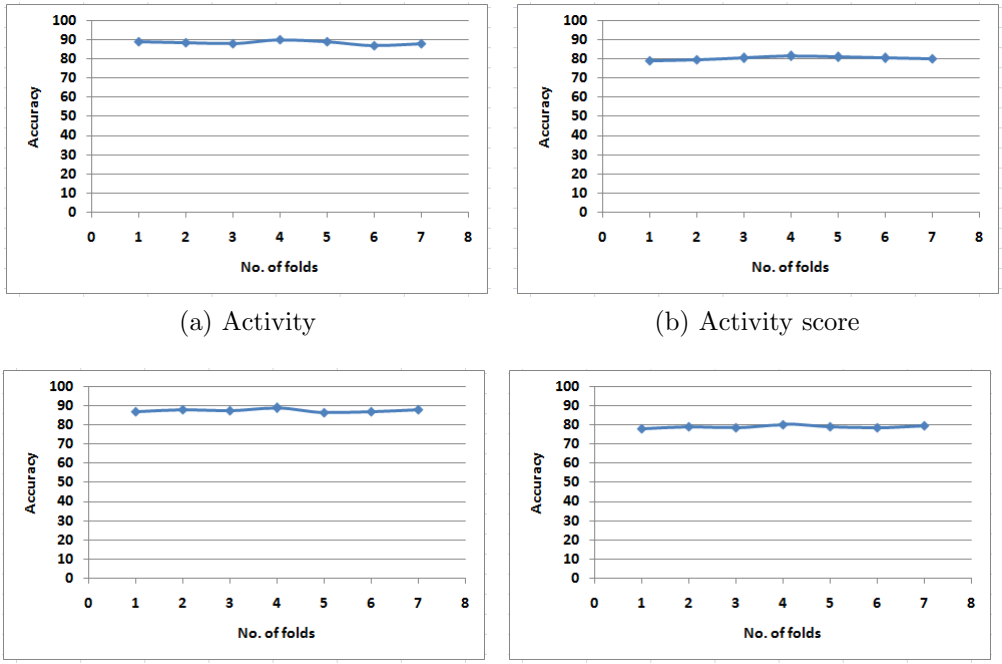
Cross-validation is a beneficial technique for assessing the performance of machine learning models. The K-fold cross-validation approach partitions the dataset into K equal-sized segments. During each execution, one of the partition is chosen for testing, while the rest of the partitions are used for training. This procedure is repeated K-times so that each partition is used for testing exactly once. In each fold, random data is provided for training and testing to measure the robustness of the model. The proposed multilevel prediction model shows the consistent performance for the accuracy measure using K-fold cross-validation [30]. Generally, the value of K is taken to be 10, but it is not a strict rule; therefore, K can have any value. Here, we are considering, K=7.

3.6 Result Analysis, Comparison, and Validation

This section presents the analysis of prediction results of five machine learning models that are random forest, linear model, support vector machine, neural network, and decision tree for classification and regression.

The comparative performance of all the classification models for the prediction of activity is analyzed by Gini coefficient, specificity, sensitivity, precision, AUC, accuracy and total time parameters. All these performance evaluation parameters are described in Section 3.5.1. Table 3.13 shows the results of these performance parameters for various models where the ensemble based random forest model is outperformed the other four models for the testing dataset. The Gini coefficient, sensitivity, specificity, precision, AUC, accuracy, and total time are calculated through random forest model which is 0.859, 0.86, 0.891, 0.94, 0.86, 88.55%, and 7.45 seconds, respectively.

The comparative performance of all the regression models for the prediction of activity score, potency, and efficacy are analysed by RMSE, correlation, R^2, and accuracy. All these model performance evaluation parameters are described in Section 3.5.2. Table 3.14 shows the results of these performance parameters for various models. RMSE has been calculated using Eq.(6), and Table 3.14 lists the RMSE of all the models. Random forest has the lowest RMSE of 5.31, 7.43, and 52.46 for the prediction of activity score, potency, and efficacy, respectively, on the testing dataset. The correlation has been calculated using Eq.(7), and Table 3.14 lists the correlation of all the models. Random forest has the almost highest correlation of 0.84, 0.71, and 0.77 for the prediction of activity score, potency, and efficacy, respectively, on the testing dataset. The R^2 has been calculated using Eq.(8), and Table 3.14 lists the R^2 of all the models. Random forest has the highest R^2 of 0.71, 0.5, and 0.59 for the prediction of activity score, potency, and efficacy, respectively, on the testing dataset. The accuracy of all regression models has been calculated using Eq.(9) with some acceptable errors, and Table 3.14 lists the accuracy of all the models. Random forest has the highest accuracy of 81.54% (with ± 7 err), 88.79% (with ± 20 err), and 80.27% (with ± 50 err) for the prediction of activity score, potency, and efficacy, respectively, on the testing dataset. Table 3.7, Table 3.8, and Table 3.9 show the total time taken by the random forest model on the testing dataset, according to the different number of features are selected. Total time taken by classification models is more because these models will execute more than 8481 drug molecules (active and inactive) and similarly, total time taken by regression models is less because of these models will run only 1084 drug molecules (active). We measure the execution time of models in seconds.



(a) Activity (b) Activity score

(c) Potency (d) Efficacy

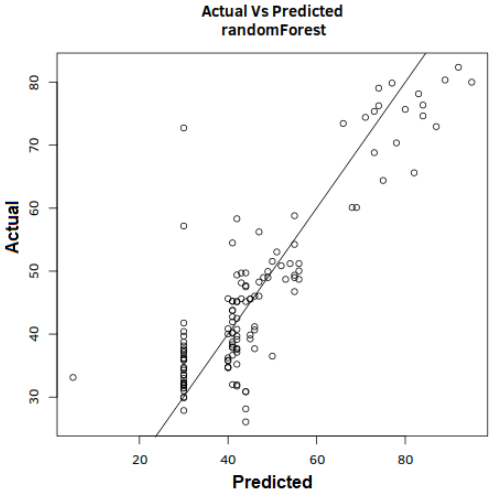
Figure 3.8: K-Fold cross-validation

Table 3.13: Classification dataset results by various decision methods

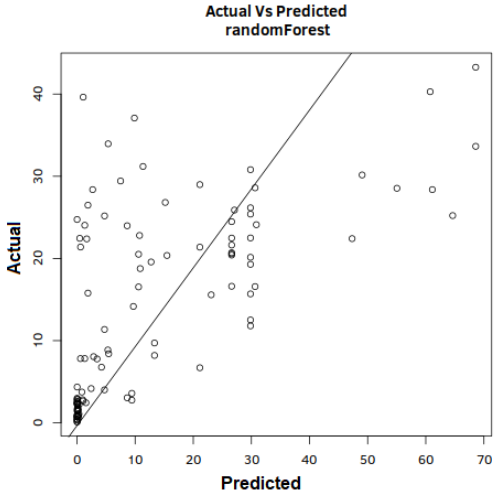
	Activity						
Model Name	**Gini coefficient**	**Sensitivity**	**Specificity**	**Precision**	**AUC**	**Accuracy**(%)	**TotalTime (Seconds)**
Random Forest	0.859	0.86	0.891	0.94	0.86	**88.55**	7.45
Decision Tree	0.749	0.815	0.737	0.942	0.79	86.78	8.86
Linear Model	0.817	0.531	0.513	0.953	0.54	83.26	6.14
Neural Network	0.813	0.895	0.884	0.944	0.83	82.34	4.06
Support Vector Machine	0.796	0.705	0.939	0.919	0.893	82.4	7.36

Table 3.14: Regression dataset results by various decision methods

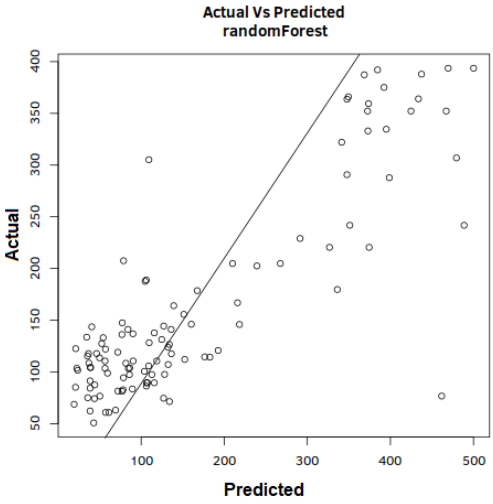
	Activity score				**Potency**				**Efficacy**			
Model Name	**r**	R^2	**RMSE**	**Accuracy** (%)	**r**	R^2	**RMSE**	**Accuracy** (%)	**r**	R^2	**RMSE**	**Accuracy** (%)
Random Forest	0.84	0.71	5.31	**81.54**	0.71	0.5	7.43	**88.79**	0.77	0.59	52.46	**80.27**
Decision Tree	0.80	0.64	6.39	77.69	0.55	0.3	8.13	84.11	0.85	0.72	52.88	79.62
Linear Model	0.61	0.37	8.1	67.69	0.63	0.4	9.08	84.00	0.52	0.27	69.21	71.96
Neural Network	0.83	0.69	6.12	78.28	0.53	0.28	8.19	82.18	0.53	0.28	71.89	69.16
Support Vector Machine	0.81	0.66	7.5	80.28	0.57	0.32	8.79	83.18	0.59	0.35	61.36	74.96



(a) Activity score



(b) Potency



(c) Efficacy

Figure 3.9: Scatter plots

To measure the robustness of the random forest, we use a 7-fold cross-validation method for all the target classes. Figure 3.8 shows the results of 7-fold cross validation for the prediction of activity, activity score, potency, and efficacy. These results of accuracy measure demonstrate the consistent performance of the random forest model for different folds of the dataset. Scatter plot is used to show the relationship between two variables. Scatter plot is sometimes called a correlation plot because it shows how two variables are correlated. Figure 3.9 shows the scatter plots for R^2 parameter. These scatter plots show the relationship between actual and predicted values of activity score, potency, and efficacy classes for random forest model, and visualized using results of the testing dataset where points that meet the line have the better correlation.

3.6.1 Validation of proposed multilevel prediction model

Validation of proposed multilevel prediction model means that we are testing the performance of this model on some new drug molecules' dataset which has similar features but not part of the actual training and testing dataset. If the prediction accuracy of this model on these new drug molecules is similar to our testing dataset to some extent, then we can say that our proposed multilevel prediction model has been validated. Here, we are validating our proposed model on some AIDS therapy's and androgen receptors' drug molecules which are unknown for our multilevel prediction model. Non-nucleoside and nucleoside reverse transcriptase inhibitors (NNRTIs) are the main drugs available to treat HIV infection. It is corresponding to Etravirine (ETR), Rilpivirine (RPV), Lersivirine are the essential medicines for AIDS therapy. These drugs show potent anti-HIV-1 activity and its modest toxicities [84]. Etravirine has been associated with hypersensitivity reactions that can be associated with liver injury including acute liver failure [85]. Rilpivirine can cause severe and life-threatening side effects which include severe skin rash, depression, mood changes, and liver problems [86]. Similarly, Lersivirine is also reported nausea, headache and skin rashes [87].

Figure 3.10 shows the complete process of validation of the proposed multilevel prediction model on the three-drug molecules of AIDS therapy namely Etravirine (ETR), Rilpivirine (RPV), and Lersivirine. We have downloaded the two-dimensional structure of these three drug molecules and some drug molecules of androgen receptors [88] from the PubChem website in SDF format. Molecular descriptors of these drug molecules have been extracted with the help of PaDEL-Descriptor [79]. Now, we applied the proposed multilevel prediction model for activity prediction of all these drug molecules. If the drug molecule is inactive, then there is no need to find its activity score, potency, and efficacy

Table 3.15: Activity, activity score, potency and efficacy prediction of some new drug molecules for validation

	Activity		Activity score		Potency		Efficacy	
Molecule Name	**Actual**	**Predicted**	**Actual**	**Predicted**	**Actual**	**Predicted**	**Actual**	**Predicted**
ETR	1	1	-	-	-	-	-	-
RPV	1	1	-	-	-	-	-	-
Lersivirine	1	1	-	-	-	-	-	-
NCGC00260548-01	1	1	44	43	2.1054	2.11	60.8775	58.258
NCGC00257038-01	1	1	41	42	30.6379	29.85	94.4017	93.57
NCGC00013012-01	0	0	NA	NA	NA	NA	NA	NA
NCGC00013015-01	0	0	NA	NA	NA	NA	NA	NA
NCGC00013037-01	0	0	NA	NA	NA	NA	NA	NA
NCGC00013042-01	0	0	NA	NA	NA	NA	NA	NA
NCGC00015060-14	0	0	NA	NA	NA	NA	NA	NA

but if a drug molecule is reported active, then there is need to run the regression model to predict its activity score, potency, and efficacy. Table 3.15 shows the output of proposed multilevel prediction model for the activity prediction of various drug molecules where our proposed model shows that drug molecules which are predicted active are actually found to be active, and the drug molecules which are predicted inactive are actually found to be inactive. These correct predictions of all the new drug molecules show the validity of the proposed model.

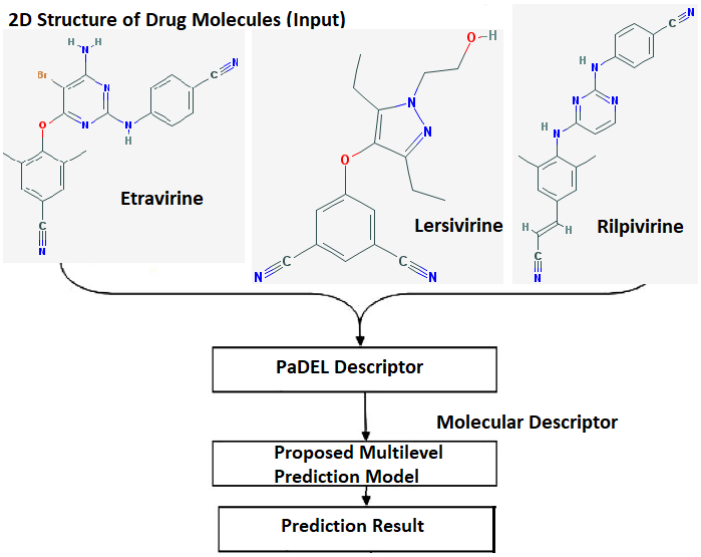


Figure 3.10: Activity prediction of AIDS therapy drug molecules using proposed multilevel prediction model

3.7 Discussion

The prediction of activity, activity score, potency, and efficacy of any drug molecule is important while deciding its estrogenic effects on human health. The results of our

proposed multilevel prediction model using random forest model are better in comparison to other existing models for both classification and regression phases. 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 3.12 is showing various models with their corresponding packages, methods, and tuned parameters. The randomForest method has 'mtry' and 'ntree' parameters for tuning, where mtry is showing 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 as large as to ensure that every input row gets predicted at least a few times. Therefore, we optimized the performance of random forest model by setting the value of mtry=2, and ntree=500 [89]. The lm method is used to fit the linear model. We improved the performance of lm by tuned the parameter 'method', which can take its two values that are method = 'qr' and method = 'model.frame', we have chosen value 'QR decomposition' because it is used to solve the linear least squares problem. The rpart method is based on recursive partitioning to build classification and regression trees. We improved the performance of rpart by tuned the parameter usesurrogate=0; then it reduced the computational time to search for surrogate splits [90] . The nnet method is used to fit a single-hidden-layer neural network which has various parameters where its 'size' parameter is showing the number of units in the hidden layer. It can be zero if there are skip-layer units. Here, we are taking the value of size=10, because at this value our nnet method is performing well [91]. The ksvm method of support vector machine can perform classification as well as regression. We improved the performance of ksvm by tuned the parameter 'kernel' and 'type'. The kernel function is used in the training and prediction, and its value can be of any function of the class kernel. The various kernels are rbfdot, polydot, vanilladot, tanhdot, laplacedot, besseldot, anovadot, stringdot, but we have chosen rbfdot (Radial Basis kernel function 'Gaussian') for better performance. The type parameter shows whether you want to perform classification or regression. C-svc and nu-svc values are for classification, while eps-svr and nu-svr values are for regression. Here, we have chosen C-svc for classification and nu-svr for regression [92]. These all parameters are tuned similarly for both the classification and regression models.

The performance analyses of binary class activity is evaluated using the metrics - Gini, sensitivity, specificity, precision, AUC, and accuracy. These are the terms which statistically measure the performance of the binary classification test, where we divide a given dataset into two categories on the basis of their common characteristic. Sensitivity indicates, how well the test predicts one category (active), and specificity measures how well the test predicts the other category (inactive). Whereas accuracy measures how well the

test predicts both categories (active and inactive). Therefore, an excellent binary classification test always results with high values for all three factors (sensitivity, specificity, and accuracy), whereas a poor binary classification test results with low values for all. But, an average binary classification test always results with average values which are almost similar for all the three factors [93]. When we consider the case of class imbalance problem where the main class of interest is rare. For example, in cancer detection applications, the class of interest (or positive class) is 'malignant', which occurs much less frequently than the negative 'benign' class. In this case, a classifier will correctly label only the benign tuples, and misclassifying all the malignant tuples. Therefore, apart from accuracy, we need of other measures, which accesses how well the classifier can predict the malignant tuples and how well it can recognize the benign tuples. For this purpose, only sensitivity and specificity measures can be used, respectively [30]. Since, the count of malignant candidates is very low and the count of benign is very high, then accuracy varies with specificity without considering sensitivity. Our dataset for classification is also imbalanced but using SMOTE algorithm, we have balanced it, and calculating its performance by accuracy. Similarly, the performance analysis of any model for continuous class classification, we always use the metrics which are correlation (r), coefficient of determination (R^2), root mean squared error (RMSE), and accuracy. The more significant value of r, R^2, accuracy and lower value of RMSE is better for any regression model. The total time to build the various models depends on the structure of the models and the size of the dataset. The classification models are taking more time because these are executing large dataset (8481 drug molecules) and regression models are taking less time because these are executing small dataset (1084 drug molecules).

3.8 Conclusion

In this paper, we proposed a novel machine learning based multilevel prediction model to the assessment of the quality of ER's drug molecules. This prediction is based on QSARs/QSPRs approaches where we have solved the problem of quality prediction of those drug molecules that can bind to estrogen receptors. The target classes for quality prediction are activity, activity score, potency, and efficacy. The dataset used in this study is exceptionally imbalanced in classes and very high in features. Initially, we balanced the dataset by SMOTE algorithm, and feature selection is performed by correlation-based feature selection and information.gain() methods. After all these data pre-processing, we developed a multilevel prediction model where we applied five models in the classification phase as well as similar five models in the regression phase. These models are decision

tree, linear model, neural network, support vector machine, and random forest. In the classification phase, all the models have been evaluated on Gini coefficient, sensitivity, specificity, precision, accuracy, and total time for the activity prediction.

Similarly, in the regression phase, all the models have been evaluated on RMSE, orrelation, R^2, and accuracy for the activity score, potency, and efficacy prediction. Through the intensive experiments, it concludes that the ensemble-based random forest method outperformed over other examined methods at both levels, and its performance is nearly linear in K-fold cross-validation. Therefore, the random forest model is used inside our proposed multilevel prediction model. Our proposed multilevel prediction model instead of having highly imbalanced data, gives the better accuracy for all the target classes as compared to the existing techniques that are a linear model, neural network, support vector machine, and decision tree. Finally, to prove the validity of the proposed model, we tested it on some AIDS Therapy's drug molecules and androgen receptor's drug molecules, which are not part of the actual dataset, where we found 100% accuracy for activity prediction. We believe that by utilizing some other feature selection method, some other data imbalance handling methods and ensemble learning with their optimized parameters may achieve better performance of classifiers. The limitation of this work is that we can use only those kinds of drug molecules on which our model has been trained for predicting its activity, activity score, potency, and efficacy. Different types of drug molecules have their different physicochemical properties (features). Therefore, different kinds of drug molecules' activity cannot be recognized by this model.

Chapter 4

Toxicity Prediction of Small Drug Molecules of Aryl Hydrocarbon Receptor using Proposed Ensemble Model

Quantitative structure-activity relationships and quantitative structure-property relationships have proved their usefulness for predicting toxicities of drug molecules regarding their biological activities. The in silico toxicity prediction techniques are essential for reducing rodents testing (in vivo), less time-consuming and 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 their 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. The pharmaceutical data exploration laboratory software is used for extracting the features of drug molecules. The dataset of 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 (non-binding) compounds of the dataset. In our proposed ensemble model, we primarily performed feature selection using Boruta library in R, after which, we resolved the class imbalance problem itself by ensemble learning where we divide the dataset into seven data frames, which have the approximately equal number 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%. The k-fold cross-validation is conducted to measure the consistency of the model. Finally, the validity of the proposed ensemble model on some drug molecules of acquired immune deficiency syndrome therapy and androgen receptor has been proved.

4.1 Introduction

Most of the drugs are small molecules that are invented to interact, 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 the cell-surface receptor (hormone receptors, neurotransmitter receptors, cell-signaling receptors, etc.), enzymes, and other functional proteins. Owing to the physiologic 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 be minor physical symptoms, such as running nose due to allergies, or life-threatening and debilitating events like depression or sepsis [12].

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 the target site [51]. The researchers must classify the active (binding) compounds from the inactive (non-binding) compounds. This strong desire can lead to the design of the 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 [76].

The data challenge of Tox21 with the collaboration of National Center for Biotechnology Information (NCBI) is held to help the researchers for understanding the chemical and compound toxicology that can disrupt biological pathways in a manner that may result in toxic effects. It was the open challenge where the researchers had to predict about compounds intervention in biochemical pathways by using only physicochemical structure data. Active drug molecules are those molecules, which can bind to one or more biochemical pathway assays and create some toxic effects into our body. These toxic effects are stress response effects (SR) and nuclear receptor effects (NR). Both the 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 [11]. 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 aryl hydrocarbon receptor. Table 4.1 shows the 12 biological pathway assays that can give distinct adverse health effects on its activation.

Table 4.1: 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, which is utilized for defining discovery and safety efforts in therapeutics [78]. 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 (1959) about the growing popularity of the 3Rs (Replacement, Reduction, Refinement) focuses on the limited use of animal and the unlimited use of computational techniques for toxicity testing [40]. The in silico models also can predict ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties in the chemical space that can reduce the dependency of chemical laboratory synthesis (in vitro) [50].

The aryl hydrocarbon receptor (AhR) is a protein, and it is a member of the family of basic helix-loop-helix transcription factors. The 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 the toxic responses with various side effects, whereas the elicitation of metabolizing enzymes results in the production of toxic metabolites. The 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 environmental persistent organic pollutants (POPs), which can cause developmental problems and immunological disorder by interfering with hormones. DLCs can also develop the disorder in the nervous system, endocrine system, reproductive functions, and even cause of cancer. The exposure of high-level dioxin to humans may result in skin lesions, such as patchy and chloracne darkening of the skin, impairment of the immune system, and modified liver function [94].

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 Boruta library in R, then the class imbalance problem is resolved

through ensemble learning method where we divide the dataset into seven data frames which have the approximately equal number 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 upon the votes of seven random forest models is created which is our proposed ensemble model that has also resolved the issue of class imbalance. The k-fold cross-validation is performed to measure the robustness of the proposed ensemble model. Finally, we have proved the validity of this model on 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 on 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:

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 chances to disrupt the processes in the human body.

3. To develop a stand-alone application for helping the researchers to predict the toxicity of the 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 the glass (in vitro).

Figure 4.1 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 and 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 [45].

The paper is composed as follows: Section 4.2 introduces a quick overview of the dataset, feature extraction using PaDEL, feature selection, and class imbalance problem. Section 4.3 clarifies the procedure of proposed ensemble model. Section 4.4 presents the description of the random forest model which is used as a base classifier for ensemble learning. Section 4.5 presents the different performance evaluation parameters of model for classification. Section 4.6 describes the investigated, compared, and validated results which are followed by the conclusion in Section 4.7.

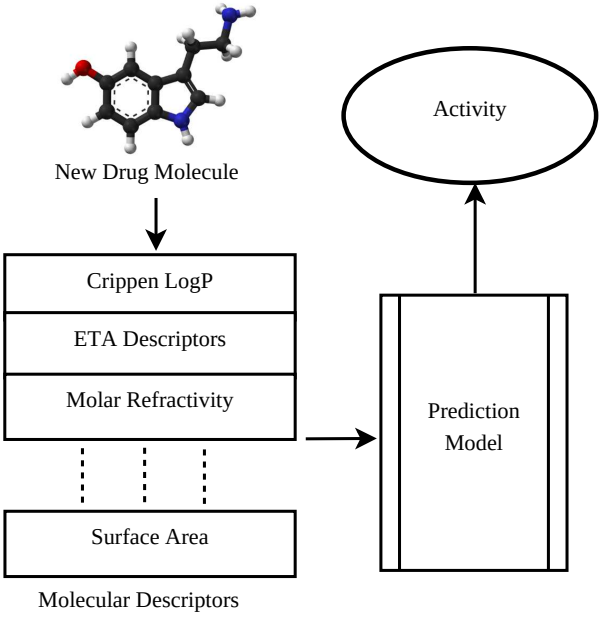


Figure 4.1: Activity prediction method

4.2 Materials and Methods

4.2.1 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 (SDF) of AhR drug molecules, and its output is a comma-separated values (CSV) file. The CSV file contains total 9008 drug molecules, and each drug molecule comprises 1444 features. The PaDEL-Descriptor is a java based free and open source software, which is similar to Dragon, MOE, MARVIN Beans, and supports more than 90 different molecular file formats like PDB, SDF, SMILE, etc. The molecular descriptors are extracted using the Chemistry Development Kit (CDK) library of java, which is related to the chemo-informatics and bio-informatics that are used internally in PaDEL. The PaDEL software can calculate 1876 molecular descriptors (1444 1D, 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 4.2 shows the graphical user interface (GUI) of PaDEL-Descriptor [2], and Figure 4.3 shows the format of the structure-data file for an active drug molecule of AhR.

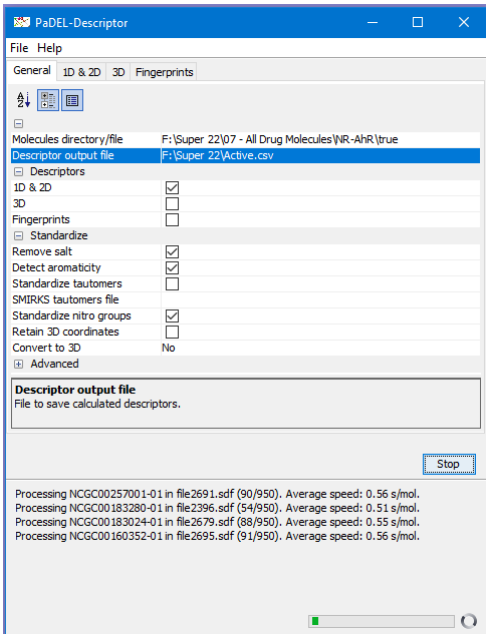


Figure 4.2: PaDEL-Descriptor GUI which calculating the molecular descriptors of AhR

4.2.2 Dataset

The AhR signaling pathway data is taken from PubChem, which is maintained by the National Center for Biotechnology Information that provides access to biomedical information from its website (https://pubchem.ncbi.nlm.nih.gov/bioassay/743122), where 743122 is the PubChem identification number for AhR. In this study, our dataset consists of total 9008 AhR's drug molecules, of which 1063 are active molecules, and the remaining 7945 are inactive molecules. All the drug molecules contain 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), etc. Table 4.2 lists some essential physicochemical properties of the drug molecules of AhR and their descriptions.

Table 4.3 shows the glance of the dataset that contains various AhR drug molecules, such as NCGC00257625-01, NCGC00259354-01, and NCGC00255335-01. The columns of the Table 4.3 shows the various molecular descriptors/features, such as ATS0m, AATSC8m, 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.

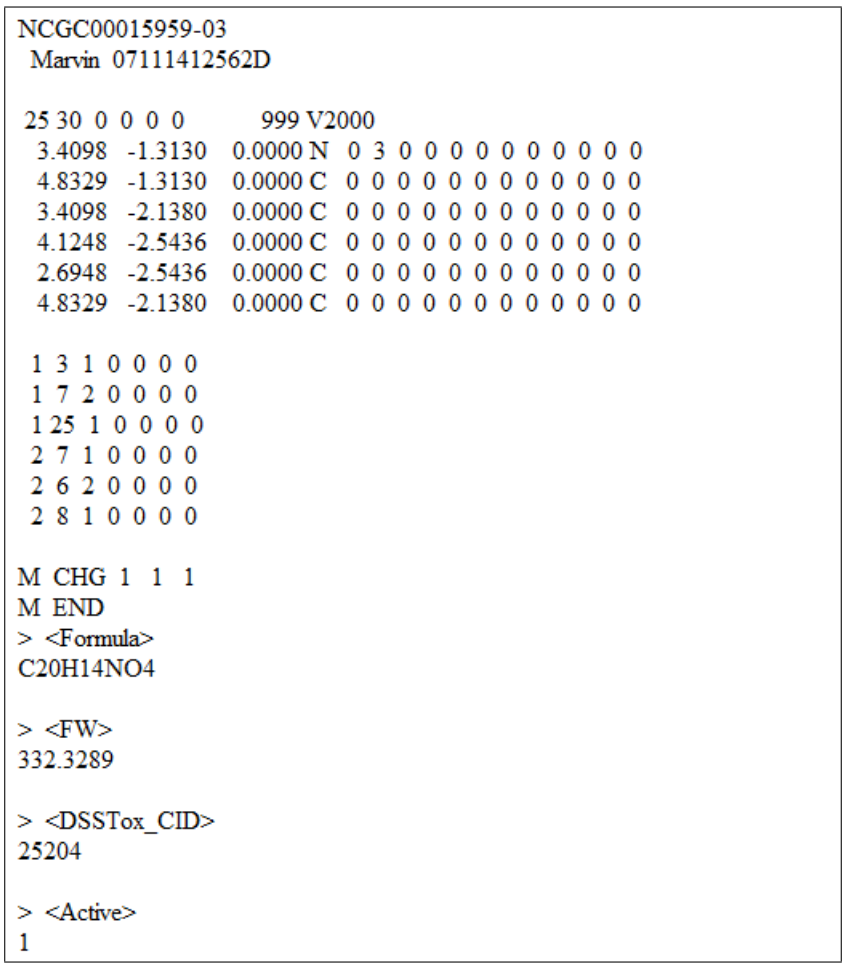


Figure 4.3: Structure-data file format for single drug molecule of AhR

4.2.3 Feature selection using Boruta algorithm

During the process of model building, the feature selection is used to filter the high 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 select 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 Boruta() function under 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 [95].

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

Table 4.2: Physicochemical properties of aryl hydrocarbon receptor's drug molecules

SN	**Name**	**Description**
1	Crippen logP	Atom based calculation of logP using Crippens method, it is also called the octanol/water partition coefficient.
2	Eccentric connectivity index	It is a distance based Atomic descriptor that is 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 is 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 relationship. It captures information about different features of molecule 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 the 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 4.3: Sample dataset of aryl hydrocarbon receptors

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

the confirmed attributes have been used for the model building. Table 4.4 shows all the important features of AhR dataset.

4.2.4 Class imbalance

Class imbalance is a problem in machine learning where the main class of interest is rare, which triggers the biasness of the classifier. Here, our dataset for the prediction of activity contains two classes; one is active, and the other is inactive. This dataset is extremely imbalanced, as the total number of active drug molecules are 1063 (minority class) and the total number of inactive drug molecules are 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 synthetic minority oversampling technique (SMOTE) [79, 96]. Here, the class imbalance problem is resolved by the ensemble learning method, as the ensemble learning

Table 4.4: Important features of aryl hydrocarbon receptor

SN	Features	SN	Features	SN	Features	SN	Features	SN	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

is more effective than data sampling techniques to enhance the classification performance of imbalanced data. It is performed by the creation of the seven data frames by dividing the dataset. These all data frames have the approximately equal number of active and inactive drug molecules [71] (refer to Phase 3 of Section 4.3 for more details).

4.2.5 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 non-toxic 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 at their early stage of drug development (pre-clinical trial) to save the lives of animals as well as money and time. Figure 4.4 shows the flowchart where our proposed ensemble-based classification model performs the categorization of a new drug molecule in active and inactive categories.

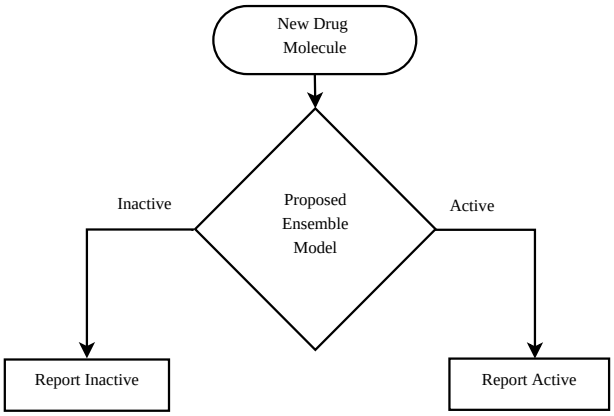


Figure 4.4: Flow chart for classification of AhR's drug molecules

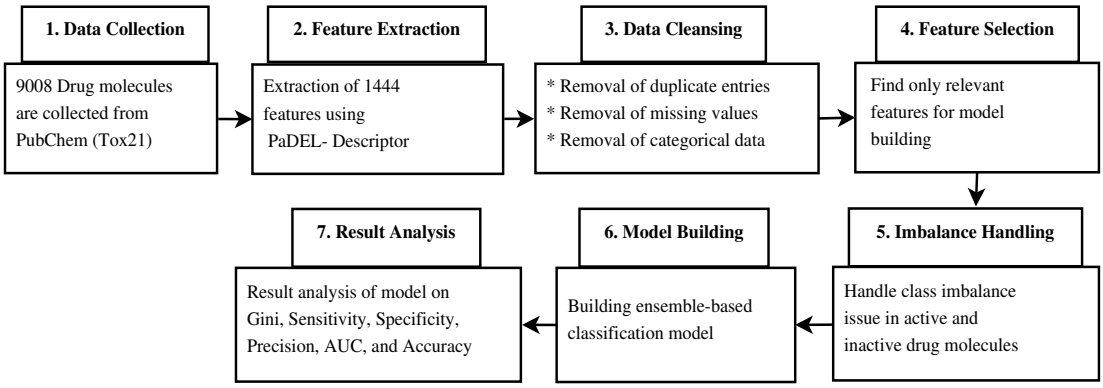


Figure 4.5: Methodology used for the proposed ensemble model

4.3 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 same training dataset or different base classifiers on the same training dataset. Here, we used first technique, where we created different splits of same training dataset, and the random forest model is applied as a base classifier on all these split datasets. This approach is to improve our classification accuracy as well as to solve the issue of class imbalance problem [71, 54]. Here, the random forest model is taken as a base classifier because the performance of this model is better than other models. Figure 4.5 shows the methodology of the proposed ensemble-based prediction model, and Figure 4.6 shows only the approach of ensemble learning which is applied in the proposed ensemble model. The following five phases are showing the methodology of our proposed ensemble model.

Phase 1: Dataset generation

The unprocessed dataset of AhR is obtained from PubChem website, which is in the structure-data files (.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 "PaDEL-Descriptor software" individually, which has generated two comma-separated values (.CSV extension) files; one for active drug molecules and the other for inactive drug molecules. Now, these two datasets are combined to form a resultant dataset, whose target variable is a binary class activity (refer to Subsection 4.2.2 for more details).

Phase 2: Data cleansing and feature selection

Data preprocessing is a technique of improving the quality of data because the high-quality input data will provide highly accurate and consistent knowledge [97]. 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 few corrupted and missing entries in our dataset. First, we have corrected these corrupted entries and then analyzed those attributes that have missing values in their cells. We have filled these missing values by the average value of that particular column. Since our dataset has 1444 features that are very high in dimensionality, therefore, we applied feature dimensionality reduction method which reduces the features as well as the execution time of the classifiers in machine learning [98]. Here, Boruta algorithm is applied to the dataset which returns only 150 features, out of 1444 features (refer to Subsection 4.2.3 for more details).

Phase 3: Class imbalance handling

The dataset found from PubChem is highly imbalanced, as it has total 9008 drug molecules of which 1063 are active, and 7945 are inactive. To resolve this problem, we primarily segregated the active and inactive drug molecules of the dataset, where we found 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 the seven data frames so that all the data frames have an approximately equal number of active and inactive drug molecules. Now, these seven data frames are different and balanced datasets which are available for model building by using ensemble learning.

Phase 4: Classification model building using ensemble learning

Now, we have seven balanced and different datasets. We have trained each dataset at

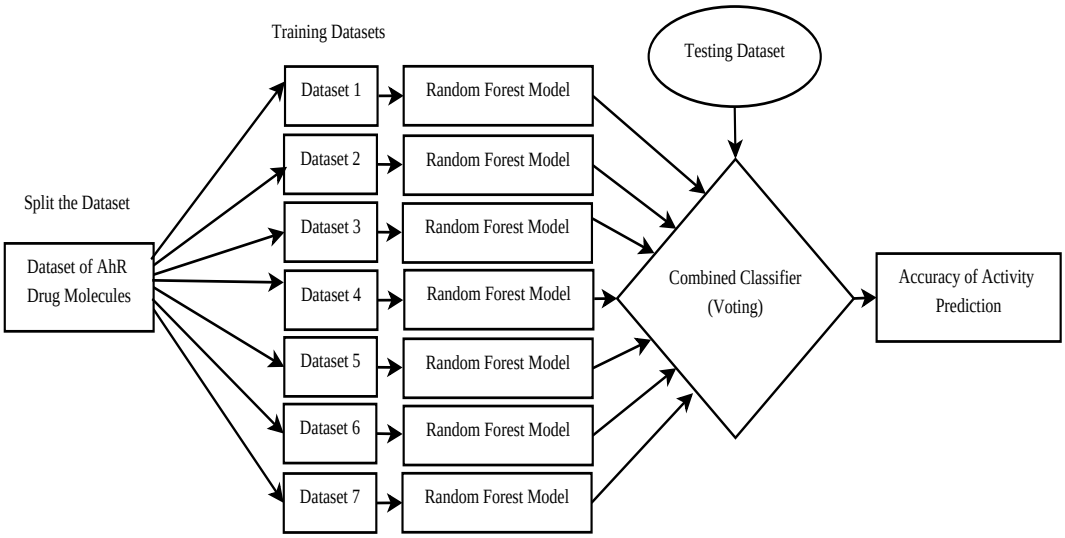


Figure 4.6: Ensemble method for activity prediction

70% data using the base classifier random forest and combined all these classifiers using the ensemble learning approach.

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 frames. The performance of the proposed ensemble model is evaluated on this testing dataset (see Figure 4.6). Now, this model can be used to predict the activity of any new drug molecule of AhR.

4.4 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 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 4.5 is showing various models with their corresponding packages, methods, and tuned parameters. Random forest with its tuned parameters has been used in the proposed ensemble model, and other models with its tuned parameters have been used for comparison with the proposed ensemble model. All the models are implemented in R, under GNU general public license. The randomForest method has “mtry” and “ntree” parameters for tuning, where mtry

Table 4.5: 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"

is showing 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 as large as to ensure that every input row gets predicted at least a few times. Therefore, we optimized the performance of random forest model by setting the value of mtry=2, and ntree=500. We used randomForest method as randomForest(formula, trainDataset, ntree=500, mtry=2); where its formula comprised 150 important features and target class as shown below:

$$Activity \sim f(AMR + nN + nAromBond + ... + WTPT.5 + WPATH) \quad (4.1)$$

Table 4.4 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 most popular class is returned during classification [30, 26]. If n is the number of records and d is the depth of tree then the time complexity of 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 [79].

4.5 Binary Classification based Performance Evaluation Parameters

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

4.5.1 Gini coefficient

Gini coefficient is used to measure the distribution inequality of data [32]. Gini values ranges between 0 and 1. The 0 and 1 values of Gini coefficient indicates perfect equality

of data and perfect inequality of data, respectively. Assuming that a model M has a Gini coefficient 0.6 and model D has a Gini coefficient 0.45, then model M is considered as a productive model in contrast to model D.

4.5.2 Sensitivity

Sensitivity (Sens) is also known as True Positive Rate (Recognition) or Recall [30]. It is the ratio of actual positives that are correctly identified as positives by the classifier. It is computed as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.2}$$

4.5.3 Specificity

Specificity (Spec) is also known as True Negative Rate [30]. It is the ratio of actual negatives that are correctly identified as negatives by the classifier. It is computed as:

$$Specificity = \frac{TN}{TN + FP} \tag{4.3}$$

4.5.4 Precision

Precision can be thought of as a measure of exactness, it means what percentage of tuples labeled as positive are actually such [30]. It is computed as:

$$Precision = \frac{TP}{TP + FP} \tag{4.4}$$

4.5.5 AUC

The area under the curve (AUC) measures the quality of the classifier. The Receiver Operating Characteristics (ROC) is a curve, which is drawn between True Positive Rate (TPR) and False Positive Rate (FPR). We can find these parameters by the confusion matrix. The area under the ROC is called AUC. The AUC value ranges between 0 and 1. The quality of a model is outstanding if it has AUC esteem close to 1. The model scoring high AUC as compared to another model is considered as an efficient model [32].

Table 4.6: 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

4.5.6 Accuracy

Accuracy is the most important criteria for measuring the exactness of any classifier [26]. Accuracy can be computed as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100 \tag{4.5}$$

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

4.6 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 4.5. 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 4.6. 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 [96] are the existing models which are used for comparison.

4.6.1 K-fold cross-validation

The k-fold cross-validation approach partitions the dataset into k equal-sized subsets or "folds". During each execution, one of the partition is chosen for testing, while the rest

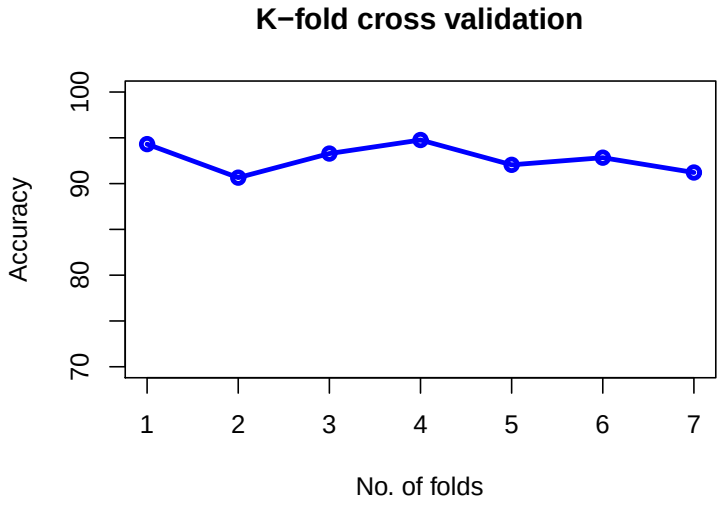


Figure 4.7: K-fold cross validation for activity prediction

Table 4.7: 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

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, the random data is provided for training and testing to measure the robustness of the model [82]. Here, we have used 7-fold cross-validation method for activity prediction, where the result of cross-validation (Figure 4.7) shows the consistent performance on all the evaluation parameters of the proposed ensemble model [45]. The value of k has selected in such a manner that each training and testing partitions 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 an equal number of drug molecules. At the value of k=7, we can take collectively any six data frames for the training of model, and remaining one data frame for the testing of the model. Table 4.7 describes the accuracy of the proposed ensemble model by applying 7-fold cross validation one time.

4.6.2 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 on 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's drug molecules and some androgen receptor's 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 the AIDS therapy [84]. These drugs show potent anti-HIV-1 activity and its 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, 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 of the body. Rilpivirine also has some side effects, which are sores in your mouth, redness or swelling of your eyes, face, lips, mouth, tongue or throat. Now, we have two-dimensional structures of all the nine drug molecules, which are downloaded as a structure-data file from the PubChem database and their molecular descriptors have been extracted with the help of PaDEL-Descriptor. Now, we applied 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 6.8. The results of the table shows that the drug molecules which are 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 4.8 shows the validation process of the proposed ensemble model on the four drug molecules of AIDS therapy.

4.7 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

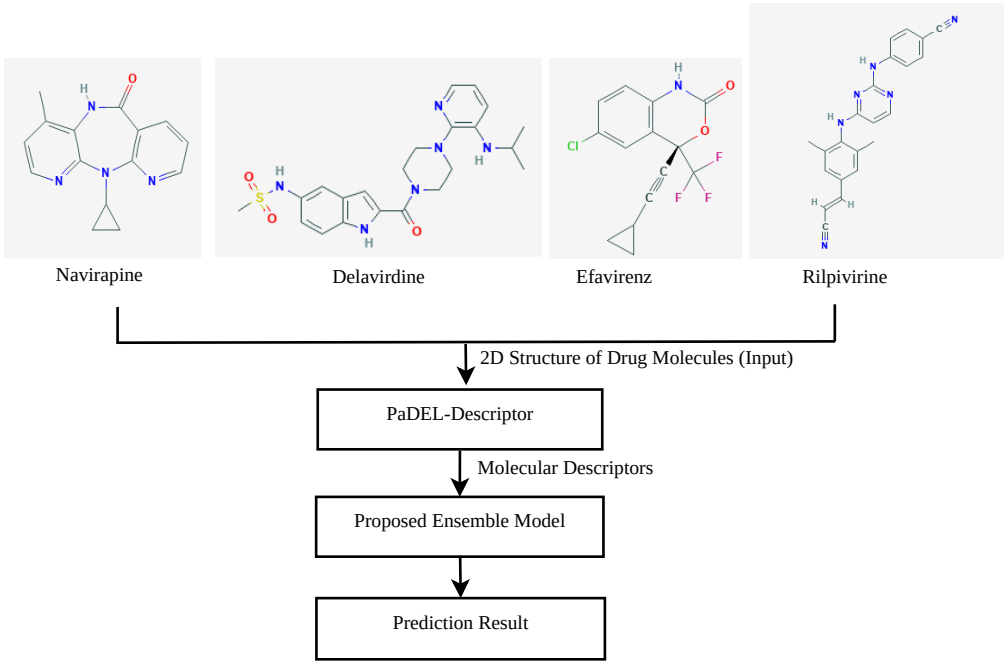


Figure 4.8: Activity prediction of AIDS therapy drug molecules using proposed ensemble model

Table 4.8: 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%

toxicity prediction by uploading the structure-data file (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 Boruta method and balanced the dataset by using an ensemble learning approach. Here, the ensemble method is used for dual purpose one, it resolves the problem of class imbalance, and second, it is used for classification. The proposed ensemble model has been evaluated on various performance parameters, i.e. 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 highly imbalance 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's 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 their different physicochemical properties or features.

Chapter 5

Toxicity Prediction of Small Drug Molecules of Androgen Receptor using Multilevel Ensemble Model

In this study, efforts are created to develop a quantitative structure-activity relationship (QSAR) based model, which are used for the prediction of toxicities to reduce testing in animals, time, and money in the early stage of drug development. An efficient machine learning model is developed to predict the toxicity of those drug molecules which binds to the androgen receptor (AR). Toxicity prediction is performed in term of their activity, activity score, potency, and efficacy by using various physicochemical properties. A multilevel ensemble model is proposed, where its first level is performed ensemble-based classification of activity, and the second level is performed ensemble-based regression of activity score, potency, and efficacy of only those drug molecules which have been found active during the classification level. The androgen receptor dataset has 10273 drug molecules where 461 are active, and 9812 are inactive, and each drug molecule has 1444 features. Therefore, our dataset is highly imbalanced and having a very large number of features. Initially, we performed feature selection then the class imbalance problem is resolved. The k-fold cross-validation is accomplished to measure the consistency of the model. Finally, our proposed multilevel ensemble model has been validated and compared with some existing models.

5.1 Introduction

Most people in their lifetime ingest various chemical, sources including medicine, household cleaning products, and food. In some cases, these chemicals may be harmful/toxic for our body. Generally, more than 30% of anticipating drugs are failing during human clinical trial (Phase II) because of their harmful or toxic effects, while these drugs already have been passed in their pre-clinical trial (Phase I) [33]. Currently, toxicity and lack of efficacy are the significant causes of drug candidate failure during development and pre-clinical or clinical trials. According to an analysis, 36% drugs are failed due to lack

of efficacy, and 43% drugs are failed due to toxicity or some adverse effects; subsequently, only <10% of the medicines survive in the market [10]. The complications of toxicity are tremendous not only regarding costs but also in the actual lives of patients. Because of these problems, there is a need to create a computational method (in silico) as an alternative of bioassays to save lives, money and time to develop new medicines. Generally, computational models are the expert systems which are based on QSAR models to predict the biological activity of various kinds of drug molecules [59].

Expert systems (rules-based system generated by human experts) have been used to predict various kinds of toxicity that include: Ames mutagenicity, rodent carcinogenicity, skin and eye irritation, running nose due to allergies, skin sensitization, acute inhalation toxicity i.e., lethal concentration (LC50), acute oral toxicity i.e., lethal dose (LD50), acute toxicity i.e., effective concentration (EC50), maximum tolerated dose (MTD), and chronic lowest observable adverse effect level (LOAEL). The most common example of toxicity is Human Ether-a-go-go-related Gene (hERG) receptor modeling where drugs will bind to the hERG receptor and cause arrhythmia and heart failure. Therefore, it is important to screen out those compounds that can create an adverse effect during drug design, as early as possible [59].

An androgen is a natural or synthetic steroid hormone that regulates male sexual development and separation; it incorporates the formation and maintenance of the function of the reproductive system. Androgens are essential in muscle advancement and related to mental or emotional attitudes about sexuality. Androgens intercede their belongings through the androgen receptor (AR) that is a nuclear hormone receptor and the member of the nuclear receptor superfamily. Transformation in the AR results in diseases such as androgen insensitivity syndrome, prostate malignancy, Kennedy's sickness, and infertility [99]. Endocrine disrupting chemicals (EDCs) and their interactions with steroid hormone receptors like AR may disrupt normal endocrine function as well as interfere with metabolic homeostasis, reproduction, developmental and behavioral functions. MDA-kb2 AR-luc cell line is used to identify the compounds that activate AR signaling [33] [88]. Disruptors of androgen are the drug molecules that interfere with the biosynthesis, metabolism or action of endogenous androgens resulting in a deflection of regular male developmental programming, toxic effects, side effects, and reproductive pathway growth [100].

The toxicology in the 21st century (Tox21) program intends to grow better toxicity appraisal techniques to rapidly and productively test whether certain chemical compounds have the potential to disrupt forms in the human body that may prompt unfavorable wellbeing impacts [33]. Tox21 was the open challenge where the researchers had to predict activity of compounds in biochemical pathways using physicochemical properties

and machine learning model. Here, data generated from nuclear receptor signaling and stress response pathway assays run against Tox21's 10k compound library to build models [101].

Apart from Tox21 other models like eToxPred and ToxiM are also available, which have different approaches to toxicity prediction. eToxPred estimates the toxicity and synthetic accessibility of small organic drug molecules, where the estimated toxicity is reported as the Tox-score, and the synthetic accessibility is evaluated with the SAscore. eToxPred used machine learning algorithms, which trained on molecular fingerprints to evaluate drug candidates. The performance of eToxPred is assessed against multiple datasets containing natural products, known drugs, synthetic bio-active compounds, and potentially hazardous chemicals [55].

ToxiM is a classification and regression model for the prediction of molecular toxicity along with the aqueous solubility and permeability of any drug molecule. Here, authors have integrated the machine-leaning and chemo-informatics approaches, where fingerprints and descriptors are input features. ToxiM used the dataset of Toxin and Toxin Target Database (T3DB) for the prediction of molecular toxicity [57].

The creation of activity prediction models using the dataset of Tox21 data challenge with the collaboration of National Center for Biotechnology Information (NCBI) was held to help the researchers to understand the chemical and compound toxicology that can disrupt biological pathways in a manner that may result in toxic effects. It consists of activity data for two panels which play essential roles in toxicological pathways [48]. Nuclear receptor signaling panel (NR) includes activity data for seven targets, which are Androgen receptor(AR), Androgen receptor ligand binding domain (AR-LBD), Estrogen receptor (ER), Estrogen receptor ligand binding domain (ER-LBD), Aryl hydrocarbon receptor (AhR), Aromatase, and Peroxisome proliferator-activated receptor gamma (PPAR-gamma). Stress Response Panel (SR) includes activity data for five targets: Antioxidant responsive element(ARE), Genotoxicity indicated by ATAD5, Heat shock factor response element (HSE), the disruption of Mitochondrial membrane potential (MMP), and p53 [47].

These are the 12 biological pathway assays which can give distinct adverse health effects on its activation. These toxic effects are stress response (SR) effects and nuclear receptor (NR) effects. Both SR and NR effects are highly critical to human health [11]. Therefore, We need to build a computational model that can predict the toxicity of the drug molecules in these 12 pathway assays based on their physicochemical properties. In this paper, we are analyzing the toxic effects only on androgen receptor (AR). This work could finally lead to wider use of large-scale chemical screening techniques to predict a

chemical compounds toxicity in humans.

In this research work, efforts are created for the development of quantitative structure-activity relationships (QSARs) with quantitative structure-property relationships (QSPRs), which are used for the prediction of varied toxicities to reduce testing in animals. We have proposed a novel machine learning model for the prediction of activity, activity score, potency, and efficacy of drug molecules of AR. Activity is a binary class which decides the toxicity of drug molecules, and activity score, potency, and efficacy are the continuous classes which determine the quality of active drug molecules only. For the prediction of all these classes efficiently, we have created an ensemble based multilevel prediction model where its first level is performing ensemble based classification for predicting activity, and its second level is performing ensemble based regression for predicting the activity score, potency, and efficacy of only those drug molecules which were found active during classification. Here, we are using the dataset of AR which is found from PubChem (Open Chemistry Database). This dataset has a total of 10271 drug molecules where 461 are active, and 9812 are inactive, and each drug molecule has 1444 features [88]. Therefore, our dataset is very high in features and very imbalanced in classes.

Here, feature selection is performed using Boruta method and information gain method of R. Subsequently, using an ensemble-based classification, we have resolved the issues of class imbalance, and classified active or inactive drug molecules of AR. In the first phase, the random forest model is used as a base classifier for ensemble-based classification. In the second phase, the prediction of activity score, potency, and efficacy are performed using the ensemble of four decision methods which are linear model, decision tree, random forest, and neural network. The objective of our proposed multilevel ensemble model is to build an efficient binary classification model for activity forecast whether a given specific compound is active (1) or inactive (0). Simultaneously, our objective is to build up an efficient regression model for the prediction of activity score, potency, and efficacy.

Figure 5.1 shows the general diagram of the proposed multilevel ensemble model, where we are taking the various structure data files (SDF) of some small drug molecules of AR. SDF is a chemical file formats to represent multiple chemical structure records and associated data fields. SDF was developed and published by Molecular Design Limited (MDL) and became the most widely used standard for importing and exporting information on chemicals [102]. SDF files are converted into molecular descriptor using PaDEL-Descriptor, and on the basis of these molecular descriptors, our proposed model is forecasting activity, activity score, potency, and efficacy of any small drug molecules of AR. We can consider this model as a decision support system where a designer of drug molecules can submit any new drug molecule of AR in CSV file format, and our proposed

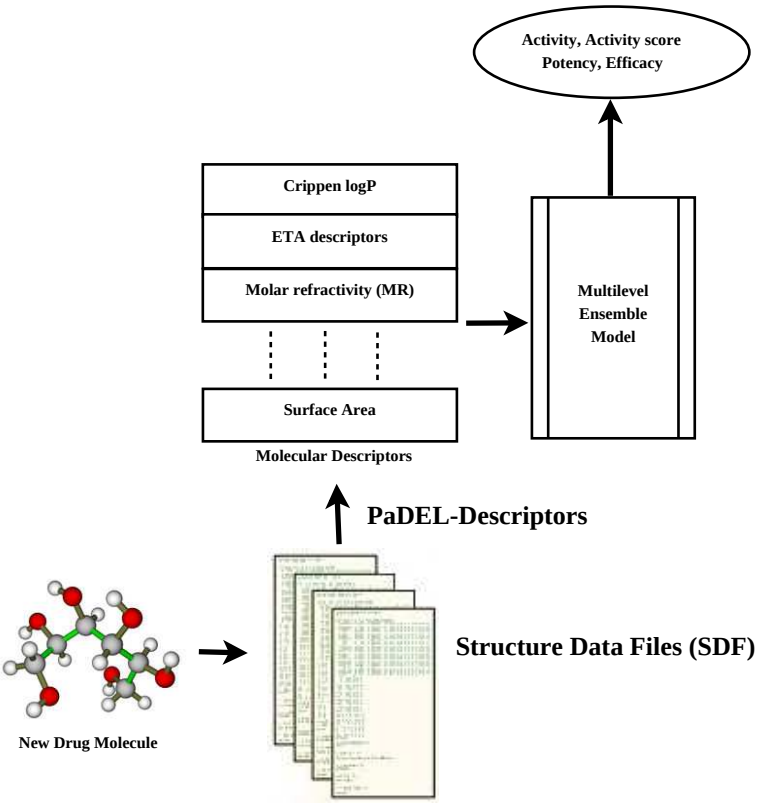


Figure 5.1: Activity, activity score, potency, and efficacy prediction method for AR's drug molecules

model will predict its target classes. If the designers of drug molecules are satisfied with the results, then they can proceed for the next level of drug development; otherwise, they can predict again or go for some other method. Figure 5.2 shows the proposed multilevel ensemble model as a decision support system.

The significant contribution of this paper is:

1. To develop better toxicity assessment features, methods, and algorithm for knowing the disruptors of AR drug molecules.

2. To develop a machine learning based model for quick and efficiently testing of

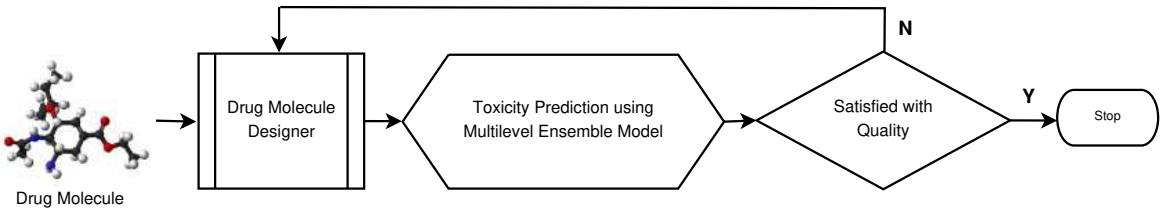


Figure 5.2: Proposed multilevel ensemble model as a decision support system

certain chemical compounds that have probable chances to disrupt processes in the human body.

3. To develop a stand-alone application that helps the researcher to predict the toxicity of the newly discovered chemical compounds and environmental chemicals using computational method (in silico), rather than inside the living organism (in vivo) or within the glass (in vitro).

The paper is composed as follows: A quick overview of the dataset, feature extraction using PaDEL, feature selection, class imbalance problem, and target classes description are introduced in Section 5.2. The procedure of the proposed multilevel ensemble model is clarified in Section 5.3. The description of various machine learning models used in this work is presented in Section 5.4. Various model evaluation parameters of classification and regression are presented in Section 5.5. Section 5.6 describes the result investigation, comparison, and validation of the proposed multilevel ensemble model. Section 5.7 shows discussion according to the performance improvement of results, which is followed by the conclusion in Section 5.8.

5.2 Materials and Methods

5.2.1 Feature extraction using pharmaceutical data exploration laboratory (PaDEL)

PaDEL-Descriptor is a java based free and open source software, which is used to calculate molecular descriptors. The descriptors are calculated using The Chemistry Development Kit (CDK), and it supports mainly MOL, SDF, SMILES file format [2]. Molecular descriptors are the encoding vectors of physicochemical properties of any drug molecules which play a fundamental role in chemistry, pharmaceutical science, toxicology, environmental protection policy, and biological activities. According to Todeschini and Consonni, the molecular descriptor is the final result of the logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number [7]. Table 5.1 describes some essential physicochemical properties and biological activities of drug molecules. Physicochemical properties play a significant role in forming of the molecular descriptors, and with the help of these molecular descriptors, it finds biological activities. It is found that the octanol/water partition coefficient, molar refractivity, dipole moment, and density are quite abundant in chemical information to encode the structural features that contribute to the various toxicities,

Table 5.1: Physicochemical properties and biological activities of drug molecules

Physicochemical Properties	Octanol/water partition-coefficient Molar refractivity Skin sensitization Boiling point Melting point Dipole moment Vapor pressure Solubility Density
Biological Activities	Anti-inflammatory activity Antidepressant activity Inhibition concentration Binding affinity Carcinogenicity Lethal dose Conductivity Retention time Mutagenicity

Table 5.2: Features of drug molecules of androgen receptor

Name	ATSC4m	ATSC1v	C3SP3	MDEC-24	nFRing	AMR	ALogp	Activity
NCGC00015479-09	321.148	195.320	0	0	1	59.7133	-0.2676	0
NCGC00015693-02	943.516	275.089	0	0	3	47.5793	1.3349	0
NCGC00016260-05	-32.789	12.412	0	0	0	15.4305	-1.0852	0
NCGC00013034-01	-634.742	281.866	4	3.748	6	82.8538	-0.1812	1
NCGC00013235-01	-664.668	276.814	3	3.044	6	62.8954	-0.4057	1
NCGC00013489-01	-388.188	123.693	4	3.231	1	15.3637	-1.6864	1

and these toxicities come under biological activities. PaDEL software takes input of SDF files of AR and produces a single CSV file which contains all the drug molecules and their 1444 features, these features are called molecular descriptors, and are used for model building.

5.2.2 Dataset and its features

The data is taken from PubChem "https://pubchem.ncbi.nlm.nih.gov/bioassay/743040", this dataset is of AR where 743040 is the PubChem identification number of androgen receptor signaling pathway. The dataset consists of total 10271 molecules out of which 461 are active molecules, and remaining 9812 are inactive molecules [88]. Table 5.2 is the subset of the dataset, where NCGC00015479-09, NCGC00015693-02, NCGC00013235-01, and so on are the names of various drug molecules and ATSC4m, AMR, ALogp, and so on are the different molecular descriptors of these drug molecules. Table 5.3 shows only the target classes for classification and regression levels.

Table 5.3: Activity, activity score, potency and efficacy classes of drug molecules of androgen receptor

Name	Activity	Activity score	Potency	Efficacy
NCGC00015479-09	0	0	-	-
NCGC00015693-02	0	0	-	-
NCGC00016260-05	0	0	-	-
NCGC00013034-01	1	44	4.216	124.545
NCGC00013235-01	1	42	4.216	52.547
NCGC00013489-01	1	42	33.492	155.466

Table 5.4: Important features of androgen receptor

SN	Features	SN	Features	SN	Features	SN	Features	SN	Features
1	nN	13	MATS8s	25	C4SP3	37	VP.7	49	MDEC.44
2	ATS5s	14	GATS3c	26	SCH.6	38	maxHBa	50	nFRing
3	ATS6s	15	GATS2e	27	VCH.6	39	ETA_BetaP	51	nFG12Ring
4	AATS1i	16	GATS1i	28	VCH.7	40	ETA_BetaP_ns	52	nTRing
5	AATS2i	17	VE2_Dzp	29	SC.6	41	ETA_EtaP	53	nTG12Ring
6	ATSC4m	18	SpMAD_Dzs	30	VC.5	42	ETA_EtaP_L	54	RotBFrac
7	ATSC4v	19	nBase	31	VC.6	43	ETA_EtaP_F_L	55	JGI2
8	ATSC1i	20	bpol	32	VPC.4	44	HybRatio	56	WTPT.5
9	AATSC4m	21	SpMin1_Bhm	33	VPC.5	45	IC1		
10	AATSC3s	22	SpMin1_Bhe	34	VPC.6	46	IC2		
11	MATS4m	23	SpMin1_Bhs	35	ASP.7	47	TIC1		
12	MATS1i	24	C3SP3	36	VP.6	48	MDEC.24		

5.2.3 Feature selection

During the process of the model building, feature selection is performed for an automatic collection of attributes which are most relevant to the predictive class. The process of feature selection is carried out using Boruta library, and FSelector library of R. Feature selection for the classification model is calculated using Boruta method [95], whose input parameters are dataset and target class is activity. The feature selection for the regression models is computed using information.gain method [103], whose input parameters are a dataset of only active drug molecules and target classes are activity score or potency or efficacy.

Boruta is a wrapper algorithm that finds relevant features by the values of meanImp, medianImp, minImp, maxImp, and normHits. Finally, Boruta() provides 56 essential features which are used to build a binary classification model for activity prediction in proposed multilevel ensemble model. Table 5.4 shows all the essential features of AR dataset.

The information.gain() is the entropy-based function which has two input parameters, that are the dataset and its target class. The target class may be anyone from activity score, potency, and efficacy. Table 5.5, Table 5.6, and Table 5.7 show the various features of dataset and their importance scores/weights for activity score, potency, and efficacy, respectively. During regression model building, only those combinations of features are se-

Table 5.5: Feature importance for activity score

SN	Features	Importance score
1	MDEC.34	0.4905828909
2	ATS4i	0.4660665782
3	GGI2	0.4620568079
.	.	.
.	.	.
.	.	.
447	nBondsT	0.0438125287
448	GATS3s	0.0436822853
449	ATSC7p	0.0418491578

Table 5.6: Feature importance for potency

SN	Features	Importance score
1	ATS4i	0.4289657857
2	ATS6i	0.4254316887
3	ATS5i	0.4197247444
.	.	.
.	.	.
.	.	.
437	nF12HeteroRing	0.0427259625
438	nT12HeteroRing	0.0427259625
439	MIC2	0.0404659457

lected which are giving the best accuracy in comparison to other combination of features. Then, finally, the first 84, 28, 149 features are selected for building ensemble based regression model for the prediction of activity score, potency, and efficacy, respectively. The cutoff.k method of FSelector package selects k best attributes from the ranked attributes [103]. Table 5.8, Table 5.9, and Table 5.10 show the performance evaluation parameters of proposed multilevel ensemble model in term of RMSE, correlation (r), coefficient of determination (R^2), and accuracy based on different combinations of features.

5.2.4 Class imbalance resolution

Class Imbalance is the issue in data related prediction where one class of data (positive) is far less than another class of data (negative). Our dataset for the prediction of activity during the classification phase is highly imbalanced because the number of active drug molecules is 461 and the number of inactive drug molecules is 9812. Therefore, the

Table 5.7: Feature importance for efficacy

SN	Features	Importance score
1	ATS4v	0.3422387541
2	SRW10	0.3408394871
3	MWC6	0.3400985047
.	.	.
.	.	.
.	.	.
343	MATS4m	0.0502219384
344	GATS4s	0.0473757888
345	AATS3i	0.0450785986

Table 5.8: Evaluation parameters for activity score based on number of features

No. of features	r	R^2	RMSE	Accuracy
84	0.92	0.85	3.48	92.11
281	0.94	0.88	3.35	90.35
429	0.93	0.86	3.63	90.35
194	0.95	0.9	3.51	89.47
262	0.92	0.85	3.63	89.47
315	0.94	0.88	3.52	89.47
359	0.94	0.88	3.87	89.47

Table 5.9: Evaluation parameters for potency based on number of features

No. of features	r	R^2	RMSE	Accuracy
28	0.77	0.59	5.27	92.66
55	0.75	0.56	5.73	92.66
115	0.71	0.5	6.26	90.83
234	0.69	0.48	5.37	90.83
247	0.64	0.41	6.23	90.83
354	0.78	0.61	5.78	90.83
393	0.7	0.49	6.54	90.83

number of active drug molecules is far less than the number of inactive drug molecules. The primary function of class balancing is to balance the class symmetry of instances. Class imbalance issue can be resolved by various approaches which are undersampling, oversampling, synthetic minority oversampling technique (SMOTE) [71], and ensemble systems. Here, the class imbalance problem is resolved by the ensemble learning method, because it is more effective than data sampling techniques to enhance the performance of the classification model. To address this issue, we primarily segregated the active and inactive drug molecules of the dataset, which is followed by dividing the dataset of inactive molecules into 21 data frames. It is done because the number of inactive drug molecules is almost 21 times higher than the active drug molecules. Subsequently, the copy of all active drug molecules is added in all the 21 data frames so that all the data frames have an equal number of active and inactive drug molecules. Now, these 21 data frames are different and balanced datasets, which are available for model building by using ensemble learning. In the regression phase, class balancing is not required because the activity score, potency, and efficacy are continuous classes.

Table 5.10: Evaluation parameters for efficacy based on number of features

No. of features	r	R^2	RMSE	Accuracy
149	0.79	0.62	47.87	82.57
110	0.83	0.69	50.5	81.65
63	0.81	0.66	55.23	81.65
123	0.86	0.74	52.74	80.73
158	0.9	0.81	42.76	80.73
233	0.86	0.74	48.97	80.73
83	0.85	0.72	47.35	80.73

5.2.5 Target class used in classification dataset

Activity is the target class which contains its two instances that are active (1) and inactive (0). Active drug molecules have the ability to bind with the androgen receptor and produce various androgenic effects by modulating its activity; these androgenic effects are toxic for a human being. Inactive drug molecules cannot bind with the androgen receptor, and these are non-toxic. It is the job of our proposed multilevel ensemble model to classify the active (binding) molecules and inactive (non-binding) molecules from the dataset of AR [88][45].

5.2.6 Target classes used in regression dataset

1. **Activity score:** There are three kinds of drug molecules which are active, inactive and inconclusive. All these kinds of drug molecules have their clinical index, which is called activity score that combines information from its target response. Active drug molecules have their activity score between 40 to 100, inactive drug molecules have their activity score only 0, and inconclusive drug molecules have their activity score between 1 and 39 [88][13].
2. **Potency:** Amount of drug which is required to produce a given response. A lowly potent drug produces a given response at high concentrations, while a highly potent drug produces the same response only at lower concentrations [88][13].
3. **Efficacy:** A maximum response that a particular drug is capable to produce is called efficacy of drug molecules. The response of the drug is achieved against an applied dose. High efficacy of the drug has an extreme ability to initiate a response after binding to any receptor [88][13].

Figure 5.3 shows the flowchart of the proposed multilevel ensemble model, where in its first level, a new drug molecule is classified in active or inactive categories. Subsequently, if a drug molecule is found active, then the second level predict its activity score, potency, and efficacy.

5.3 Proposed Multilevel Ensemble Model

Ensemble learning is a technique to improve the accuracy of classification or regression model by combining the series of base classifiers. In the ensemble-based classification all the base classifiers vote/recommended any new data tuple, and based on these votes,

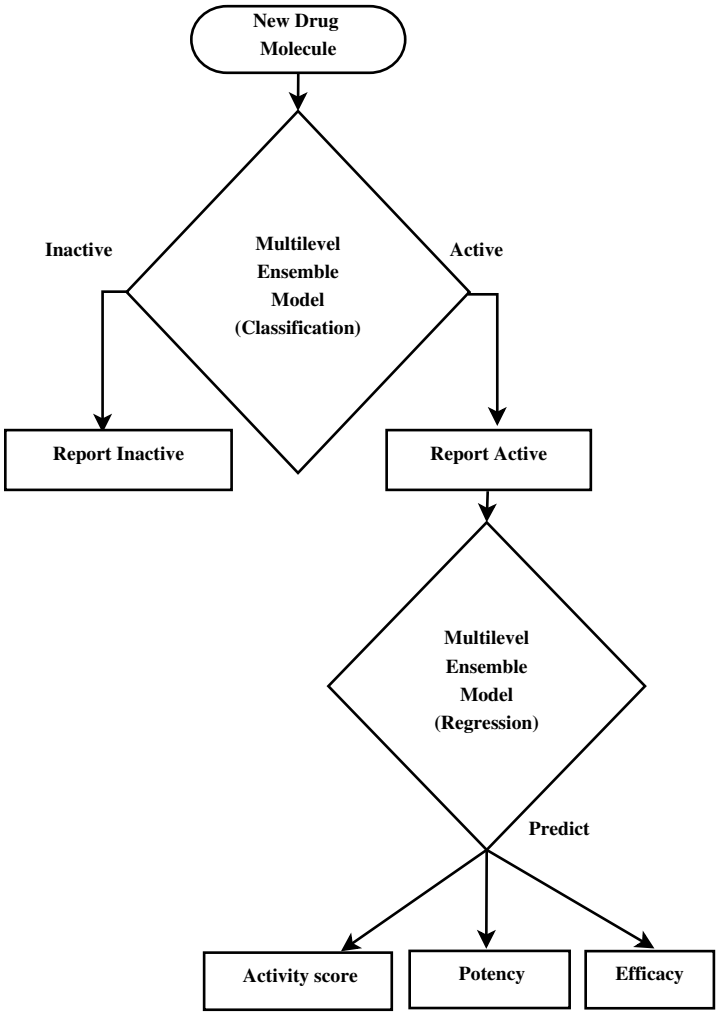


Figure 5.3: Flowchart of proposed multilevel ensemble model

a class label prediction is returned. The ensemble-based classification model can build using the same base classifiers on different splits of same training dataset or different base classifiers on the same training dataset [104]. Here, we used the first technique to perform the ensemble-based classification, where we created different divisions of the same training dataset, and the random forest model is applied as a base classifier on all these split datasets. Random forest model is taken as a base classifier because the performance of this model is better than other models. Random forest is itself an ensemble based classifier, which is the collection of several decision trees where each tree votes and most popular class is returned during classification and regression [30][13]. Therefore, our proposed ensemble approach using random forest model has improved the performance of classification as well as regression, along with solving the issue of class imbalance problem [96][105].

We used the second technique to perform ensemble-based regression, where we are taking

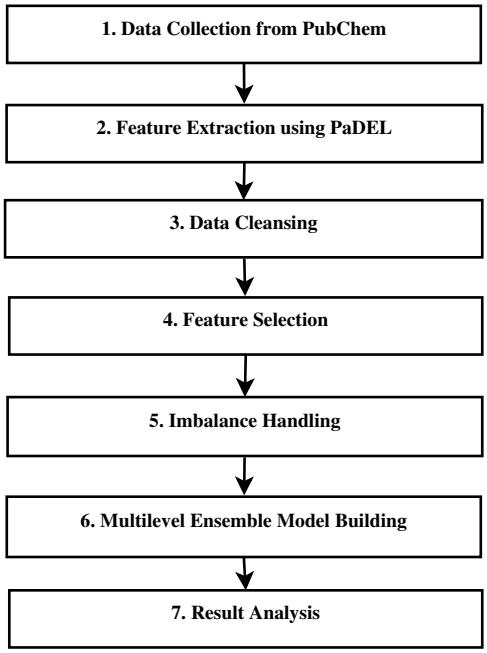


Figure 5.4: Methodology used for the proposed multilevel ensemble model

the combination of four base regression models which are linear model, random forest, neural network, and decision tree. Here, ensembling is done as taking the average of predictions of all these base classifiers. Figure 5.4 shows the methodology of the proposed multilevel ensemble model, and it is described in the following seven steps:

Step 1: Dataset generation:

The unprocessed dataset of AR is obtained from PubChem website, which is in the structure-data files (.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

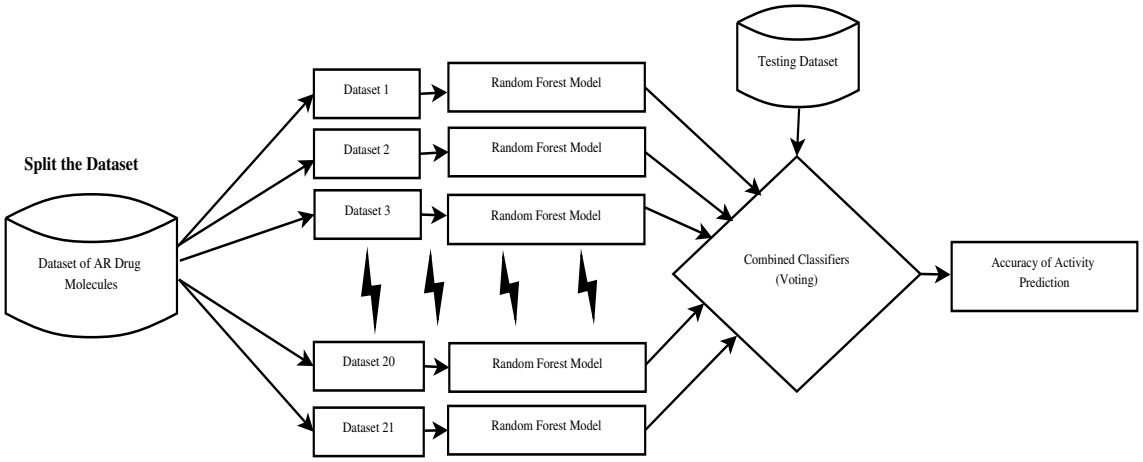


Figure 5.5: Ensemble method for activity prediction

molecules of AR. These two directories are given as input to 'PaDEL-Descriptor software' individually, which generated two comma-separated values (.CSV extension) files; one for active drug molecules and the other for inactive drug molecules. Now, these two files are combined to form the main dataset, whose target variable is a binary class activity. Here, active and inactive drug molecules are 461 and 9812, respectively, which is the causes of highly imbalance dataset.

Step 2: Determination of training and testing dataset for classification:

For the determination of the testing dataset of classification, we take 138 active drug molecules out of total 461 drug molecules, which are around 30% of whole active drug molecules, and remaining around 70% of active drug molecules are for training dataset. Similarity, we take 3029 inactive drug molecules out of total 9812 drug molecules for testing, which are around 30% of whole inactive drug molecules, and remaining around 70% of inactive drug molecules are for training dataset. Now, the total of 7106 drug molecules are available for the training dataset, and the total of 3167 are available for the testing dataset. These training and testing dataset contains both kinds of drug molecules.

Step 3: Classification model building using ensemble learning:

To remove the class imbalance problem from the training dataset, we have divided the training dataset of inactive drug molecules into 21 data frames, because the inactive drug molecules are 21 times higher than active drug molecules. Subsequently, a copy of all active drug molecules is added in all data frames of inactive drug molecules. Now, all these 21 data frames have an equal number of active and inactive drug molecules, which are available for model building by using ensemble learning. Since all these data frames are balanced and different datasets; therefore, we have trained each dataset using the base classifier random forest and combined all these classifiers using the ensemble learning method.

Step 4: Voting system:

An ensemble model for classification is developed which is based on votes of 21 random forest models, and its performance is evaluated on 30% testing tuples, which are already separated from the main dataset. Now, this ensemble model will be our final prediction model to predict the activity of any new drug molecule, whether it is active or inactive.

Step 5: Determination of training and testing dataset for regression:

After activity prediction, all the predicted active drug molecules are shortlisted. Sub-

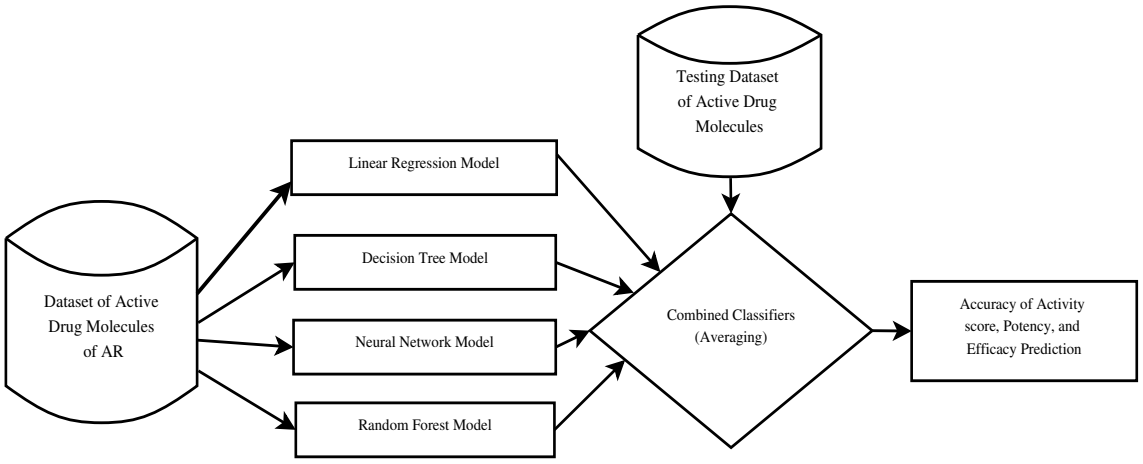


Figure 5.6: Ensemble method for activity score, potency and efficacy prediction

sequently, the prediction is performed only on regression dataset to predict the activity score, potency, and efficacy of only active drug molecules. All these target classes are predicted individually. Here, we are taking 70% data for training and 30% data for testing from the total dataset of active drug molecules.

Step 6: Regression model building using ensemble learning:

The target classes activity score, potency, and efficacy are predicted using an ensemble of four base classifiers which are decision tree, linear model, random forest, and neural network. These all target classes predicted individually.

Step 7: Average system:

An ensemble model for regression is developed which is based on the average of prediction results of all four models. Here, each model is trained at 70% of the dataset, and its performance is evaluated on the remaining 30% of the dataset which is already separated from the full dataset of predicted active drug molecules. Now, this ensemble model will be our final prediction model to predict the activity score, potency and efficacy of any new active drug molecule.

The combination of a classification ensemble model and a regression ensemble model is called the multilevel ensemble model. Figure 5.5 and Figure 5.6 show that how ensemble learning is performed during the classification level and regression level, respectively, in the proposed multilevel ensemble model.

5.4 Machine Learning Models

The combination of random forest (randomForest) models are used for the classification of activity class [89], and the combination of random forest (randomForest), linear model (lm) , decision tree (rpart), and neural networks (nnet) are used for the regression of activity score, potency, and efficacy classes . The success of our predictive model depends on the quality of a training dataset, the descriptive power of molecular descriptors, and selecting and tuning machine learning algorithms. Some parameters of the models are optimized to find a better prediction outcome. Table 5.11 shows the various models with their required packages, methods and tuning parameters. These models are used in the proposed multilevel ensemble model and implemented in R, under GNU general public license [90][91][30].

Table 5.11: 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
Neural Network (NN)	nnet	nnet	size=10
Linear Model (LM)	none	lm	method = “qr”

5.5 Model Evaluation Parameters

Various parameters are calculated to evaluate the performance of the different models of classification and regression. K-fold cross-validation is performed to check the consistency of the proposed model. According to Boruta algorithm, only 56 features are selected and the remaining 1388 features are discarded from the dataset during classification level as mentioned in Section 5.2.3. The formula for the training process of the proposed

multilevel ensemble model during the classification level is:

$$Activity \sim f(nN + ATS5s + ATS6s + AATS1i + AATS2i + ATSC4m + ATSC4v + ATSC1i + AATSC4m + AATSC3s + MATS4m + MATS1i + MATS8s + GATS3c + GATS2e + GATS1i + VE2\_Dzp + nBase + SpMAD\_Dzs + bpol + SpMin1\_Bhm + SpMin1\_Bhe + SpMin1\_Bhs + C3SP3 + C4SP3 + SCH.6 + VCH.6 + VCH.7 + SC.6 + VC.5 + VC.6 + VPC.4 + VPC.5 + VPC.6 + ASP.7 + VP.6 + maxHBa + ETA\_BetaP + VP.7 + ETA\_BetaP\_ns + ETA\_EtaP + ETA\_EtaP\_L + TIC1 + ETA\_EtaP\_F\_L + HybRatio + IC1 + IC2 + MDEC.24 + MDEC.44 + JGI2 + nFRing + nFG12Ring + nTRing + nTG12Ring + RotBFrac + WTPT.5) \tag{5.1}$$

5.5.1 Classification parameters

To check the performance of proposed and existing binary classification models, some specific evaluation parameters like Gini coefficient, specificity, sensitivity, AUC, precision, F-score, and accuracy are used for binary classes. These parameters are calculated using the following equations. All these performance metrics are essential for any binary classifiers and can be found with the help of confusion/error matrix [30], which is shown in Table 5.12.

Table 5.12: Confusion matrix for activity prediction

		Predicted class	
		Active	Inactive
Actual class	Active	TP	FN
	Inactive	FP	TN

$$Gini = 2 * AUC - 1 \tag{5.2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5.4}$$

$$Precision = \frac{TP}{FP + TP} \tag{5.5}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100 \tag{5.6}$$

- TP: Active drug molecules are classified as active drug molecules i.e., True Positive.
- TN: Inactive drug molecules are classified as inactive drug Molecules i.e., True Negative.
- FP: Inactive drug molecules are classified as active drug molecules i.e., False Positive.
- FN: Active drug molecules are classified as inactive drug molecules i.e., False Negative.

5.5.1.1 Area under the curve (AUC)

An area under the curve is also called Receiver Operating Characteristics (ROC). It is drawn between true positive rate (sensitivity) and false positive rate (100-specificity), which is found by a confusion matrix. Each point in the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The upper left corner shows 100% sensitivity, and 100% specificity, which reflects the 100% accuracy of the test. Therefore, the closing value of the ROC curve towards the upper left corner is better than other [106]. Figure 5.7 shows the ROC chart of activity classification using the proposed ensemble model, where its AUC value is 0.947.

5.5.1.2 F-score

The F-score is also called F-measure or F-distribution, which is the harmonic mean of Precision and Recall. It gives equal weight to Precision and Recall [30]. It is an alternative way to use Precision and Recall in combine them into a single measure.

$$F - score = \frac{2 * precision * recall}{precision + recall} \tag{5.7}$$

5.5.2 Regression parameters

To check the performance of proposed and existing regression models, some specific evaluation parameters like correlation (r), root mean square error (RMSE), coefficient of determination (R^2), and accuracy is evaluated for continuous classes. These parameters are explained below.

5.5.2.1 Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}} \tag{5.8}$$

where A = actual target and P = predicted target, RMSE should be reduce as low as possible [82][107].

5.5.2.2 Correlation (r)

Correlation is a statistical technique that shows how the actual values and the predicted values are related. It ranges from -1 to 1. If r is close to 0, it means there is no relationship between the variables, and if the value of r is one then it is considered as a good correlation [108]. If r is negative, it means that as one gets larger, the other gets smaller and vice versa. It is defined as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.9}$$

Where x is the actual value, y is the predicted value, $\overline{x}$ and $\overline{y}$ are the means of all the actual values and the mean of the all predicted values, respectively, and n is the total number of instances.

5.5.2.3 Coefficient of determination (R^2)

The R^2 is a statistical analyzer that show how efficient a model predicts the future outcomes [108].

$$R^2 = r * r \tag{5.10}$$

Where r is the correlation coefficient.

5.5.2.4 Accuracy

It is evaluated as the percentage deviation of the actual target to the predicted target with some acceptable error.

$$Accuracy = \frac{100}{n} \sum_{i=1}^{n} q_i$$
$$q_i = \begin{cases} 1 & \text{if } abs(P_i - A_i) \leq err \\ 0 & \text{otherwise} \end{cases} \tag{5.11}$$

where A is the actual target, P is the predicted target, err is the some acceptable error, and n is the total number of instances [82].

5.6 Result Analysis, Comparison, and Validation

The distribution of data in the training-testing experiment has been set to 70% and 30%, respectively. The Gini coefficient, specificity, sensitivity, precision, AUC, F-score, and accuracy of our proposed multilevel ensemble model during classification is 0.894, 0.979, 0.828, 0.92, 0.947, 0.948 and 92.86%, respectively, on the testing dataset of AR. The comparative performance of our proposed multilevel ensemble model with some existing classification models for the activity prediction is analysed by these metrics as shown in Table 5.13. The accuracy of the random forest, AdaBoost, SVM, and decision tree is 83.63%, 82.06%, 79.79%, and 74.73%, respectively, which are lower than our proposed ensemble model. Figure 5.8 shows the histogram for the comparison of accuracy of activity prediction using our proposed multilevel ensemble model as well as some existing models. The results show that our proposed multilevel ensemble model has outperformed the other existing machine learning models.

RMSE, correlation (r), R^2, and accuracy are the performance metrics of our proposed multilevel ensemble model during the regression phase. Table 5.15 shows the comparative performance of various models. The results show that our proposed multilevel ensemble model has outperformed the other existing machine learning models.

The proposed multilevel ensemble model has the lowest RMSE of 3.48, 5.27, and 47.87 for the prediction of activity score, potency, and efficacy, respectively, on the testing dataset. The proposed multilevel ensemble model has the highest r of 0.92, 0.77 and 0.79 for the prediction of activity score, potency, and efficacy, respectively, on the testing dataset. The proposed multilevel ensemble Model has the highest R^2 of 0.85, 0.59 and 0.62 for the prediction of activity score, potency, and efficacy, respectively, on the testing

Table 5.13: Performance comparison of proposed multilevel ensemble model with existing models in classification phase

Decision Method	Gini	Sensitivity	Specificity	Precision	AUC	F-score	Accuracy(%)
Multilevel Ensemble Model	0.894	0.979	0.828	0.92	0.947	0.948	**92.86**
Random Forest	0.78	0.759	0.912	0.895	0.89	0.821	83.63
AdaBoost	0.683	0.738	0.92	0.917	0.842	0.818	82.06
Support Vector Machine	0.741	0.652	0.942	0.917	0.871	0.763	79.79
Decision Tree	0.509	0.663	0.836	0.81	0.754	0.729	74.73

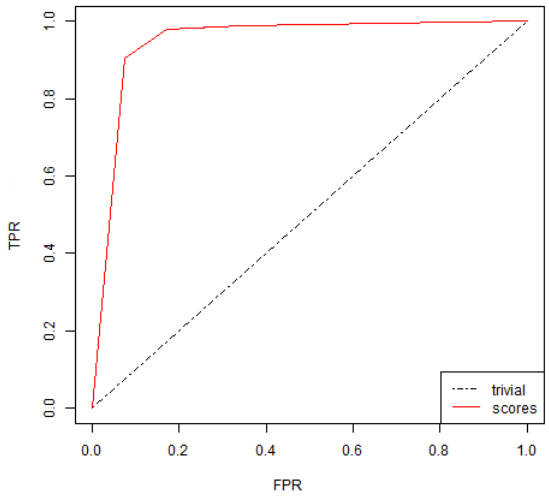


Figure 5.7: ROC performance of multilevel ensemble model on testing dataset, AUC: 0.947

dataset.

The proposed multilevel ensemble model has the highest accuracy of 92.11% (with ± 9 err), 92.66% (with ± 25 err), and 82.57% (with ± 45 err) for the prediction of activity score, potency, and efficacy, respectively, on the testing dataset. RMSE, correlation (r), R^2, and accuracy are shown in Eq. (5.8), Eq. (5.9), Eq. (5.10), Eq. (5.11), respectively, where accuracy has been calculated with some acceptable error. Table 5.15 lists the RMSE, correlation, R^2, and accuracy of all the models.

5.6.1 K-fold cross-validation

To measure the robustness of the model, the K-fold cross-validation technique shows the stable performance for the accuracy of the proposed ensemble model [30]. Here, we have used 10-fold (K=10) cross-validation for the prediction of activity, activity score, potency, and efficacy classes. In this case, at a time nine data frames are used for training and one data frame is used for testing. Table 5.14 describes the accuracy, and Figure 5.9 shows

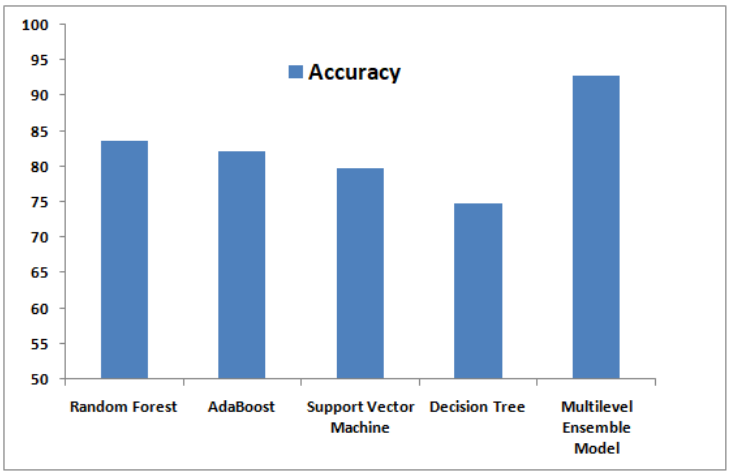


Figure 5.8: Performance comparison of proposed multilevel ensemble model with other models for activity prediction

the accuracy in the form of a line graph of the proposed ensemble model in 10-fold cross-validation for the prediction of all the target classes. These results of cross-validation show the consistent performance of the proposed ensemble model on different folds of the dataset [45]

Table 5.14: Results of 10-fold cross validation for activity, activity score, potency, and efficacy prediction

Folds	Accuracy	Accuracy score	Potency	Efficacy
1	93.29	92.97	92.52	80.56
2	94.98	90.81	91.63	81.87
3	91.71	92.35	92.92	80.80
4	94.83	92.67	91.17	80.60
5	92.65	92.19	91.06	81.34
6	94.72	92.88	92.22	80.06
7	91.40	90.63	91.44	81.53
8	93.70	90.71	91.30	82.22
9	93.73	92.22	92.84	81.79
10	94.48	90.10	92.89	80.27

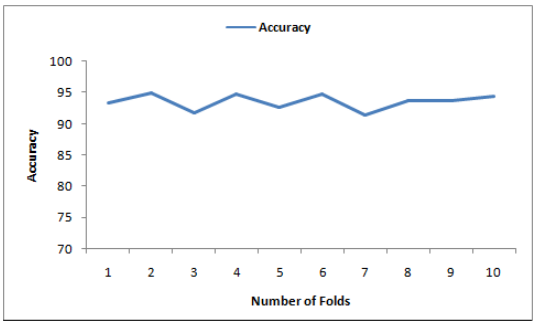
The scatter plots of R^2 are shown in Figure 5.10. It shows the relationship between actual values and predicted values of activity score, potency, and efficacy classes, which are predicted by the proposed multilevel ensemble model. These scatter plots are visualized using results of the testing dataset where points that meet the line have a better correlation.

5.6.2 Validation of proposed multilevel ensemble model

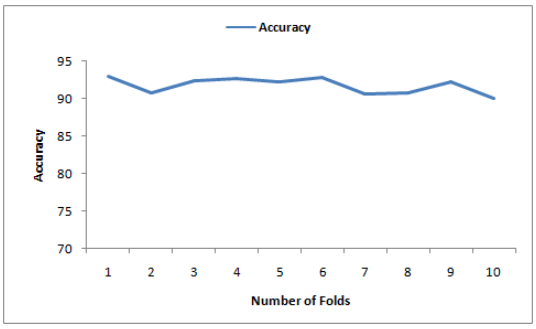
We have validated our proposed ensemble model on two validation datasets. The first dataset contains eight drug molecules, which are related to general food additives, cos-

Table 5.15: Performance comparison of proposed multilevel ensemble model with existing models in regression phase

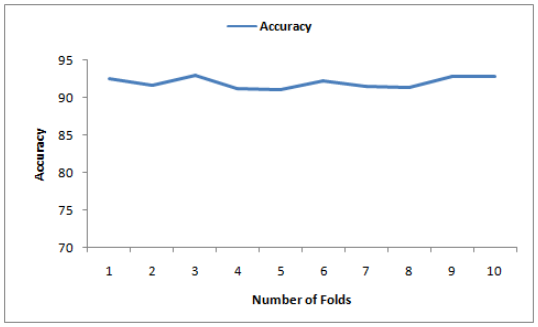
	Activity score				**Potency**				**Efficacy**			
Model Name	**r**	R^2	**RMSE**	**Accuracy** (%)	**r**	R^2	**RMSE**	**Accuracy** (%)	**r**	R^2	**RMSE**	**Accuracy** (%)
Multilevel Ensemble Model	0.92	0.85	3.48	**92.11**	0.77	0.59	5.27	**92.66**	0.79	0.62	47.87	**82.57**
Random Forest	0.94	0.88	4.57	89.72	0.62	0.38	8.03	75.69	0.8	0.64	60.7	49.54
Decision Tree	0.90	0.81	3.69	85.09	0.49	0.24	10.45	72.48	0.68	0.46	67.32	66.97
Linear Model	0.84	0.71	5.54	79.82	0.39	0.15	13.23	73.39	0.16	0.03	199.1	53.21
Neural Network	0.91	0.83	3.95	86.84	0.2	0.04	12.52	81.65	0.72	0.52	68.52	71.56



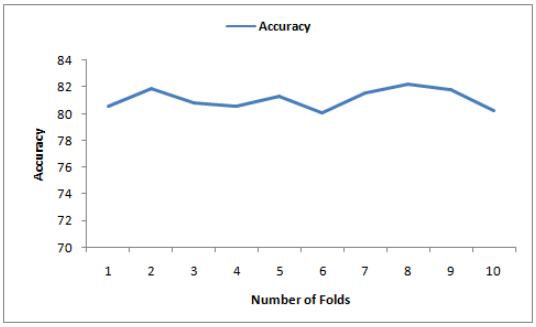
(a) Activity



(b) Activity score



(c) Potency

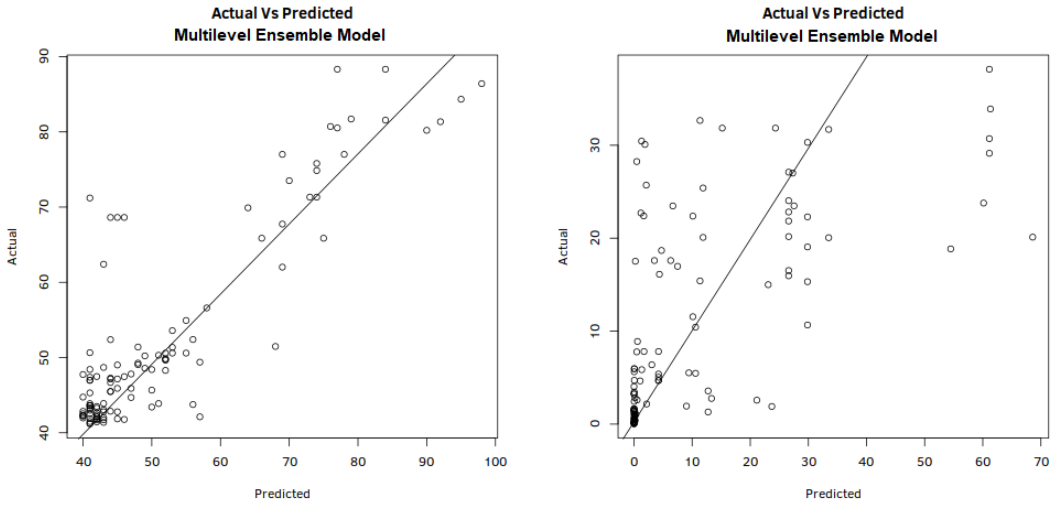


(d) Efficacy

Figure 5.9: K-fold cross validation for activity, activity score, potency, and efficacy classes

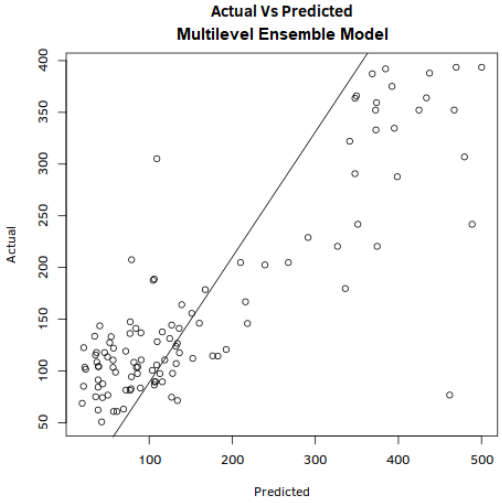
metics, detergents and preservatives [57]. The second dataset contains seven small drug molecules of estrogen receptor [73]. These, all drug molecules are neither part of training dataset nor part of the testing dataset, if the performance of our proposed multilevel ensemble model is satisfactory on these new drug molecules then it accomplishes the validation process of our proposed model. Our proposed multilevel ensemble model predicted all the withdrawn drugs "toxic," if they are toxic and "non-toxic," if they are non-toxic.

Food additives such as aspartame, saccharin, and monosodium glutamate (MSG) are predicted to be toxic by this proposed model. Aspartame and saccharin are artificial



(a) Activity score

(b) Potency



(c) Efficacy

Figure 5.10: Scatter plots for activity score, potency, and efficacy classes

non-carbohydrate calorie-free sweeteners, which are commonly used in the commercial market. Pesticides are being widely used to control insects, rodents and other pests in the agriculture fields. The continued usage of pesticides has highly detrimental environmental impacts on air, water, soil, and food, and could be toxic to humans. Their exposure has been linked to hormone disruption, cancer, reproductive development, and neurological effects like loss of memory. This model predicted the highly used pesticides likes Dimethyl tetrachloroterephthalate acid (DCPA) and Ethylenediaminetetraacetic acid (EDTA) [109][110] as toxic. Butyl hydroxy butyl nitrosamine and Sodium tetra decane sulfonate compounds, which are present in beauty and cosmetic products are predicted to be toxic by our proposed ensemble model. Imidazolidinyl urea is also used in

Table 5.16: Validation of proposed multilevel ensemble model on some new drug molecules

	Activity		**Activity score**		**Potency**		**Efficacy**	
Target Drug Molecules	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted
Aspartame (C14H18N2O5)	1	1	-	-	-	-	-	-
Saccharin (C7H5NO3S)	1	1	-	-	-	-	-	-
MSG (C5H8NNaO4)	1	1	-	-	-	-	-	-
DCPA (C10H6Cl4O4)	1	1	-	-	-	-	-	-
EDTA (C10H16N2O8)	1	1	-	-	-	-	-	-
Butyl hydroxy butyl nitrosamine	1	1	-	-	-	-	-	-
Sodium tetra decane sulfonate	1	1	-	-	-	-	-	-
Imidazolidinyl (C11H16N8O8)	0	0	-	-	-	-	-	-
NCGC00015829-02	1	1	45	43.25	10.12	10.31	182.71	168.25
NCGC00015830-02	1	1	41	40.39	11.35	12.38	29.79	30.57
NCGC00015864-02	0	0	NA	NA	NA	NA	NA	NA
NCGC00015872-05	0	0	NA	NA	NA	NA	NA	NA
NCGC00015876-07	0	0	NA	NA	NA	NA	NA	NA
NCGC00015877-06	0	0	NA	NA	NA	NA	NA	NA
NCGC00015882-09	0	0	NA	NA	NA	NA	NA	NA

cosmetics as an antimicrobial preservative due to its high solubility in water, but it is predicted to be non-toxic by our proposed ensemble model [57].

Along with these drug molecules, we have also validated our proposed model on seven drug molecules of estrogen receptor (ER). Now, we have two-dimensional structures of all 15 drug molecules, which are downloaded as a structure-data file (SDF) from the PubChem database and their molecular descriptors have been extracted with the help of PaDEL-Descriptor. After it, we applied our proposed ensemble model for activity prediction of aspartame, saccharin, MSG, DCPA, EDTA, Butyl hydroxy butyl nitrosamine, Sodium tetra decane sulfonate, Imidazolidinyl, and seven drug molecules of the estrogen receptor. The output of the proposed ensemble model is summarized in Table 5.16. The results of the table demonstrate that the predicted values are almost similar to actual values. These correct predictions of all the drug molecules show the validity of the proposed ensemble model.

5.7 Discussion

The detection of activity, activity score, potency, and efficacy of any drug molecule is important while deciding its androgenic effects on the human being. Our proposed multilevel ensemble model is the combination of 21 random forest base classifiers during its classification phase, and it is the combination of four base models which are random forest, neural network, decision tree, and linear model during its regression phase. In both levels, The performance of our proposed multilevel ensemble model is compared with some existing models, where our proposed model outperformed the other models. In

the classification level, we compared our proposed model with random forest, AdaBoost, decision tree, and support vector machine because these compared models perform better for the binary classification. In regression level, we compared our proposed model with random forest, decision tree, linear model, and neural network because these models perform better for the continuous data classes.

5.8 Conclusion

To find out the toxic drug molecules from a similar kind of drug molecules' dataset is a complex task. There is a severe need to design a computational method to predict the toxicity of drug molecules during the pre-clinical trial to save the life of animals, time and cost. In this paper, we have proposed a novel machine learning based multilevel ensemble model to the assessment of the toxicity of AR's drug molecules. This prediction is based on QSARs/QSPRs/QSTRs approach where we have solved the problem of toxicity prediction of those drug molecules that can bind to androgen receptors and produce toxic effects. The target classes for toxicity prediction are activity, activity score, potency, and efficacy. The dataset used in this study is very high in features and exceptionally class imbalanced. Initially, feature selection is performed using Boruta() and information.gain() methods and balanced the dataset by dividing it. After all these data pre-processing approaches, we have developed a multilevel ensemble model where we applied ensemble learning in its classification phase as well as in its regression phase. Our proposed ensemble model is evaluated on different performance parameters, and compared with other peer models for classification phase and for regression phase, individually. Through the intensive experiments and comparisons, it concludes that our proposed multilevel ensemble model, in spite of having highly imbalanced data, outperformed other examined methods in its both levels and its performance is almost consistent in k-fold cross-validation. Finally, to prove the validity of the proposed ensemble model, we have tested it on some new drug molecules where it performed very well.

Chapter 6

Ensemble Technique for Toxicity Prediction of Small Drug Molecules of the Antioxidant Response Element Signalling Pathway

The in silico toxicity prediction techniques are useful to reduce rodent testing (in vivo). Authors have proposed a computational method (in silico) for the toxicity prediction of small drug molecules using their various physicochemical properties (molecular descriptors), which can bind to the antioxidant response elements (AREs). The software PaDEL-Descriptor is used for extracting the different features of drug molecules. The ARE dataset has total 7439 drug molecules, of which 1147 are active and 6292 are inactive, and each drug molecule contains 1444 features. We have proposed a novel ensemble-based model that can efficiently classify active (binding) and inactive (non-binding) compounds of the dataset. Initially, we performed feature selection using random forest importance algorithm in R, and subsequently, we have resolved the class imbalance issue by ensemble learning method itself, where we divided the dataset into five data frames, which have an almost equal number of active and inactive drug molecules. An ensemble model based upon the votes of four base classifiers is proposed, which gives an accuracy of 97.14%. The K-fold cross-validation is conducted to measure the consistency of the proposed ensemble model. Finally, the proposed ensemble model is validated on some new drug molecules and compared with some existing models.

6.1 Introduction

The drug is a combination of chemicals that prevent disease or assist in restoring the health of the human being. One of the primary means of conserving human health is maintaining by administrating the small molecule that is the chemical construction of drug. Drug designing is an inventive process; a new medication is found on the basis of the knowledge of biological target [3]. Organically drug is a small molecule (a lower molecular weight is <900 Daltons) that inhibits or activates bio-molecule such as protein.

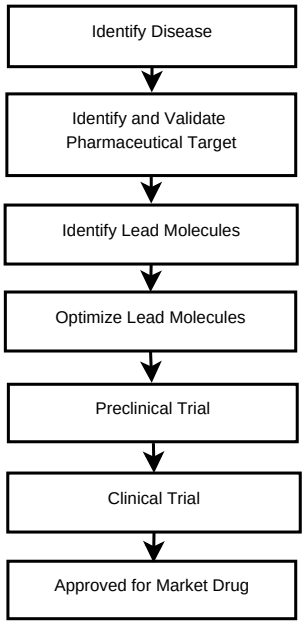


Figure 6.1: Major milestones in drug development

The more exact term for designing a molecule is ligand design which binds precisely to its target. Before a ligand design, we need to optimize many properties of it such as side effects, metabolic half-life, etc., for the safety and effectiveness of the drug.

Successful drug molecules bind to specific components (usually proteins) of the target cells. After it, these cells activate or deactivate those components which lead to modify cells behaviour or destruct them. Introducing a novel drug is not an easy job, therefore, the drug development industries have been spending huge efforts and time not only to discover new drug but also to improve existing medicines. Averagely, the discovery of new drug development takes 10-15 years and costs 400-800 million US dollars [4]. In contemporary drug discovery, Figure 6.1 shows the major milestones of drug development.

During drug development, first, developers perform pre-clinical trials on the animals then go for clinical trials on humans. In spite of pre-clinical studies of drugs, >30% promising pharmaceutical get failed in human clinical trials because they are determined to be toxic [10]. There are >40 million compounds that are available for virtual screening on the ZINC database [111] and PubChem database [112]. These databases are also used for commercial purposes. The goal of this work is to quickly and efficiently test whether certain chemical compounds have the potential to disrupt processes in the human body that may lead to adverse health effects.

Physicochemical properties of chemical compounds always guide to determine its activity (active or inactive); therefore, it has been rigorously used to classify a toxic drug and non-toxic drug. In this work, we try to apply four machine learning models which are

Table 6.1: Various NRs and SR pathways with their PubChem ID

Assay ID	**Assay**	**PubChem ID**
NR-AR	Androgen receptor, full length	743040
NR-AR-LBD	Androgen receptor, LBD	743053
NR-ER	Estrogen receptor alpha, full length	743079
NR-ER-LBD	Estrogen receptor alpha, LBD	743077
NR-AhR	Aryl hydrocarbon receptor	743122
NR-PPAR-gamma	Peroxisome proliferator-activated receptor gamma	743140
NR-aromatase	Aromatase	743139
SR-Nrf2/ARE	Nuclear factor-like 2/antioxidant responsive element	743219
SR-HSE	Heat shock factor response element (HSE)	743228
SR-ATAD5	Genotoxicity indicated by ATAD5	720516
SR-MMP	Mitochondrial membrane potential	720637
SR-p53	p53	720552

the random forest, AdaBoost, decision tree and support vector machine to predict the toxicity of the chemical compound using their physicochemical properties. There are more than 8000 molecular descriptors (physicochemical properties) that are identified for chemical compounds that will help in train and test the predictive model. To calculate the molecular descriptors several open source and commercial tools are available.

Active drug molecules are those molecules, which can bind to one or more biochemical pathway assays and create some toxic effects into our body. These toxic effects are stress response (SR) effects and nuclear receptor (NR) effects. Both the SR and NR effects are highly relevant to human health because the activation of NRs can disrupt endocrine system function, and the activation of SR pathways can lead to liver injury or cancer [11]. The role of a drug is to correct the functioning of these SR signalling pathways or NR signalling pathways [12]. We can build computational models to predict the activity of the drug molecules in antioxidant response elements (AREs) based on their physicochemical properties [13]. Table 6.1 shows the 12 biological pathway assays along with their PubChem identification number; these bioassays can give distinct negative health effects on its activation. PubChem is the world's largest database of freely accessible chemical information. We can search for any chemical by its name, molecular formula, structure, and other identifiers. PubChem identification number is a unique identifier of a chemical compound in its database.

AREs are the central part of the signal transduction pathway in eukaryotic cells that respond to oxidative stress [14]. Reactive oxygen species (ROS) and reactive nitrogen species (RNS) are generated in the body from internal metabolism. In normal cells, ROS are generated in a controlled manner and serve some useful purposes. Oxidants formed in response to physiological sign act as important signalling molecules to regulate these processes as inflammation, cell division, autophagy, immune function and SR. Uncontrolled oxidants production is the consequence of oxidative stress that impairs cellular functions and produces a risk of toxicity, cancer and chronic disease [113].

Cigarette, asbestos, petrol, kerosene, coal tar, and diesel exhausted particles are the pro-oxidant in nature and contain free radicals, electrophiles, and carcinogens, which include polycyclic aromatic hydrocarbons, that are the consequence of airway diseases. These airway diseases include asthma, chronic obstructive pulmonary disease (COPD), fibrosis, emphysema, acute respiratory distress syndrome (ARDS), and bronchial carcinogenesis [114].

AREs are well established as significant regulators of redox homeostasis and activators of cytoprotection during oxidative stress [14]. Oxidative stress has been implicated in the pathogenesis of a variety of diseases ranging from cancer to neurodegeneration. The ARE signalling pathway plays an important role in the amelioration of oxidative stress.

The CellSensor ARE-bla HepG2 cell line (Invitrogen) can be used for analyzing the Nrf2/antioxidant response signalling pathway. Nrf2 (nuclear factor erythroid 2-related factor 2) and Nrf1 (nuclear factor erythroid 2-related factor 1) are transcription factors that bind to AREs and activate these genes. The CellSensor ARE-bla Hep G2 cell line contains a beta-lactamase reporter gene under control of the antioxidant response element (ARE) stably integrated into HepG2 cells. This cell line has been used to screen the Tox21 10K compound library to identify agonists that induce oxidative stress [115].

Here, we are developing a model based on the quantitative structure-activity relationship (QSAR) model. QSAR is a classification or regression model to predict the biological activity of the drug molecule using its physicochemical properties. We are creating a binary classifier for the toxicity prediction of small drug molecules of ARE. Toxicology is the art of identifying unexpected human health effects and risk analysis based on data from pre-clinical animal models and physicochemical properties. Here, we are providing a conceptual overview of predictive toxicology, which relates to building a toxicology model on the basis of scientific advances in the molecular, cellular and computational sciences.

Systems biology, bioinformatics, and bioassay technologies are helpful to scientists to understand how pathways or cellular networks in the human body carry out normal functions, which are keys to maintaining health. The causes of adverse health effects depend on the alteration of important pathways by chemical exposures. However, these effects only occur when exposures are of sufficient duration and intensity. Therefore, we can say that a new toxicity-testing system relies mainly on understanding toxicity pathways, the cellular response pathways that can result in adverse health effects. Such a system would evaluate biologically significant alterations without the use of whole animals.

In this paper, we have proposed a novel ensemble-based binary classification model to predict the activity of ARE 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 random forest importance algorithms, then the class imbalance problem is resolved through the ensemble learning method where we have divided the dataset into five data frames, which have an equal number of active or inactive drug molecules. Subsequent to this, four data frames are used for training and one data frame is used for testing. Therefore, we trained the model at 80% and tested at 20%. An ensemble model based upon the votes of random forest, AdaBoost, decision tree, and SVM models is created that is our proposed ensemble model as well as it has also resolved the issue of class imbalance. The K-fold cross-validation is performed to measure the robustness of the proposed ensemble model. Finally, we have proved the validity of the proposed ensemble model on some new drug molecules, which are neither part of the training dataset nor part of the testing dataset, and in the last, we have compared our proposed ensemble model with some existing models. The major highlights of this work are the following:

1. This study focused on predicting the toxicity of drug molecules on the basis of ensemble learning approach.

2. We focused on accuracy as the most important criteria of activity prediction. Apart from it, other metrics like Gini, sensitivity, specificity, precision, kappa, AUC, F-score, and MCC are also evaluated.

3. The performance of our proposed ensemble model is compared with various standard classifiers like SVM, decision tree, random forest, and AdaBoost, where our proposed ensemble model outperformed the others.

4. We have validated our proposed ensemble model on some other drug molecules, which are neither part of training dataset nor part of the testing dataset. These drug molecules are related to food additives, cosmetics, detergents, preservatives, and ATAD5.

Figure 6.2 shows the general diagram of the prediction model, where the various physicochemical properties of any small drug molecules of AREs are taken and their activities are predicted through our prediction model.

The paper is composed as follows: Section 6.2 introduces a quick overview of the dataset, PaDEL, feature selection, and class imbalance problem. Section 6.3 clarifies the procedure of the proposed ensemble model. Section 6.4 presents the description of the various models, which are used as the base classifier for ensemble learning. Section 6.5 presents

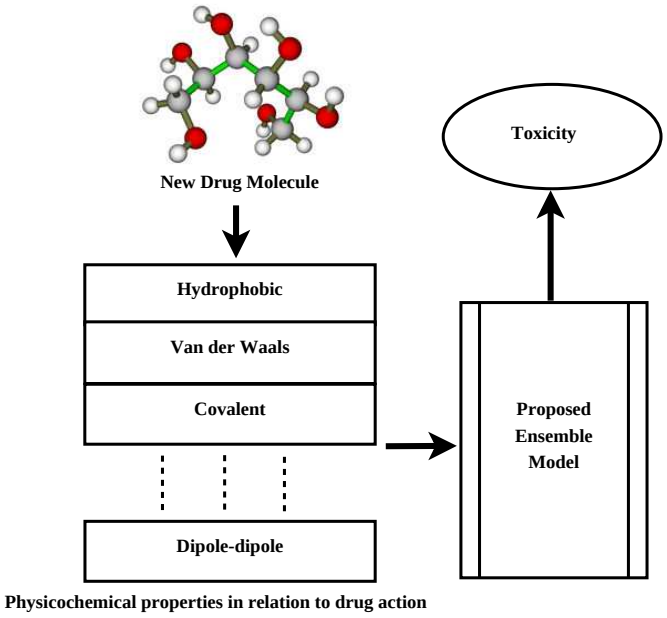


Figure 6.2: Prediction model for ARE

the different binary classification based performance metrics. Section 6.6 describes the result investigation, comparison, and validation of the proposed ensemble model. Section 6.7 shows discussion regarding the performance evaluation parameters, which is followed by the conclusion in Section 6.8.

6.2 Materials and Methods

6.2.1 Feature extraction using pharmaceutical data exploration laboratory (PaDEL)

PaDEL-Descriptor is a free and open source software developed in Java language, and it is used to calculate molecular descriptors. It supports mainly MOL, structure-data file (SDF) and SMILES file formats [2]. Table 6.2 shows the format of the SDF for an active drug molecule of ARE. According to Todeschini and Consonni [7], the molecular descriptor is the final result of the logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number. Table 6.3 describes some essential molecular descriptors' type, class, definition and names of the individual values. Physicochemical properties play a significant role in forming of the molecular descriptors, and the help of these molecular descriptors can find biological activities. It is found that the octanol/water partition coefficient, molar

Table 6.2: SDF format for the single active drug molecule of ARE

```
NCGC00166121-03
  OpenBabel06141707472D

  19 17 0 0 0 0 0 0 0 0999 V2000
   5.7273 -0.7167 0.0000 N 0 3 0 0 0 0 0 0 0 0 0 0
   5.0041 -0.2970 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
   6.4569 -1.1300 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
   5.3205 -1.4399 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
   6.1470 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
   4.2874 -0.7167 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
   2.8475 -1.5432 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
   3.5642 -1.9500 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
   5.9081 -1.6530 0.0000 Br 0 5 0 0 0 0 0 0 0 0 0 0
   1 2 1 0 0 0 0
   1 3 1 0 0 0 0
   1 4 1 0 0 0 0
   1 5 1 0 0 0 0
   2 6 1 0 0 0 0
   6 13 1 0 0 0 0
   7 8 1 0 0 0 0
   7 18 1 0 0 0 0
M CHG 2 1 1 19 -1
M END
> <Formula>
C17H38BrN

> <FW>
336.3943 (256.4898+79.9045)

> <DSSTox_CID>
24367

> <SR-ARE>
1
```

refractivity, dipole moment, and density are quite abundant in chemical information to encode the structural features that contribute to the various toxicities, and these toxicities come under biological activities [116]. The input of PaDEL-Descriptor is the SDF of ARE drug molecules, and its output is comma-separated values (CSV) file that contains a total of 7439 drug molecules of ARE. The PaDEL-Descriptor calculated 1444 1D and 2D molecular descriptors for these drug molecules, which are the features of our proposed ensemble model.

6.2.2 Dataset

The ARE signalling pathway dataset is taken from PubChem, where 743219 is the PubChem identification number for ARE. PubChem is a freely available chemical database of small organic molecules and has information about their biological activities [115]. Its website is 'https://pubchem.ncbi.nlm.nih.gov/bioassay/743219'. PubChem is organized and maintained by the National Center for Biotechnology Information (NCBI) and has

Table 6.3: Molecular descriptors

Descriptor type	Descriptor class	Definition	Descriptor name
ALOGP	Constitutional Descriptor	Calculates atom additive logP and molar refractivity values as described by Ghose and Crippen	ALogP, ALogp2, AMR
APol	Electronic Descriptor	Descriptor that calculates the sum of the atomic polarizabilities	apol
AminoAcidCount	Protein Descriptor	Returns the number of amino acids found in the system	nA, nR, nN, nD, nC, nF, nQ, nE, nG, nH, nI, nP, nL, nK, nM, nS, nT, nY, nV, nW
AromaticAtomsCount	Constitutional Descriptor	Descriptor based on the number of aromatic atoms of a molecule	naAromAtom
AutocorrelationMass	Topological Descriptor	The Moreau-Broto autocorrelation descriptors using atomic weight	ATSm1, ATSm2, ATSm3, ATSm4, ATSm5
BondCount	Constitutional Descriptor	Descriptor based on the number of bonds of a certain bond order	nB
KappaShapeIndices	Topological Descriptor	Descriptor that calculates Kier and Hall kappa molecular shape indices	Kier1, Kier2, Kier3
PetitjeanNumber	Topological Descriptor	Descriptor that calculates the Petitjean Number of a molecule	Petitjean Number
RuleOfFive	Constitutional Descriptor	This Class contains a method that returns the number failures of the Lipinski's Rule Of Five	Lipinski Failures
WeightedPath	Topological Descriptor	The weighted path descriptors characterize molecular branching.	WTPT1, WTPT2, WTPT3 WTPT4, WTPT5

Table 6.4: Dataset of ARE signalling pathway

Name	Activity	ALogP	GATS3s	GATS1e	CrippenLogp	MATS1e	IC2
NCGC00179658-04	0	-0.0719	0.828973952	0.556220751	3.95677	0.083016158	4.461091334
NCGC00257825-01	0	-0.7037	0.650664462	0.68378286	2.08888	-0.071277463	2.852217001
NCGC00256532-01	0	-1.2426	0.43035223	0.282995597	1.0297	0.404075338	2.582793613
NCGC00255471-01	1	0.6148	1.276277114	0.813611134	1.656	-0.109233375	3.607475391
NCGC00255854-01	1	0.2762	1.05806013	1.207252358	0.35038	-0.133621527	3.46132014
NCGC00257536-01	1	-1.0852	1.326692928	0.887052279	2.18458	-0.142264794	3.58418372

three linked databases within the NCBI's Entrez information retrieval system. These are PubChem Compounds, PubChem Substances, and PubChem BioAssays. We are dealing with the PubChem BioAssays only. Here, our dataset consists a total of 7439 ARE's drug molecules, of which 1147 are active molecules and the remaining 6292 are inactive molecules. All the drug molecules contain 1444 features, and also known as molecular descriptors. The most common molecular descriptors are the partition coefficient (AlogP), volume, molar refractivity (AMR), elements count, ETA descriptors, autocorrelation, nBase, nRing, apol, number of carbon atoms (nC), and number of hydrogen atoms (nH) and so on.

Table 6.4 shows the glance of the dataset that contains various ARE drug molecules, such as NCGC00179658-04, NCGC00257825-01, and NCGC00256532-01. The columns of the Table 6.4 shows the various molecular descriptors/features, such as ALogP, GATS3s and CrippenLogP. These all features and drug molecules are extracted from the SDFs by using PaDEL-Descriptor. Here, activity is a target class, which shows whether a drug molecule is active or inactive.

Table 6.5: Important features of antioxidant response element

S.No	Features	Mean decrease accuracy	S.No	Features	Mean decrease accuracy
1	SpMax1_Bhm	9.0766	27	IC2	5.6021
2	GATS8e	8.9279	28	GATS2e	5.5501
3	gmin	8.7925	29	GATS2s	5.4207
4	ALogP	8.5133	30	MIC1	5.3805
5	minddssS	8.1146	31	R_TpiPCTPC	5.2156
6	GATS3s	7.8081	32	GATS6e	5.1135
7	BCUTc.1	7.7205	33	minsCH3	5.1013
8	AATS5p	7.7131	34	SPMin6_Bhv	5.0730
9	minsOm	7.4152	35	MATS1s	5.0473
10	GATS1e	7.1466	36	GATS2p	5.0281
11	SpMax5_Bhm	7.1183	37	MaxHssNH	4.9721
12	TIC1	6.9687	38	MAXDN	4.9512
13	maxsOm	6.7789	39	AATS4p	4.9218
14	SsOm	6.4716	40	SpDiam_dzi	4.9038
15	IC1	6.3548	41	AATS7p	4.8437
16	VE1_Dze	6.2444	42	GATS3c	4.7560
17	SpMax4_Bhm	6.1907	43	AATSC1s	4.7362
18	CrippenLogp	6.0897	44	nsBr	4.7181
19	MATS1e	5.9891	45	ATSc3i	4.6617
20	ATSC4m	5.9777	46	GATs8s	4.6488
21	GATS8m	5.8385	47	GATs5s	4.6475
22	GATS8c	5.7944	48	minHother	4.5705
23	GATS6s	5.7650	49	GATs4c	4.4819
24	Kier2	5.7191	50	AATS6v	4.4683
25	CrippenMR	5.6395	51	SpDiam_Dzp	4.4593
26	BCUTc.1h	5.6311	52	JGI10	4.4256

6.2.3 Feature selection using random forest importance algorithm

Feature selection is a process of select the important features that may improve the performance of the model and remove those attributes that have redundant and irrelevant information [13]. Our dataset has 1444 features, which are very high in quantity; therefore, it will increase the time and space complexity of the model and also create a complex model. Here, the process of feature selection is carried out using random.forest.importance() function under FSelector library in R. The algorithm finds weights of attributes using randomForest algorithm [80].

The input parameters of random.forest.importance() function are the dataset of 1444 features of ARE and the target variable is the activity. After the execution of this algorithm, it calculated importance of each descriptor using the mean decrease in accuracy value and mean decrease in node impurity. We have selected the attribute based on the mean decrease in accuracy value. After applying the random forest feature selection algorithm on the dataset 1392 features are discarded, and only 52 features are considered as important features using cutoff.k function of R. Table 6.5 shows all the important features of ARE dataset.

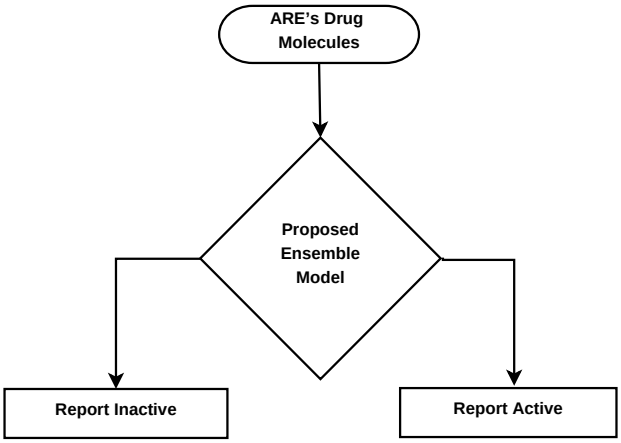


Figure 6.3: Workflow for the classification of ARE drug molecules

6.2.4 Class imbalance

Here, the used dataset is highly imbalanced, as the total number of active drug molecules is 1147 (minority/positive class) and the total number of inactive drug molecules is 6292 (majority/negative class). Then we can say that active drug molecules are far less than the inactive drug molecules. Therefore, before going to the model formation, we have managed the class symmetry of active instances and inactive instances by using the ensemble learning method, as the ensemble learning is more effective than data sampling techniques (oversampling, undersampling and SMOTE). It is performed by the creation of five data frames by dividing the main dataset (refer to Step 5 of Section 6.3 for more details), these all data frames have almost an equal number of active and inactive drug molecules of ARE [96].

6.2.5 Target class

Activity is the target class which has binary instances i.e. active (1) and inactive (0). Active compounds have the capability to bind with ARE and produce toxic effects by modulating its activity, and inactive compounds are non-toxic and can not bind with ARE. The intensity of the toxic effects of an active drug molecule can be analysed by its activity score, potency and efficacy. The active drug molecules are harmful and can disrupt the processes in the human body. Therefore, we should remove these kinds of molecules at their pre-clinical stage of drug development to save the lives of animals as well as money and time. Figure 6.3 shows the flowchart where our proposed ensemble-based classification model is classifying ARE's drug molecules in active and inactive categories.

6.3 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 and based on these votes, a class label prediction is returned. Here, we have made an ensemble-based classification model using random forest, SVM, decision tree model and AdaBoost. This approach is to improve our classification accuracy as well as to solve the issue of class imbalance [30][104]. Here, random forest, SVM, decision tree model, and AdaBoost are taken as the base classifiers because the performance of these models are better than other models for binary classification. Figure 6.4 shows the methodology of the proposed ensemble-based prediction model, and Figure 6.5 shows the approach of ensemble learning which is applied in proposed model. The following seven steps are showing the methodology of our proposed model.

Step 1: Dataset generation
The unprocessed dataset of ARE is obtained from PubChem website, which is in the SDF 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 ARE. These two directories are given as input to 'PaDEL-Descriptor software' individually, which has generated two CSV files; one for active drug molecules and the other for inactive drug molecules. Now, these two datasets are combined to form a resultant dataset, whose target variable is a binary class activity (refer to Subsection 6.2.2 for more details).

Step 2: Feature extraction using PaDEL
Feature extraction is performed using PaDEL software that extracts 1444 1D and 2D molecular descriptors. These molecular descriptors are the encoded form of physicochemical properties of various drug molecules, which are the features of our dataset (refer to Subsection 6.2.1 for more details).

Step 3: Data cleaning
Data preprocessing can be performed by using data cleaning and feature selection. Before model formation, we have cleaned the dataset in order to remove the various discrepancies and to improve the quality of input data to achieve highly accurate and consistent knowledge [97]. Initially, we found a few corrupted and missing entries in our dataset. Therefore, we have corrected these corrupted entries and then analysed those attributes that have missing values in their cells. We have filled these missing values by the average value of that particular column.

Step 4: Feature selection

Our dataset has 1444 features, which are very high in dimensionality; therefore, feature selection is required. Feature selection is a way for feature dimensionality reduction that improves the performance of the classifiers in machine learning [30]. Here, random.forest.importance algorithm is applied on the dataset, which selected only 52 important features (refer to Subsection 6.2.3 for more details). The following formula is showing the target class and its corresponding features, which is used in our proposed ensemble model as well as those models that are used for comparison:

$$\begin{aligned} Activity \sim f(&SpMax1\_Bhm + GATS8e + gmin + ALogP + minddssS + GATS3s + \\ &BCUTc.1l + AATS5p + minsOm + GATS1e + SpMax5\_Bhm + TIC1 + maxsOm + \\ &SsOm + IC1 + VE1\_Dze + SpMax4\_Bhm + CrippenLogP + MATS1e + ATSC4m + \\ &GATS8m + GATS8c + GATS6s + Kier2 + CrippenMR + BCUTc.1h + IC2 + \\ &GATS2e + GATS2s + MIC1 + R\_TpiPCTPC + GATS6e + minsCH3 + SpMin6\_Bhv + \\ &MATS1s + GATS2p + maxHssNH + MAXDN + AATS4p + SpDiam\_Dzi + AATS7p + \\ &GATS3c + AATSC1s + nsBr + ATSC3i + GATS8s + GATS5s + JGI10 + minHother + \\ &GATS4c + AATS6v + SpDiam\_Dzp) \end{aligned} \tag{6.1}$$

Step 5: Class imbalance handling

The dataset found from PubChem is highly imbalanced, as it has a total of 7439 drug molecules out of which 1147 are active and 6292 are inactive. To resolve this problem, we have primarily segregated the active and inactive molecules of the dataset, which is followed by dividing the dataset of inactive molecules into five data frames. This is done as the number of inactive drug molecules is more than five times of the active drug molecules. Subsequently, the copy of all active molecules is added in all the five data frames, so that all the data frames have almost an equal number of active and inactive drug molecules. Now, these five data frames are different and balanced datasets that are available for model building by using ensemble learning.

Step 6: Classification model building using ensemble learning

We have five small datasets (data frames), which are individual and different. Out of these five data frames, We have utilized four data frames (80% of the total dataset) for training and the remaining one data frame (20% of the total dataset) for testing. The training dataset's first, second, third, and fourth data frames are trained using the base

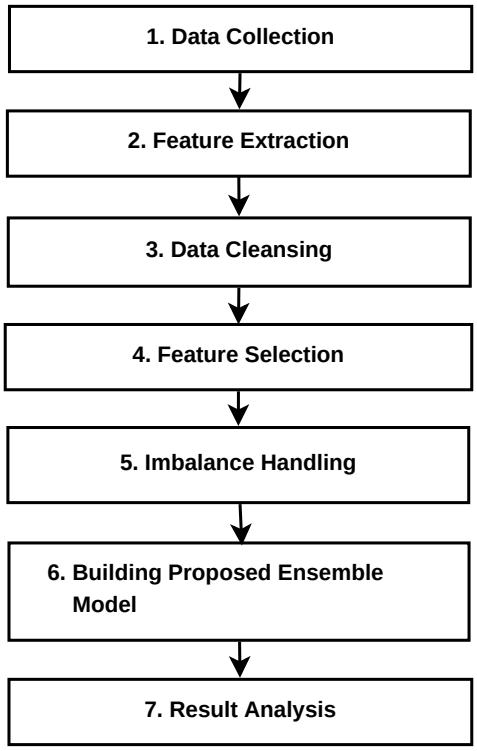


Figure 6.4: Methodology used

classifiers random forest, SVM, decision tree and AdaBoost, respectively, and combined all these classifiers using the ensemble learning method.

Step 7: Prediction of results by voting system
An accuracy of ensemble model is evaluated at 20% testing tuples (one data frame); it is evaluated based on the votes of four base models which are the random forest, decision tree, SVM and AdaBoost. Therefore, this ensemble model is the combination of four base models and will be our final prediction model to predict the activity of any new drug molecule, whether it is active or inactive. This ensemble model perfectly predict the testing samples, and provides reliable and accurate results for them because these samples are predicted through the voting of four base classifiers.

6.4 Machine Learning Models

Models used for the classification of activity are explained below. Different models describe their required packages, methods and tuned parameters. Some required parameters of all used models are tuned to get a better prediction outcome. Decision tree, SVM, random forest and AdaBoost with its tuned parameters have been used in our proposed

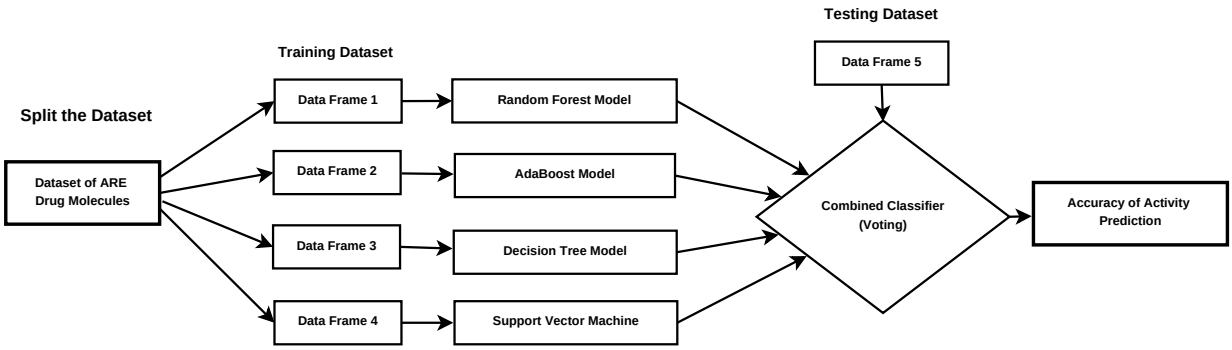


Figure 6.5: Proposed ensemble method for activity prediction

ensemble model. All the models are implemented in R, under GNU general public licence.

6.4.1 Decision tree (rpart)

It is a machine learning model that predicts the value of the activity target class based on given input features. It looks like a flowchart where each interior node represents a test on an input feature, there are edges to children represents an outcome of the test and each leaf node or terminal node represents a class label [26].

The rpart package contains the method rpart(), which is used in R for decision tree induction. It contains various parameters, where we improved the performance of rpart by tuning the parameters usesurrogate and maxsurrogate. When these values are set to 0 then computation time will be reduced because approximately half of the computational time is used to search for surrogate splits [90].

6.4.2 Support vector machine (ksvm)

It is used mostly in linear classification problems. Here, we create a hyperplane that separates the two classes (active or inactive) in n-dimensional space very well, where n represents the number of ARE's features [30].

The kernlab package contains the method ksvm() in R; the ksvm method of support vector machine can perform classification. We improved the performance of ksvm by tuning the parameter 'kernel' and 'type'. The kernel function is used in the training and prediction;

we have taken radial basis kernel function 'Gaussian' (rbfdot) for better performance. The type parameter of it shows whether you want to perform classification or regression. Here, SVM is used only for classification; therefore, we have selected classification based parameter (type=C-svc) [92].

6.4.3 Random forest (randomForest)

Random forest is an ensemble based model, which is the collection of various decision trees. Here, a random forest model provides the result based on voting of various decision trees for a particular class. A sample is predicted in the favour of that class, which found maximum votes [30].

The randomForest package contains the method randomForest() in R. The randomForest() has various parameters, where 'mtry' and 'ntree' parameters are used for tuning, mtry is the number of variables randomly sampled as candidates during each split, and ntree is the number of trees to grow. We found the better performance of random forest model by assigning the value of mtry=2 and ntree=500 [89].

6.4.4 AdaBoost (ada)

It is the boosting algorithm for binary classification, which is also called adaptive boosting. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one.

The ada package contains the method ada() in R. The AdaBoost model performs better for discrete data; therefore, we have selected its type parameter 'discrete', which performs discrete boosting. We have chosen the value of iter=50, which shows the efficient boosting iteration of AdaBoost, and we have selected the value of shrinkage nu=0.5 for performance boosting [117].

6.5 Binary Classification Based Performance Metrics

Performance of proposed and existing binary classification models are evaluated by some specific binary classification performance metrics like Gini coefficient, sensitivity, speci-

ficity, precision, F-score, MCC, kappa, AUC and accuracy. These parameters are described below and can be found with the help of confusion/error matrix.

- Area under the curve (AUC):

The Receiver operating characteristics (ROC) is a curve, which is also called an area under the curve. It is drawn between true positive rate (sensitivity) and false positive rate (1-specificity), which is found by a confusion matrix. Each point in the ROC curve represents a specificity/sensitivity pair corresponding to a particular decision threshold. The closing value of the ROC curve towards the upper left corner is better than other values[106].

- Gini coefficient:

$$Gini = 2 * AUC - 1 \tag{6.2}$$

- Sensitivity:

$$Sensitivity = \frac{TP}{TP + FN} \tag{6.3}$$

- Specificity:

$$Specificity = \frac{TN}{TN + FP} \tag{6.4}$$

- Precision:

$$Precision = \frac{TP}{TP + FP} \tag{6.5}$$

- Accuracy:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100 \tag{6.6}$$

- Matthews correlation coefficient (MCC):

The matthews correlation coefficient is a performance metric, which is used to measure the quality of a binary classification problem, it provides a balanced measure even if a dataset is imbalanced. It takes true or false positive and true or false negative parameters from the confusion matrix. It gives a result between -1 and +1, where +1 shows a perfect agreement between prediction and observation, -1 indicates total disagreement between prediction and observation, and 0 shows no better than random prediction [118].

Accuracy is not useful when the two classes are of very different sizes. Our activity class is very imbalanced, but the value of MCC is not differing too much from accuracy because we have already balanced the instances of class.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6.7}$$

FP: Inactive drug molecules are classified as active drug molecules i.e. false positive
FN: Active drug molecules are classified as inactive drug molecules i.e. false negative
TP: Active drug molecules are classified as active drug molecules i.e. true positive
TN: Inactive drug molecules are classified as inactive drug molecules i.e. true negative

- F-score:

The F-score is the harmonic mean of precision and recall, which is an alternative way to use precision and recall that measures them into a combined form. This approach is also known as the F-distribution or F-measure. It gives equal weight to precision and recall:

$$F - score = \frac{2 * precision * recall}{precision + recall} \tag{6.8}$$

- Kappa:

Cohens kappa is similar to accuracy measure. Kappa statistic is a very good measure that can handle very well of both multi-class and imbalanced class problems:

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{6.9}$$

where p_o and p_e are the observed agreement and the expected agreement, respectively. Generally, it tells you how much your classifier is performing better over the performance of another classifier that simply guesses at random according to the frequency of each class. Cohens kappa is always less than or equal to 1. The classifier is useless when the value of kappa is 0 [119].

6.6 Result Analysis, Comparison and Validation

The Gini coefficient, sensitivity, specificity, precision, F-score, MCC, kappa, AUC and accuracy are the model performance evaluation parameters for any binary classification

Table 6.6: Performance comparison of proposed ensemble model with existing classification models

Decision method	Gini	Sensitivity	Specificity	Precision	F-score	MCC	Kappa	AUC	Accuracy(%)
Proposed ensemble model	0.990	0.973	0.968	0.984	0.978	0.936	0.943	0.995	**97.14**
Random forest	0.987	0.972	0.938	0.972	0.972	0.910	0.908	0.993	96.19
Decision tree	0.905	0.953	0.921	0.963	0.958	0.869	0.862	0.952	94.29
SVM	0.988	0.942	0.972	0.985	0.963	0.898	0.895	0.994	95.24
AdaBoost	0.990	0.972	0.956	0.979	0.975	0.924	0.920	0.995	96.67

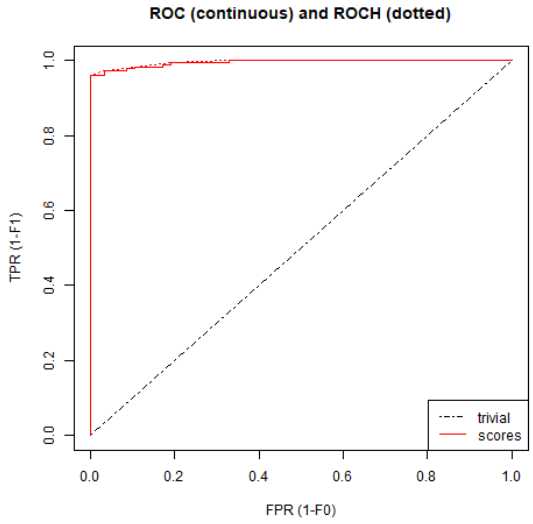


Figure 6.6: ROC performance of multilevel ensemble model on the testing dataset, AUC: 0.995

model, which are described in Section 6.6. Activity prediction is evaluated by these parameters for our proposed ensemble model as well as for some existing models. The Gini coefficient, sensitivity, specificity, precision, F-score, MCC, kappa, AUC, and accuracy of our proposed ensemble model are 0.990, 0.973, 0.968, 0.984, 0.978, 0.936, 0.943, 0.995 and 97.14%, respectively. The comparative performance of some existing classification models and our proposed ensemble model for the prediction of activity is analysed by these parameters as shown in Table 6.6. The results show that our proposed ensemble model has outperformed the other models for the testing dataset of ARE.

6.6.1 K-fold cross-validation

To measure the robustness of the model, the K-fold cross-validation technique shows the stable performance for the accuracy of the proposed ensemble model [30]. Here, we have used 5-fold (K=5) cross-validation for the prediction of activity because our dataset has a total of five data frames. In this case at a time four data frames are used for training

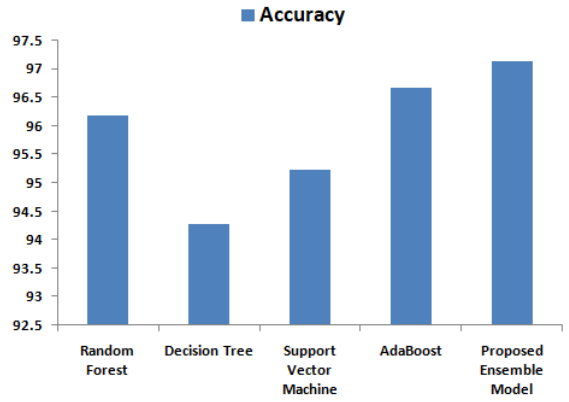


Figure 6.7: Bar chart for the performance comparison of proposed ensemble model with existing models

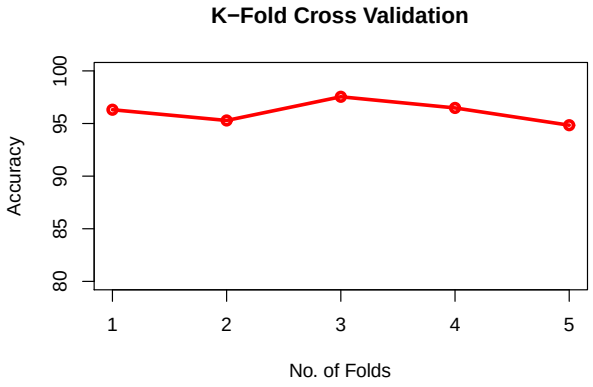


Figure 6.8: K-fold cross-validation for the activity prediction of ARE

and one data frame is used for testing. Table 6.7 describes the accuracy, and Figure 6.8 shows the accuracy in the form of a line graph of the proposed ensemble model in 5-fold cross-validation for the prediction of activity. These results of cross-validation show the consistent performance of the proposed ensemble model on different folds of the dataset [45]

Table 6.7: 5-fold cross validation of proposed ensemble model

Folds	Accuracy
1	96.31
2	95.29
3	97.54
4	96.48
5	94.84

6.6.2 Performance validation of proposed ensemble model

We have validated our proposed ensemble model on two validation datasets. The first dataset contains 12 drug molecules, which are related to general food additives, cosmetics, detergents and preservatives [57]. The second dataset contains eight small drug molecules that induce genotoxicity in human embryonic kidney cells expressing luciferase-tagged ATAD5 (ATAD5) [120]. All these drug molecules are neither part of training dataset nor part of the testing dataset; if the performance of our proposed ensemble model is satisfactory on these new drug molecules then it accomplishes the validation process of our proposed ensemble model.

Food additives such as aspartame, saccharin and monosodium glutamate (MSG) are predicted to be toxic by our proposed ensemble model. Aspartame and saccharin are artificial non-carbohydrate calorie-free sweeteners, and MSG is used for the flavour enhancer.

Pesticides are being widely used to control insects rodents in the agriculture fields. The continued usage of pesticides has highly detrimental environmental impacts on air, water, soil and food and could be toxic to humans. It disrupts your hormones, causes cancer, disturbs reproductive development, and create neurological effects like loss of memory. Pesticides likes dimethyl tetrachloroterephthalate acid (DCPA) and ethylenediaminetetraacetic acid (EDTA) [109][110] are predicted toxic by our proposed ensemble model.

The ingredients of beauty and cosmetic products like butyl hydroxy butyl nitrosamine and sodium tetra decane sulfonate compounds are predicted to be toxic by our proposed model. Imidazolidinyl urea is also used in cosmetics as an antimicrobial preservative due to its high solubility in water, but it is predicted to be non-toxic by our proposed model. Benzethonium chloride is commonly used in cosmetics, medicaments, deodorants and mouthwash because of its antiseptic and antimicrobial properties. It is predicted toxic by our proposed model [116].

Another commonly used synthetic product is polysorbate-80, which is used as an emulsifier in vitamins, vaccines, medicines, surfactant in soaps and cosmetics, defoamer in the fermentation of wine and binding agent in ice cream, which was also predicted to be toxic by our proposed model.

Sodium hypochlorite is commonly known as bleaching powder, which is used in bleaching, surface purification, and disinfection of water. Sodium hypochlorite was predicted to be non-toxic by our proposed model. Asbestos, which is commonly used in construction works because of its thermal insulation and fire protection, is also predicted toxic by our

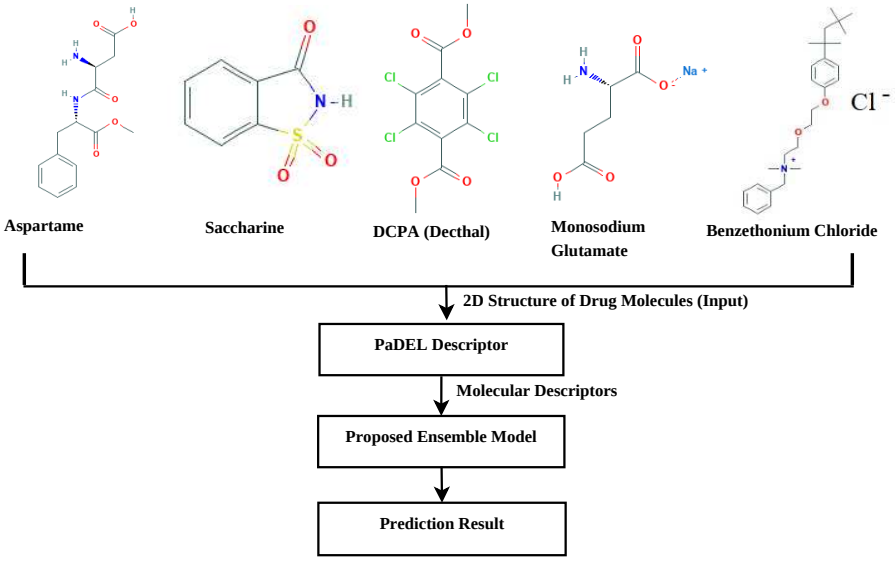


Figure 6.9: Activity prediction of some new drug molecules using the proposed ensemble model for performance validation

proposed model [57].

Now, we have 2D structures of 12 drug molecules, which are related to food additives, cosmetics, detergents, preservatives and eight drug molecules of ATAD5. All these drug molecules are downloaded from the PubChem database 'https://pubchem.ncbi.nlm.nih.gov' in the form of SDF format, and later their molecular descriptors have been calculated by PaDEL-Descriptor software. Then, we applied our proposed ensemble model on aspartame, saccharin, MSG, DCPA, EDTA, butyl hydroxy butyl nitrosamine, sodium tetra decane sulfonate, imidazolidinyl, benzethonium chloride, polysorbate-80, sodium hypochlorite, asbestos, and eight drug molecules of ATAD5 to predict the activity. The output of the proposed ensemble model is summarized in Table 6.8. The results of the table show that the drug molecules that are predicted active are found to be active, and the drug molecules that are predicted inactive are found to be inactive. The correct predictions of all these drug molecules show the validity of the proposed ensemble model. Figure 6.9 shows the validation process of the proposed ensemble model only on some drug molecules of food additives, cosmetics, detergents, and preservatives.

6.7 Discussion

In binary classification, we have divided the given dataset into two categories on the basis of the common characteristics of the data. Statistical performance analysis of any binary classifier is evaluated on the basis of the performance metrics like Gini coefficient, sensi-

Table 6.8: Activity prediction results of some new drug molecules for performance validation

Target drug molecule	**Actual class**	**Predicted class**	**Accuracy(%)**
Aspartame	1	1	100%
Saccharin	1	1	100%
Monosodium glutamate (MSG)	1	1	100%
DCPA(dacthal)	1	1	100%
EDTA	1	1	100%
Butyl hydroxy butyl nitrosamine	1	1	100%
Sodium tetra decane sulfonate	1	1	100%
Imidazolidinyl	0	0	100%
Benzethonium chloride	1	1	100%
Polysorbate-80	1	1	100%
Sodium hypochlorite	0	0	100%
Asbestos	1	1	100%
NCGC00260056-01	1	1	100%
NCGC00160391-01	1	1	100%
NCGC00166123-01	1	1	100%
NCGC00163700-01	1	1	100%
NCGC00094121-01	0	0	100%
NCGC00258875-01	0	0	100%
NCGC00254309-01	0	0	100%
NCGC00256470-01	0	0	100%

tivity, specificity, precision, F-score, MCC, kappa, AUC and accuracy. Here, sensitivity represents the performance of the prediction of positive class (active), while specificity represents the performance of the prediction of negative class (inactive). Accuracy is the most important measurement that predicts how well the model predicts both classes (active and inactive). Therefore, a good binary classifier gives a high value of sensitivity, specificity and accuracy [93].

Our dataset of activity classification is highly imbalanced, where the positive class (active) is rare than the negative class (inactive). In the case of class imbalanced, generally, a classifier correctly classifies the inactive class and incorrectly classifies the active class. Therefore, we should not believe only on accuracy but also use some other performance metrics like sensitivity and specificity [30].

Since the active count is too low and inactive count is too high, then the accuracy varies with specificity without considering sensitivity. Similarly, when the active count is very high and inactive count is very low, then the accuracy tends to vary with sensitivity without considering specificity. Our dataset is very imbalanced, but by the division of dataset in five data frames, which have an almost equal number of active and inactive drug molecules and using ensemble approach, we have nullified the effect of data imbalance. Therefore, our ensemble-based binary classifier gives almost similar results for all these three performance metrics.

6.8 Conclusion

In this paper, we have proposed an ensemble-based efficient computational method, which has solved the problem of toxicity prediction of those drug molecules which can bind to the AREs. The target class for the toxicity prediction is the activity. The dataset used in this study is very high in features and extremely imbalanced in classes. Initially, we have performed feature selection by random forest importance algorithm in R and balanced the dataset by dividing the dataset into five data frames, which have almost an equal number of active and inactive drug molecules. Further, the ensemble method is used which provides a combined result of all data frames; therefore, it resolves the class imbalance issue as well as classifying the drug molecules. After the creation of the proposed ensemble model, we have evaluated this model on various performance parameters, i.e. Gini coefficient, sensitivity, specificity, precision, AUC, F-score, MCC, kappa and accuracy for the activity prediction. Through the intensive experiments, it is found that our proposed ensemble model, in spite of highly imbalanced in classes, has given better accuracy than other existing models, which are decision tree, SVM, AdaBoost, and random forest, 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 some new drug molecules which are neither part of the training dataset nor part of the testing dataset, where we have found 100% accuracy.

Chapter 7

Conclusions and Future Works

This chapter concludes the thesis and proposes some suggestions where the present work can be extended. Section 7.1 brings out an overall conclusion of the research work which are presented in this thesis, and Section 7.2 describes some ideas regarding future research directions and a possible extension of the present work.

7.1 Conclusion

In this thesis, an attempt has been made to improve the toxicity prediction of small drug molecules of androgen receptor (AR), estrogen receptor (ER), aryl hydrocarbon receptor (AhR), and antioxidant response element (ARE) using physicochemical properties and computational intelligence approaches. We have developed the multilevel prediction model to predict the activity, activity score, potency, and efficacy of ER nuclear receptor by using the random forest model. An ensemble-based classification model is developed to predict the activity, and an ensemble-based regression model is developed to predict the activity score, potency, and efficacy of AR nuclear receptor. We have also developed an ensemble-based classification model to predict the activity of AhR nuclear receptor, and another ensemble-based classification model to predict the activity of ARE stress response pathway. The main contribution of this thesis is described in the following manner:

1. We have developed a computational method (in silico) for checking the activity/toxicity of drug molecules rather than inside the living animal (in vivo) or in glass (in vitro) to save the lives of animals and money.

2. This study focused on predicting the activity/toxicity of drug molecules based on machine learning approaches as well as ensemble-based classification.

3. We have developed a stand-alone application(s) that help the researchers to predict the activity, activity score, potency, and efficacy of the newly discovered chemical compounds as well as environmental chemicals, which have the probable chances to disrupt the processes in the human body.

4. We have developed better activity, activity score, potency, and efficacy assessment features, methods, and algorithms for drug molecules of AR, ER, AhR, and ARE.

5. In machine learning models, the most important phase is data collection. In this thesis, active and inactive drug molecules have been collected from one reliable and authenticated resource, i.e. PubChem (refer to Subsection 3.2.2, Subsection 4.2.2, Subsection 5.2.2, and Subsection 6.2.2).

6. The preprocessing of the data, which includes feature extraction, data cleansing, feature selection, and splitting the dataset into a training set and test set, is performed.

7. Molecular descriptors are the encoded form of physicochemical properties of drug molecules. Here, drug molecules are in the structured-data file (SDF) format, which contains all physicochemical properties. We have calculated (extracted) the molecular descriptor of drug molecules of AR, ER, AhR, and ARE by using PaDEL descriptor software, which is a java based free and open-source software. These molecular descriptors are the actual features of any drug molecules (refer to Subsection 3.2.1, Subsection 4.2.1, Subsection 5.2.1, and Subsection 6.2.1).

8. The feature selection process has been performed by using different techniques, which includes CFS, information gain, Boruta, and random forest importance algorithm (refer to Subsection 3.2.4, Subsection 4.2.3, Subsection 5.2.3, and Subsection 6.2.3).

9. The dataset of AR, ER, AhR, and ARE are highly imbalanced in classes. Therefore, we resolved the class imbalance issue of ER dataset using SMOTE algorithm, and for AR, AhR, and ARE, we divided their datasets into multiple data frames, which have the approximately equal number of active and inactive drug molecules. After that, we combined them and classified by using ensemble learning approaches (refer to Subsection 3.2.3, Subsection 4.2.4, Subsection 5.2.4, and Subsection 6.2.4).

10. We have predicted the activity, activity score, potency, and efficacy by using physicochemical properties and machine learning models. In this thesis, random forest, decision tree, support vector machine, linear model, neural network, and AdaBoost model are used separately as well as in the combination as the base classifiers to create ensemble-based classification models (refer to Section 3.4, Section 4.4, 5.4, and Section 6.4).

11. The parameters of machine learning models which are used in our study are tuned for the better prediction of binary as well as continuous classes (refer to Section

3.4, Section 4.4, Section 5.4, and Section 6.4).

12. In this thesis, three ensemble learning approaches are proposed. In each, proposed ensemble learning different machine learning models are used as the base classifiers. Here, the bagging technique is used to develop different ensemble-based classification or regression models (refer to Figure 4.6, Figure 5.5, Figure 5.6, and Figure 6.5).

13. A multilevel prediction model is developed, which has its two levels, on its first level classification of activity of ER is performed, and on its second level regression of activity score, potency and efficacy of ER are performed by using physicochemical properties and various individual machine learning models where random forest model is performed well (refer to Chapter 3).

14. An ensemble-based classification model is proposed for the prediction of the activity of AhR where our proposed ensemble model outperformed the other models (refer to Chapter 4).

15. A multilevel ensemble model is developed, which has its two levels, on its first level classification of activity of AR is performed, and on its second level regression of activity score, potency and efficacy of AR are performed by using physicochemical properties and ensemble-based machine learning model. Our proposed multilevel ensemble model outperformed the other models (refer to Chapter 5).

16. One another ensemble-based classification model is proposed for the prediction of the activity ARE where our proposed ensemble model outperformed the other models (refer to Chapter 6).

17. We have focused on accuracy as the most important criteria for the prediction of binary class activity. Apart from it, other metrics like Gini coefficient, sensitivity, specificity, precision, kappa, AUC, F-score, and MCC are also considered (refer to Subsection 3.5.1, Section 4.5, Subsection 5.5.1, and Section 6.5).

18. We have focused on accuracy as the most important criteria for the prediction of continuous class activity score, potency, and efficacy. Apart from it, other metrics like root mean square error (RMSE), correlation (r), and coefficient of determination (R^2) are also considered (refer to Subsection 3.5.2, and Subsection 5.5.2).

19. The consistency of prediction of the proposed multilevel prediction model for ER, proposed multilevel ensemble model for AR, and proposed ensemble model for AhR and ARE has been validated by performing K-fold cross-validation (refer to Subsection 3.5.3, Subsection 4.6.1, Subsection 5.6.1, and Subsection 6.6.1).

20. The performance of our proposed multilevel prediction model for ER is compared with various standard classifiers like SVM, decision tree, neural network, and linear model where the random forest model outperformed the others (refer to Table 3.1.3, and Table 3.1.4).

21. The performance of our proposed ensemble-based models for AhR, AR and ARE is compared with various standard classifiers like SVM, decision tree, random forest, neural network, linear model, and AdaBoost, where our proposed ensemble model outperformed the others (refer to Table 4.6, Table 5.13, Table 5.15, and Table 6.6).

22. We have validated all our proposed models on some other drug molecules, which are neither part of training dataset nor part of the testing dataset. These drug molecules are of AIDS therapies, food additives, cosmetics, detergents, preservatives, androgen receptor, estrogen receptor, and ATAD5 (refer to Subsection 3.6.1, Subsection 4.6.2, Subsection 5.6.2, and Subsection 6.6.2).

7.2 Scope for Future Work

Research is an iterative and continuous procedure. The work presented in this thesis focused on the toxicity prediction of pre-clinical trial drugs using physicochemical properties of drug molecules and machine learning approaches. There are several directions in which this work can be extended. Some of the suggestions for future work are:

1. The proposed ensemble-based approaches can be applied to perform beneficial roles in different biological areas. These roles include prediction of chronic diseases, protein structure prediction, allergy prediction, cancer prediction, diabetes prediction, and activity prediction of those drug molecules which can bind to some other nuclear receptors or stress response pathways.

2. The toxicity prediction is performed using 1444 features. More features, and some better feature selection algorithm need to be explored for more accurate prediction.

3. In this thesis, six machine learning models are used for the toxicity prediction of small drug molecules. Various new machine learning models are also available; if we explore these models, then they may provide more accurate and fast predictions.

4. For processing the high volume of data, we are suggesting to calibrate this framework by implementing it on the top of modern big data techniques like Hadoop, HBase, Spark, and so on.

5. We believe that by utilizing some other data imbalance handling methods and different kinds of ensemble learning with their optimized parameters may achieve better performance of classifiers.

6. Boosting is another kind of ensemble learning approach, which can provide some better predictions.

List of Publications

1. Vishan Kumar Gupta and Prashant Singh Rana, *"Activity Assessment of Small Molecules in Estrogen Receptor using Multilevel Prediction Model"*, IET System Biology, 13(3): 147-158, 2019. [SCIE Indexed, Impact Factor 1.392]

2. Vishan Kumar Gupta and Prashant Singh Rana, *"Toxicity Prediction of Small Drug Molecules of Aryl Hydrocarbon Receptor using Proposed Ensemble Model"*, Turkish Journal of Electrical Engineering and Computer Science, 27(4): 2833-2849, 2019. [SCIE Indexed, Impact Factor 0.625]

3. Vishan Kumar Gupta and Prashant Singh Rana, *"Toxicity Prediction of Small Drug Molecules of Androgen Receptor using Multilevel Ensemble Model"*, Journal of Bioinformatics and Computational Biology, World Scientific Publishers, 17(5): 1950033, 2019. [SCIE Indexed, Impact Factor 0.991]

4. Vishan Kumar Gupta and Prashant Singh Rana, *"Ensemble Technique for Toxicity Prediction of Small Drug Molecules of the Antioxidant Response Element Signalling Pathway"*, The Computer Journal, Oxford University Press, 2020. [SCIE Indexed, Impact Factor 0.98, In press]

References

[1] Chemoinformatics - a quick review - scientific figure on researchgate, https://www.researchgate.net/figure/drug-discovery-and-development-process_fig2_252015829 (accessed on june 17, 2019).

[2] Chun Wei Yap. Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, 32(7):1466–1474, 2011.

[3] Tommy Liljefors, Povl Krogsgaard-Larsen, and Ulf Madsen. *Textbook of drug design and discovery*. CRC Press, 2002.

[4] Joseph A DiMasi, Ronald W Hansen, and Henry G Grabowski. The price of innovation: new estimates of drug development costs. *Journal of health economics*, 22(2):151–185, 2003.

[5] The drug development process, https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process (accessed on june 17, 2019).

[6] Ellen K. Silbergeld. *Toxicology, Encyclopedia of occupational health and safety*. 4th edition.

[7] Roberto Todeschini and Viviana Consonni. *Molecular descriptors for chemoinformatics*, volume 41. John Wiley & Sons, New Jersey, United States, 2009.

[8] Graham F Smith. Designing drugs to avoid toxicity. In *Progress in medicinal chemistry*, volume 50, pages 1–47. Elsevier, 2011.

[9] James Inglese and Douglas S Auld. High throughput screening (hts) techniques: applications in chemical biology. *Wiley Encyclopedia of Chemical Biology*, pages 1–15, 2007.

[10] Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature reviews Drug discovery*, 3(8):711, 2004.

[11] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. Deeptox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80, 2016.

[12] SC Rastogi, Parag Rastogi, and Namita Mendiratta. *Bioinformatics Methods And Applications: Genomics Proteomics And Drug Discovery*. PHI Learning Pvt. Ltd., 3rd edition, 2008.

[13] Vishan Kumar Gupta and Prashant Singh Rana. Activity assessment of small drug molecules in estrogen receptor using multilevel prediction model. *IET systems biology*, 13(3):147–158, 2019.

[14] Azhwar Raghunath, Kiruthika Sundarraj, Raju Nagarajan, Frank Arfuso, Bian Jinsong, Alan P Kumar, Gautam Sethi, and Ekambaram Perumal. Antioxidant response elements: discovery, classes, regulation and potential applications. *Redox biology*, 17:297–314, 2018.

[15] TG Dietterich. Ensemble learning. the handbook of brain theory and neural networks. *Arbib MA*, 2002.

[16] Robert E Schapire. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer, 2003.

[17] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[18] Kun Chang Lee and Heeryon Cho. Performance of ensemble classifier for location prediction task: emphasis on markov blanket perspective. *International Journal of u-and e-Service, Science and Technology*, 3(3), 2010.

[19] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006.

[20] Thomas G Dietterich. Machine-learning research. *AI magazine*, 18(4):97–97, 1997.

[21] David H Wolpert, William G Macready, et al. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

[22] Cha Zhang and Yunqian Ma. *Ensemble machine learning: methods and applications*. Springer, 2012.

[23] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[24] Robert E Schapire, Yoav Freund, Peter Bartlett, Wee Sun Lee, et al. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.

[25] Magdalena Graczyk, Tadeusz Lasota, Bogdan Trawiński, and Krzysztof Trawiński. Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal. In *Asian conference on intelligent information and database systems*, pages 340–350. Springer, 2010.

[26] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education, India, 2018.

[27] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2009.

[28] Hiu C Law. Clustering, dimensionality reduction, and side information. Technical report, Michigan State Univ East Lansing Dept of Computer Science and Engineering, 2006.

[29] How to handle imbalanced classification problems in machine learning, https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem (accessed on mar 16, 2017).

[30] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, USA, 2011.

[31] Linear model, https://www.mathworks.com/discovery/linear-model.html (accessed on july 07, 2019).

[32] Divya Khanna and Prashant Singh Rana. Multilevel ensemble model for prediction of iga and igg antibodies. *Immunology letters*, 184:51–60, 2017.

[33] Tox21, 'national institute of health - toxicology in the 21st century', https://ncatsncats.nih.gov/tox21 (accessed on mar 7, 2017).

[34] Tox21 data challenge 2014, https://tripod.nih.gov/tox21/challenge/data.jsp (accessed on mar 7, 2017).

[35] Adolf Grauel, LA Ludwig, I Renners, and F Berk. Computational intelligence and predictive toxicology. In *Proc. AAAI*, 1999.

[36] Kunal Roy, Supratik Kar, and Rudra Narayan Das. *A primer on QSAR/QSPR modeling: fundamental concepts*. Springer, 2015.

[37] Saeed Yousefinejad and Bahram Hemmateenejad. Chemometrics tools in qsar/qspr studies: A historical perspective. *Chemometrics and Intelligent Laboratory Systems*, 149:177–204, 2015.

[38] James Inglese, Douglas S Auld, Ajit Jadhav, Ronald L Johnson, Anton Simeonov, Adam Yasgar, Wei Zheng, and Christopher P Austin. Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proceedings of the National Academy of Sciences*, 103(31):11473–11478, 2006.

[39] Thomas Unterthiner, Andreas Mayr, Günter Klambauer, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, and Sepp Hochreiter. Deep learning as an opportunity in virtual screening. In *Proceedings of the deep learning workshop at NIPS*, volume 27, pages 1–9, 2014.

[40] Jerrold Tannenbaum and B Taylor Bennett. 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*, 54(2):120–132, 2015.

[41] Christoph Helma. In silico predictive toxicology: the state-of-the-art and strategies to predict human health effects. *Current opinion in drug discovery & development*, 8(1):27–31, 2005.

[42] Subhash C Basak, Denise Mills, Moiz M Mumtaz, and Krishnan Balasubramanian. Use of topological indices in predicting aryl hydrocarbon receptor binding potency of dibenzofurans: A hierarchical qsar approach. 2003.

[43] Elena Lo Piparo, Konrad Koehler, Antonio Chana, and Emilio Benfenati. Virtual screening for aryl hydrocarbon receptor binding prediction. *Journal of medicinal*

chemistry, 49(19):5702–5709, 2006.

[44] Antonio Cassano, Alberto Manganaro, Todd Martin, Douglas Young, Nadège Piclin, Marco Pintore, Davide Bigoni, and Emilio Benfenati. Caesar models for developmental toxicity. In *Chemistry Central Journal*, volume 4, page S4. Springer, 2010.

[45] Stephen J Capuzzi, Regina Politi, Olexandr Isayev, Sherif Farag, and Alexander Tropsha. Qsar modeling of tox21 challenge stress response and nuclear receptor signaling toxicity assays. *Frontiers in Environmental Science*, 4:3, 2016.

[46] Mark TD Cronin, Steven J Enoch, Mark Hewitt, and Judith C Madden. Formation of mechanistic categories and local models to facilitate the prediction of toxicity. *ALTEX-Alternatives to animal experimentation*, 28(1):45–49, 2011.

[47] Malgorzata Natalia Drwal, Vishal Babu Siramshetty, Priyanka Banerjee, Andrean Goede, Robert Preissner, and Mathias Dunkel. Molecular similarity-based predictions of the tox21 screening outcome. *Frontiers in Environmental science*, 3:54, 2015.

[48] Filip Stefaniak. 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*, 3:77, 2015.

[49] Huanxiang Liu, Fang Bai, Ding Hong, Yingying Lu, Cunlu Xu, and Xiaojun Yao. Prediction of the antioxidant response elements' response of compound by deep learning. *Frontiers in Chemistry*, 7:385, 2019.

[50] Mahmud Tareq Hassan Khan. Predictions of the admet properties of candidate drug molecules utilizing different qsar/qspr modelling approaches. *Current drug metabolism*, 11(4):285–295, 2010.

[51] YZ Chen and CY Ung. 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*, 20(3):199–218, 2001.

[52] Rafał Adamczak and Włodzisław Duch. Neural networks for structure-activity relationship problems. In *5th Conference on Neural Networks and Soft Computing, Zakopane*, pages 669–674, 2000.

[53] Hu Li, CW Yap, CY Ung, Y Xue, ZR Li, LY Han, HH Lin, and Yu Zong Chen. Machine learning approaches for predicting compounds that interact with therapeutic and admet related proteins. *Journal of pharmaceutical sciences*, 96(11):2838–2860, 2007.

[54] Jinjian Jiang, Nian Wang, Peng Chen, Jun Zhang, and Bing Wang. Drugecs: An ensemble system with feature subspaces for accurate drug-target interaction prediction. *BioMed Research International*, 2017, 2017.

[55] Limeng Pu, Misagh Naderi, Tairan Liu, Hsiao-Chun Wu, Supratik Mukhopadhyay, and Michal Brylinski. etoxpred: a machine learning-based approach to estimate the toxicity of drug candidates. *BMC Pharmacology and Toxicology*, 20(1):2, 2019.

[56] Yunyi Wu and Guanyu Wang. Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. *International journal of molecular sciences*, 19(8):2358, 2018.

[57] Ashok K Sharma, Gopal N Srivastava, Ankita Roy, and Vineet K Sharma. Toxim: A toxicity prediction tool for small molecules developed using machine learning and chemoinformatics approaches. *Frontiers in pharmacology*, 8:880, 2017.

[58] George E Dahl, Navdeep Jaitly, and Ruslan Salakhutdinov. Multi-task neural networks for qsar predictions. *arXiv preprint arXiv:1406.1231*, 2014.

[59] David Hecht and Gary B Fogel. Computational intelligence methods for admet prediction. *Front Drug Des Discov*, 4:351–377, 2009.

[60] Ciprian-Daniel Neagu, Emilio Benfenati, Giuseppina Gini, Paolo Mazzatorta, and Alessandra Roncaglioni. Neuro-fuzzy knowledge representation for toxicity prediction of organic compounds. In *ECAI*, pages 498–502, 2002.

[61] Omar Deeb and Mohammad Goodarzi. In silico quantitative structure toxicity relationship of chemical compounds: Some case studies. *Current drug safety*, 7(4):289–297, 2012.

[62] Steven W Baertschi, Karen M Alsante, and Robert A Reed. *Pharmaceutical stress testing: predicting drug degradation*. CRC Press, 2016.

[63] Distributed strucure-searchable toxicity (dsstox) database, https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database (accessed on july 18, 2019).

[64] Richard K Harrison. Phase ii and phase iii failures: 2013–2015. *Nature Review Drug Discovery*, 15(12):817–8, 2016.

[65] Dorian Pyle. *Data preparation for data mining*. morgan kaufmann, 1999.

[66] Oded Maimon and Lior Rokach. Data mining and knowledge discovery handbook. 2005.

[67] Rui J Almeida, Uzay Kaymak, and Joao MC Sousa. A new approach to dealing with missing values in data-driven fuzzy modeling. In *International Conference on Fuzzy Systems*, pages 1–7. IEEE, 2010.

[68] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *ACM Sigmod Record*, volume 29, pages 427–438. ACM, 2000.

[69] AY Chung Liu. The effect of oversampling and undersampling on classifying imbalanced text datasets. *The University of Texas at Austin*, 2004.

[70] Nitesh V Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 875–886. Springer, 2009.

[71] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[72] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[73] Estrogen receptor, national center for biotechnology information. pubchem database. aid=743079, https://pubchem.ncbi.nlm.nih.gov/bioassay/743079 (accessed oct 12, 2017).

[74] Jerome C Nwachukwu, Sathish Srinivasan, Yangfan Zheng, Song Wang, Jian Min, Chune Dong, Zongquan Liao, Jason Nowak, Nicholas J Wright, René Houtman, et al. Predictive features of ligand-specific signaling through the estrogen receptor. *Molecular systems biology*, 12(4):864, 2016.

[75] Erin K Shanle and Wei Xu. Endocrine disrupting chemicals targeting estrogen receptor signaling: identification and mechanisms of action. *Chemical research in toxicology*, 24(1):6–19, 2010.

[76] Drug activity prediction, https://www.kaggle.com/c/drugactivityprediction (accessed on 22 april 2016).

[77] Arja H Asikainen, Juhani Ruuskanen, and Kari A Tuppurainen. Consensus knn qsar: a versatile method for predicting the estrogenic activity of organic compounds in silico. a comparative study with five estrogen receptors and a large, diverse set of ligands. *Environmental science & technology*, 38(24):6724–6729, 2004.

[78] Andrey A Toropov, Alla P Toropova, Ivan Raska Jr, Danuta Leszczynska, and Jerzy Leszczynski. Comprehension of drug toxicity: Software and databases. *Computers in biology and medicine*, 45:20–25, 2014.

[79] Nishtha Hooda, Seema Bawa, and Prashant Singh Rana. B2fse framework for high dimensional imbalanced data: A case study for drug toxicity prediction. *Neurocomputing*, 276:31–41, 2018.

[80] Piotr Romanski and Lars Kotthoff. Fselector: Selecting attributes; 2018. r package version 3.4.0.

[81] Asha Gowda Karegowda, AS Manjunath, and MA Jayaram. Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271–277, 2010.

[82] Prashant Singh Rana, Harish Sharma, Mahua Bhattacharya, and Anupam Shukla.

Quality assessment of modeled protein structure using physicochemical properties. *Journal of bioinformatics and computational biology*, 13(02):1550005, 2015.

[83] Bonnie Fremgen. *Probability & Statistics for Engineers & Scientists*. Prentice Hall, 2011.

[84] Iris Usach, Virginia Melis, and José-Esteban Peris. Non-nucleoside reverse transcriptase inhibitors: a review on pharmacokinetics, pharmacodynamics, safety and tolerability. *Journal of the International AIDS Society*, 16(1):18567, 2013.

[85] Etravirine - national center for biotechnology information. pubchem compound database', https://pubchem.ncbi.nlm.nih.gov/compound/193962, (accessed on sep 02, 2018).

[86] Rilpivirine - national center for biotechnology information. pubchem compound database', https://pubchem.ncbi.nlm.nih.gov/compound/6451164 (accessed on sep 02, 2018).

[87] Lersivirine - national center for biotechnology information. pubchem compound database', https://pubchem.ncbi.nlm.nih.gov/compound/16739244 (accessed on sep 02, 2018).

[88] Androgen receptor, national center for biotechnology information. pubchem database. aid=743040, https://pubchem.ncbi.nlm.nih.gov/bioassay/743040 (accessed oct 12, 2017).

[89] randomforest - the r package for statistical computing, https://cran.r-project.org/web/packages/randomforest/randomforest.pdf (accessed on oct 12, 2017).

[90] rpart - the r package for statistical computing, https://cran.r-project.org/web/packages/rpart/rpart.pdf (accessed on oct 12, 2017).

[91] nnet - r package, https://cran.r-project.org/web/packages/nnet/nnet.pdf (accessed on oct 12, 2017).

[92] kernlab - the r package for statistical computing, https://cran.r-project.org/web/packages/kernlab/kernlab.pdf (accessed on oct 12, 2017).

[93] S.N Dr. Achuthsankar and B Aswathi. Sensitivity, specificity, accuracy and the relationship between them, available online: http://www.lifenscience.com/bioinformatics/ (accessed on jan 7, 2019), 2018.

[94] Brigitta Stockinger. Beyond toxicity: Aryl hydrocarbon receptor-mediated functions in the immune system. *Journal of biology*, 8(7):61, 2009.

[95] Miron B Kursa, Witold R Rudnicki, et al. Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010.

[96] Wei Feng, Wenjiang Huang, and Jinchang Ren. Class imbalance ensemble learning based on the margin theory. *Applied Sciences*, 8(5):815, 2018.

[97] Anitha Arumugam. A predictive modeling approach for improving paddy crop productivity using data mining techniques. *Turkish Journal of Electrical Engineering & Computer Sciences*, 25(6):4777–4787, 2017.

[98] Hidayet Takci. Improvement of heart attack prediction by the feature selection methods. *Turkish Journal of Electrical Engineering & Computer Sciences*, 26(1):1–10, 2018.

[99] Yoshio Takei, Hironori Ando, and Kazuyoshi Tsutsui. *Handbook of hormones: comparative endocrinology for basic and clinical research*. Academic Press, 2015.

[100] Doug C Luccio-Camelo and Gail S Prins. Disruption of androgen receptor signaling in males by environmental chemicals. *The Journal of steroid biochemistry and molecular biology*, 127(1-2):74–82, 2011.

[101] Tox21 data challenge 2014, national center for advancing translational sciences (ncats), https://tripod.nih.gov/tox21/challenge/about.jsp (accessed on mar 7, 2017).

[102] BioTech-FYI-Center. Chemical file format - structure data format(sdf), http://biotech.fyicenter.com/resource/sdf_format.html (accessed on jan 02, 2018).

[103] informatio.gain(), 'fselector package', http://cran.r-project.org/web/packages/fselector/fselector.pdf (accessed on may 25, 2017).

[104] Ensemble methods: Elegant techniques to produce improved machine learning results, https://www.toptal.com/machine-learning/ensemble-methods-machine-learning (accessed on jan 22, 2018).

[105] Vishan Kumar GUPTA and Prashant Singh RANA. Toxicity prediction of small drug molecules of aryl hydrocarbon receptor using a proposed ensemble model. *Turkish Journal of Electrical Engineering & Computer Sciences*, 24(4):2833–2849, 2019.

[106] Mark H Zweig and Gregory Campbell. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.

[107] Yadunath Pathak, Prashant Singh Rana, PK Singh, and Mukesh Saraswat. Protein structure prediction (rmsd<=å) using machine learning models. *IJDMB*, 14(1):71–85, 2016.

[108] Sharon L. Myers Keying Ye Ronald E. Walpole, Raymond H. Myers. *Probability & Statistics for Engineers & Scientists*. Pearson India Education Services Pvt. Ltd., India, 2016.

[109] Caroline Cox. Dcpa (dacthal). *Journal of pesticide reform: a publication of the Northwest Coalition for Alternatives to Pesticides*, 11(3):17–20, 1991.

[110] C Barton. Edta (ethylenediaminetetraacetic acid). *Encyclopedia of Toxicology*,

2:147–148, 2014.

[111] Zinc database, https://zinc.docking.org/substances/home/ (accessed on nov. 17, 2019).

[112] Pubchem database, https://pubchem.ncbi.nlm.nih.gov/ (accessed on nov. 17, 2019).

[113] Qiang Ma. Role of nrf2 in oxidative stress and toxicity. *Annual review of pharmacology and toxicology*, 53:401–426, 2013.

[114] Sekhar P Reddy. The antioxidant response element and oxidative stress modifiers in airway diseases. *Current molecular medicine*, 8(5):376–383, 2008.

[115] Antioxident response element/nrf2, national center for biotechnology information. pubchem database. aid=743219, https://pubchem.ncbi.nlm.nih.gov/assay/743219 (accessed on apr. 19, 2019).

[116] Vishan Kumar Gupta and Prashant Singh Rana. Toxicity prediction of small drug molecules of androgen receptor using multilevel ensemble model. *Journal of Bioinformatics and Computational Biology*, 17(5):1950033, 2019.

[117] The r package 'ada' for stochastic boosting, https://cran.r-project.org/web/packages/ada/ada.pdf (accessed on oct 12, 2017).

[118] Matthews correlation coefficient, https://www.vcalc.com/wiki/emilyb/ (accessed on oct 12, 2018).

[119] Cohen's kappa, https://thedatascientist.com/performance-measures-cohens-kappa-statistic/ (accessed on sep 24 2018).

[120] Genotoxicity in human embryonic kidney cells expressing luciferase-tagged atad5, national center for biotechnology information. pubchem database. aid=720516, https://pubchem.ncbi.nlm.nih.gov/assay/720516 (accessed on apr. 19, 2019).