DEVELOPING COMPUTATIONAL TOXICOLOGY MODELS TO ASSESS THE RISK OF ENVIRONMENTAL POLLUTANTS ON ECOSYSTEM HEALTH

A PROJECT REPORT

Submitted by

Mutyala Sushma Chowdary
(Reg. No. CH.SC.U4AIE23034)

Garapati Mohitha
(Reg. No. CH.SC.U4AIE23016)

Mopuri Rishitha
(Reg. No. CH.SC.U4AIE23030)

Madhumita K
(Reg. No. CH.SC.U4AIE23027)

Dharshini R
(Reg. No. CH.SC.U4AIE23043)

Ridhi Verma
(Reg. No. CH.SC.U4AIE23046)

In partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING

Under the guidance of

Dr. I R Oviya

Submitted to

AMRITA
VISHWA VIDYAPEETHAM
श्रद्धावान् लभते ज्ञानम्
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF UGC ACT, 1956

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
AMRITA SCHOOL OF COMPUTING
AMRITA VISHWA VIDYAPEETHAM
CHENNAI - 601103
APRIL 2025

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF UGC ACT 1956
SCHOOL OF COMPUTING

BONAFIDE CERTIFICATE

This is to certify that this project report entitled **“DEVELOPING COMPUTATIONAL TOXICOLOGY MODELS TO ASSESS THE RISK OF ENVIRONMENTAL POLLUTANTS ON ECOSYSTEM HEALTH”** is the bonafide work of **“ M Sushma Chowdary (Reg. No. CH.SC.U4AIE23034), G Mohitha (Reg. No. CH.SC.U4AIE23016), M Rishitha (Reg. No. CH.SC.U4AIE23030), Madhumita K (Reg. No. CH.SC.U4AIE23027), Dharshini R (Reg. No. CH.SC.U4AIE23043), Ridhi Verma (Reg. No. CH.SC.U4AIE23046)”** who carried out the project work under my supervision as a part of End semester project for the course 22BIO211 - Intelligence of Biological Systems 2 .

SIGNATURE

Name Signature

Dr. I R Oviya
Assistant Professor (Sr.Gr.)
Department of Computer Science and Engineering
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Chennai Campus

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF UGC ACT, 1956
SCHOOL OF COMPUTING

DECLARATION BY THE CANDIDATE

We declare that the report entitled **“DEVELOPING COMPUTATIONAL TOXICOLOGY MODELS TO ASSESS THE RISK OF ENVIRONMENTAL POLLUTANTS ON ECOSYSTEM HEALTH”** submitted by us for the degree of Bachelor of Technology is the record of the project work carried out by us as a part of the End Semester project for the course 22BIO211 - Intelligence of Biological Systems 2 under the guidance of **Dr. I R Oviya**. This work has not formed the basis for the award of any course project, degree, diploma, associateship, fellowship, or title in this or any other university or similar institution. We also declare that this project will not be submitted elsewhere for academic purposes.

S.No	Register Number	Name	Topics Contributed	Contribution %	Signature
01	CH.SC.U4AIE23034	M Sushma	SVM Model	16.6%	
02	CH.SC.U4AIE23016	G Mohitha	Hybrid Model (RF + KNN)	16.6%	
03	CH.SC.U4AIE23030	M Rishitha	KNN Model	16.6%	
04	CH.SC.U4AIE23027	Madhumita K	Random Forest	16.6%	
05	CH.SC.U4AIE23043	Dharshini R	Hybrid Model (SVM + RF)	16.6%	
06	CH.SC.U4AIE23046	Ridhi Verma	SNN	16.6%	

SIGNATURES

M Sushma Chowdary
(Reg. No. CH.SC.U4AIE23034)

G Mohitha
(Reg. No. CH.SC.U4AIE23016)

M Rishitha
(Reg. No. CH.SC.U4AIE23030)

Madhumita K
(Reg. No. CH.SC.U4AIE23027)

Dharshini R
(Reg. No. CH.SC.U4AIE23043)

Ridhi Verma
(Reg. No. CH.SC.U4AIE23046)

ACKNOWLEDGEMENT

This project work would not have been possible without the contribution of many people. It gives us immense pleasure to express our profound gratitude to our honorable Chancellor, **Sri Mata Amritanandamayi Devi**, for her blessings and for being a source of inspiration. We are indebted to extend our gratitude to our Director, **Mr. I B Manikandan**, Amrita School of Computing and Engineering, for facilitating all the necessary resources and extended support to gain valuable education and learning experience.

We register our special thanks to **Dr. V. Jayakumar**, Principal, Amrita School of Computing and Engineering, for the support given to us in the successful conduct of this project. We would like to express our sincere gratitude to **Dr. I R Oviya**, Assistant Professor (Sr.Gr.), Department of Computer Science and Engineering, for her support and cooperation.

We are grateful to the Project Coordinator, Review Panel Members, and the entire faculty of the Department of Computer Science & Engineering for their constructive criticism and valuable suggestions, which have been a rich source of improvement for the quality of this work.

M Sushma Chowdary
(Reg. No. CH.SC.U4AIE23034)

G Mohitha
(Reg. No. CH.SC.U4AIE23016)

M Rishitha
(Reg. No. CH.SC.U4AIE23030)

Madhumita K
(Reg. No. CH.SC.U4AIE23027)

Dharshini R
(Reg. No. CH.SC.U4AIE23043)

Ridhi Verma
(Reg. No. CH.SC.U4AIE23046)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

SVM	Support Vector Machine
KNN	K Nearest Neighbors
SNN	Self Organizing Neural Networks
AUC-ROC	Area Under the Receiver Operating Characteristic
SRF	Synthetic Risk Factor
RQ	Risk Quotient
SMILE	Simplified Molecular Input Line Entry System
RBF Kernel	Radial Basis Function Kernel
SHAP	SHapley Additive exPlanations

ABSTRACT

Biodiversity, Aquatic Ecosystems, and Plant Health are negatively affected by various pollutants, which include both chemical agents and heavy metals. This at higher doses disturb physiological functions and have negative effects on both growth and reproduction, or at worst even induce plant mortality. Classic toxicity tests that were established to test the effects of contaminants on environmental matrices are time-consuming, expensive, and only capable of estimating a few long-term or delayed environmental effects. An upcoming possibility is using machine-learning algorithms to make chemical-structure-based predictions of toxicity to provide an answer to these limitations. Initially, our vision was to utilize Machine Learning Models, like Self-Organizing Neural Networks (SNN), k-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest model, in predicting the toxicity of chemical pollutants. The independent accuracies of each model came up to SVM (53%), KNN (49%), Random Forest (50%), and SNN (43%), while a hybrid model made up of SNN, KNN, and SVM achieved the highest accuracy of 94%. These models were trained and validated using practical toxicity datasets to assess their predictive capabilities. A comparative analysis of the algorithms identified the most reliable model for toxicity prediction. The results demonstrate that machine learning algorithms can effectively generate fast and accurate toxicity assessments, which may be instrumental in controlling environmental risks and regulating pollutants.

Keywords: Toxicity, Neural Networks, Support Vector Machine, Random Forest, K-Nearest Neighbor, Pollutants

CHAPTER 1

INTRODUCTION

Environment pollutants pose a serious threat to the environment health primarily, The destruction of biodiversity, the pollution of water, and the destabilization of systems. Human activities, especially, industries have led to the increase in dumping of harmful chemicals into the environment which require effective assessment and mitigation strategies. Computational Toxicology hence becomes a stronger approach for the assessment of hazards in the environment with predictions for results on ecosystems. Artificial Intelligence and Machine Learning approaches have proven to have wonderful potential for enhancing toxicity prediction. A huge dataset and a strong rating model are associated with these methods to predict toxicity endpoints. They offer quick and cost-effective alternatives in both terms of complexity, For example, the US Environmental Protection Agency (EPA) has been at the forefront of computational toxicology implementation creating high-throughput decision support systems for screening and assessing chemical exposure to hazard and risk [1]. These technologies illustrate how computational methods can handle the challenges of chemical toxicity.

The recent developments in the Machine Learning Toxicity Prediction have shown that these data-driven approaches can improve the precision of risk assessments. Such advancements will allow scientists to focus directly on a specific toxicity endpoint, which would lead to improved chemical pollution management and better-tailored therapies. Future opportunities may also be opened by AI technologies integrated into chemical management systems. Any chemical risk management is best done through scalable systems that fast-track toxicity predictions [2]. Environmental toxins have environmental impacts, but they also interfere with the transmission of diseases through groups of consumers, thereby impacting public health in ripple effects [3]. Knowledge of these linkages is crucial for building integrated approaches for risk assessment that consider environmental and human health concerns. Furthermore, recent developments have indicated the need to assess the toxicity of environmental pollutants, especially in compound exposures with mixtures of chemicals. The conventional modes of defining such classification typically do not respond to the cumulative effects arising from many pollutants. These models, however, provide a better approach to overcome these limitations and offer tools that estimates the toxicity of mixtures of chemicals more accurately. [4].

CHAPTER 2

LITERATURE SURVEY

Artificial intelligence (AI) is being used more and more to forecast the toxicity of environmental chemicals using methods such as quantitative structure–activity relationship (QSAR) models. Studies have shown that molecular fingerprints and descriptors, and machine learning methods such as Random Forest and deep learning, improve predictive power [5]. In addition, high-throughput screening technologies like ToxCast and Tox21 offer quick and affordable alternatives to traditional toxicity testing.However, for all their advantages, AI models themselves have issues such as data imbalance and limited applicability in real-world scenarios. The impacts of environmental pollutants have also been subject to further investigation, particularly that on the outcomes of nano-ZnO and polyethylene microplastics combined exposure in the mosquito fish. Excessive bioaccumulation as well as oxidative stress were indicated, highlighting pollutant synergic effects and regulatory implications for environment safety [6]. However, limited exposure time, focus on one species, and lack of field validation imply further research is required. In an effort to increase environmental risk assessment methods, the Synthetic Risk Factor (SRF) methodology was established through an integration of pollutant persistence to provide a superior evaluation than using Risk Quotient (RQ) techniques.

Although SRF is shown to possess robust predictive powers in aquatic habitats, its use of existing information introduces uncertainties and its extension to other environmental habitats needs to be confirmed [7]. Computational toxicology, which is a field that is continuously evolving, utilizes machine learning and deep learning to estimate contaminant toxicity and ecosystem effects. Ensemble learning and molecular fingerprints have been highlighted as valuable in reproductive toxicity prediction, specifically in the ability to deal with high-dimensional information [8]. Transfer learning also came into picture as a rising star, offering the potential of filling the gaps between computational predictions and actual real-world drug toxicity prediction [9]. Deep learning has applications other than chemical toxicity, including measurements of biological systems. Convolutional neural networks have also been used in the case of automated inflammatory response assessment in pollutant-exposed zebrafish as a non-invasive and cost-saving tool for toxicity prediction [10]. Furthermore, these AI-based approaches have aided in the designing of new tyrosine kinase inhibitors and have predicted their bioactivity simultaneously. Knowledge graphs incorporated with large language models have greatly transformed hazardous chemical data management.

These technologies situate complicated data in context for chemical safety evaluation [11]. Another novel computational method is the application of graph-based isomorphism networks to predict peptide toxicity. A model that can evaluate peptide toxicity from sequences and structures has been created, which could be generalized to pollution research with other agents [12]. Beyond small molecules, deep learning is revolutionizing pharmacokinetics and toxicity prediction. A method called Deep-PK has been proposed, increasing interpretability through the recovery of meaningful chemical features from neural networks [13]. In a similar vein, progress in geometric graph learning methods has helped to evaluate nanomaterials' toxicity, assisting with safer material design for environmental use [14]. Nanotoxicology continues to be an important area of interest for computational models. Machine learning algorithms have been used to forecast the dynamic toxic impacts of engineered nanomaterials, showing their applicability in environmental risk assessment [15]. Moreover, in silico methods have been used to forecast the toxicity of new psychoactive substances, yielding important information on both the ecological and forensic relevance of these chemicals [4]. Through these advancements, computational toxicology with AI remains to optimize environmental risk assessments through prediction accuracy while decreasing traditional testing reliance. Issues of data limitations, real-world validity, and generalization remain, requiring further studies and model validation for validating their predictions on environmental impact.

CHAPTER 3

DATASET

The Tox21 dataset is regarded as an essential resource in the field of computational toxicology, and it is used in this study. This collection provides molecular data through SMILES (Simplified Molecular Input Line Entry System) strings, which efficiently represent the chemical structures of diverse molecules. Even further, the dataset consists of binary toxicity labels by which each molecule is said to be either toxic or non-toxic, depending on the effect of the molecule as such on biological systems. The dataset Tox21 is mainly used for measuring Environmental and Human Damages caused by chemicals. It comprises a wide variety of chemicals and helps generalize a trained model across different pollutant categories. Because of this reason, the completeness and credibility of this dataset make it attractive for finding predictive models within the scope of computational toxicology. [16, 17]

The Tox21 data set comprises 12,060 training samples and 647 test samples that represent chemical compounds. There are 801 ”dense features” that represent chemical descriptors, such as molecular weight, solubility or surface area, and 272,776 ”sparse features” that represent chemical substructures (ECFP10, DFS6, DFS8; stored in Matrix Market Format). Machine learning methods can either use sparse or dense data or combine them. For each sample there are 12 binary labels that represent the outcome (active/inactive) of 12 different toxicological experiments. Dataset Resource available in the following link:

`https://www.bioinf.jku.at/research/DeepTox/tox21.html`

CHAPTER 4

ARCHITECTURE DIAGRAM

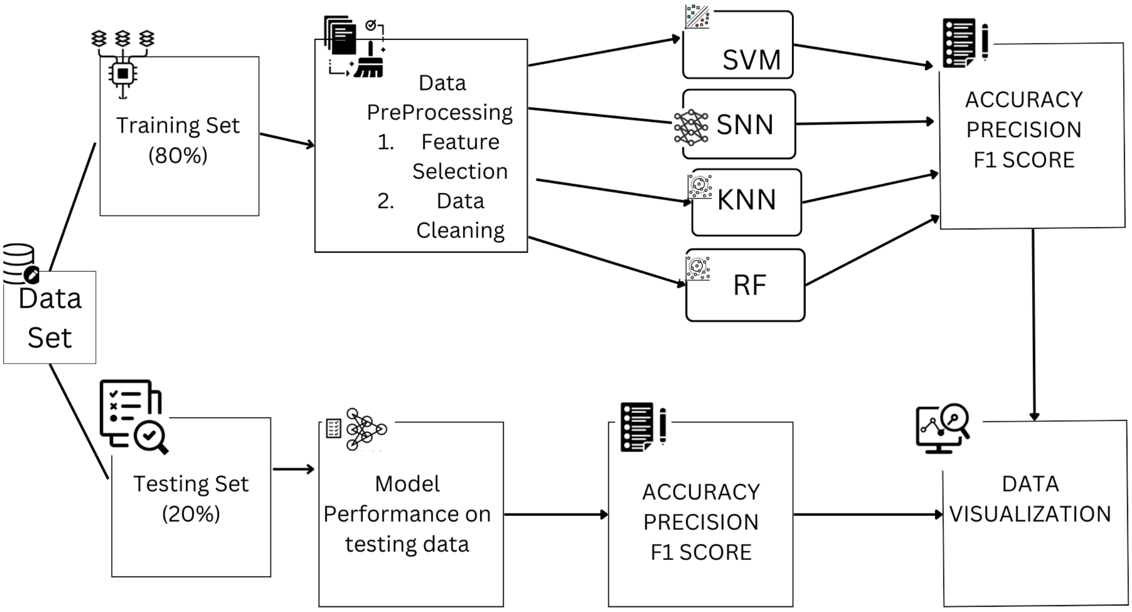

Figure 4.1: Data-Driven Model Evaluation Pipeline: Preprocessing, Training, and Performance Analysis

The architectural diagram that explains a machine learning workflow beginning from splitting the dataset into training set (80 percent) and testing set (20 percent). The training data itself is subjected to preprocessing, which encompasses feature selection as well as one or the other forms of data cleaning that can add to the model's performance. Four models, namely the Support Vector Machine (SVM), Simple Neural Network (SNN), K-Nearest Neighbors (KNN), and Random Forest (RF), were trained on the data that underwent processing. The hybrid model is made by merging three different kinds of machine-learning algorithms, specifically Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). This method utilizes the strength of each model for better performance of predictions.The model evaluation was based on accuracy, precision, and F1 score. Upon checking the performances of the models on the testing set, it was ensured that the models generalized well on unseen data. The final evaluation results are then visualized for better insight and comparison of the model performance.

CHAPTER 5

METHODOLOGY

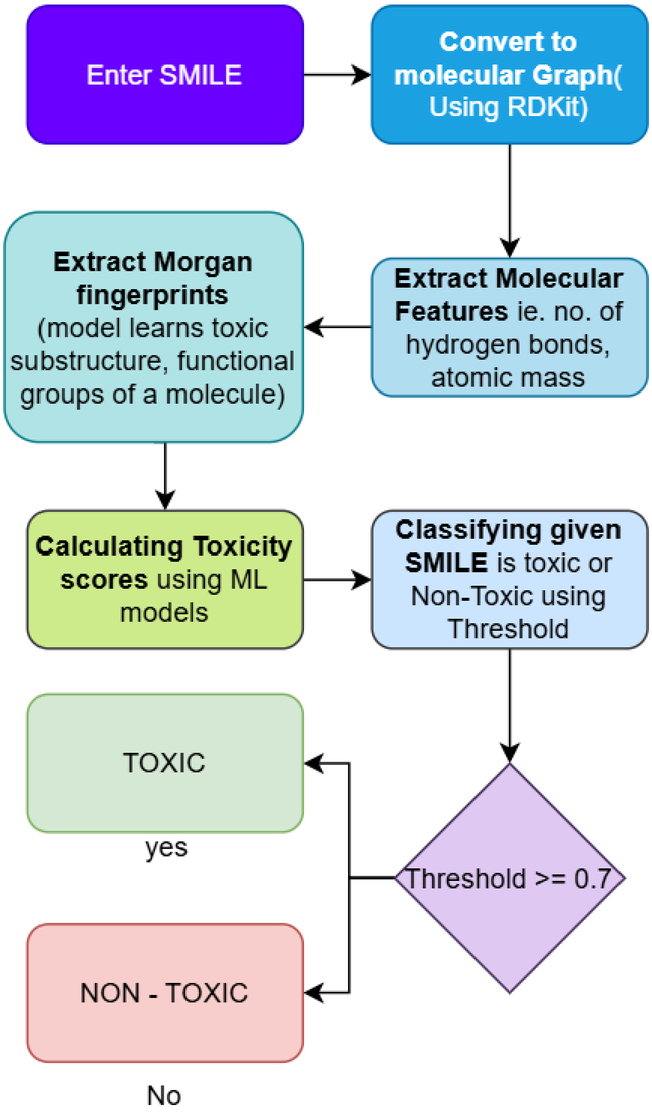
Enter SMILE
Convert to molecular Graph(Using RDKit)
Extract Morgan fingerprints (model learns toxic substructure, functional groups of a molecule)
Extract Molecular Features ie. no. of hydrogen bonds, atomic mass
Calculating Toxicity scores using ML models
Classifying given SMILE is toxic or Non-Toxic using Threshold
TOXIC
yes
Threshold >= 0.7
NON - TOXIC
No

Figure 5.1: Computational Toxicology Pipeline

The preparation of the data was undertaken to increase data integrity through deletion of duplicate records, instances of missing or inconsistent data. Significant chemical descriptors were selected, followed by normalization to ensure equal impacts of all features on model training. To conform with Machine Learning Algorithms, toxicity levels, labeled ”toxic” or ”non-toxic,” were converted into Numerical Entries. The descriptors such as molecular weight, the number of atoms, number of rotatable bonds, and LogP (Partition coefficient) were computed using

RDKit to provide a lay of machine learning features. This provided to analyze the relationship between the molecular structure and toxicological effects. The figure [Fig.1] represents a Computational Toxicology Pipeline, where SMILES notation is converted into molecular graphs, features are extracted, toxicity scores are computed using ML models, and classification is done based on a threshold. Moreover, four machine learning algorithms plus a hybrid model were employed for the prediction of the toxicity of chemical pollutants. [19, 20]

5.1 SUPPORT VECTOR MACHINE

The Support Vector Machine (SVM) is a supervised machine learning classifier that labels compounds as toxic or non-toxic by finding a best hyperplane in high-dimensional feature space. SVM does it by mapping the data points to a higher feature space where data points are linearly separable and the margin of classes is maximum. Radial Basis Function (RBF) kernel enhances the model performance by introducing a non-linear mapping such that the model can detect more intricate data relationships. [21, 19] Mathematically, the RBF kernel is defined as:

$$K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right), \tag{5.1}$$

where γ controls how much influence a single training sample has. The SVM decision function is given by:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right), \tag{5.2}$$

where $\alpha_\mathbf{i}$ are the model coefficients, $\mathbf{y_i}$ are the class labels, and $\mathbf{b}$ is the bias. This Model uses SVC(kernel=”rbf”, probability=True), meaning it applies the RBF kernel and estimates class probabilities using Platt scaling.

5.2 K NEAREST NEIGHBORS

The k-Nearest Neighbors (KNN) algorithm classifies molecules based on structural similarities, relying on a majority vote of the closest training examples. The similarity between points is typically measured using Euclidean distance:

$$d(x_i, x_j) = \sum_{m=1}^{M} (x_{im} - x_{jm})^2 \tag{5.3}$$

where M is the number of features. Given a test sample, KNN selects the K nearest training points and assigns the most common label among them. The classification decision is given by:

$$\hat{y} = \arg\max_{c} \sum_{i \in \mathcal{N}_K} \mathbf{1}(y_i = c) \tag{5.4}$$

where $\mathcal{N}_\mathbf{K}$ represents the K nearest neighbors. This Model uses KNeighborsClassifier(n_neighbors=5), meaning it considers the five closest training points for classification.

5.3 SELF ORGANIZING NEURAL NETWORKS

Self-Organizing Neural Networks (SNN) relied on unsupervised learning in classifying compounds with identical structural features, which, in turn, played role in toxicity pattern analysis.

5.4 RANDOM FOREST

The Random Forest algorithm is an ensemble learning method technique for building multiple decision trees and merging their predictions for higher accuracy and consistency. Each tree recursively partitions the feature space into binary splits maximizing information gain or Gini impurity reduction. While a single tree can overfit, Random Forest prevents this through the application of bootstrap sampling and a random subset of features at each split. The final prediction is determined by taking a majority vote.

$$\hat{y} = \arg\max_{c} \sum_{t=1}^{T} \mathbf{1}(h_t(x) = c) \tag{5.5}$$

where $\mathbf{h_t(x)}$ is the prediction of the $\mathbf{t}$ -th tree. This Model uses RandomForestClassifier(n_estimators=100, random_state=42), meaning it builds 100 decision trees. Random Forest is highly effective for molecular descriptor-based classification, handling non-linearity, reducing overfitting, and performs well on imbalanced datasets.

5.5 HYBRID MODEL

This hybrid model is an ensemble learning method that combines three machine learning algorithms—Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) through a VotingClassifier with soft voting in order to improve predictive performance. Each model contributes differently, Random Forest uses several decision trees for strong classification, KNN classifies on similarity to neighbor-data points, and SVM obtains best decision boundaries. The soft voting mechanism procedure allocates weighted probabilities (40% for Random Forest, 30% each for KNN and SVM) to compute the final prediction. This integration uses each model's strength ensuring precise and unbiased classification while it generalizes its performance across various cases.. Moreover, the model can give toxicity predictions for unseen chemical compounds as well as probability scores and in this way, it can be of significant benefit for real-life uses.

5.5.1 FEASIBILITY ANALYSIS

The feasibility of the hybrid model can be analyzed using ensemble learning theory, probability-weighted voting, and the bias-variance tradeoff. The soft voting mechanism in the Voting Classifier ensures that the final prediction depends on combining weighted probabilities of the individual base models. Weighted averaging lowers the effect of the individual models' drawbacks and helps the overall stability. Another critical aspect ensuring feasibility is the bias-variance tradeoff, where the generalization error is defined as:

$$E[(y - \hat{y})^2] = \text{Bias}^2 + \text{Variance} + \sigma^2 \tag{5.6}$$

Here, Random Forest has low bias but high variance, KNN is locally sensitive, and SVM gives a robust decision boundary with low variance. The ensemble benefits from error compensation, where the strengths of one model cancel out the weaknesses of another, resulting in an optimal tradeoff between bias and variance. From a theoretical perspective, the ensemble error is always lower than the error of individual models, as per the Bayes Optimality Bound:

$$E_{\text{ensemble}} \leq E_{\text{best base model}} \tag{5.7}$$

This follows Condorcet's Jury Theorem, which states that combining multiple independent classifiers reduces the probability of error, provided that each model performs better than random guessing. Additionally, the ambiguity decomposition states that ensemble error is reduced by model diversity:

$$E_{\text{ensemble}} = \frac{1}{N} \sum_{i=1}^{N} E_i - \text{diversity} \tag{5.8}$$

Since Random Forest, KNN, and SVM come from different learning modes, they make diverse predictions, which in turn improve information generalization. These mathematical parameters reinforce the opinion that the hybrid model provides not only a solution but also is a theoretically empirical. justified choice to achieve better classification results. Both models were used to make the prediction, while the performance improvement was achieved through the tuning of the hyperparameter by means of grid search cross-validation. The model computation was performed using accuracy, precision, recall, F1 score, and AUC- ROC, among others. The study of the generalization of the models was based on the prediction of the toxicity of the molecular structures that were not previously seen in the form of SMILES notation and which were transcribed into molecular descriptors using RDKit. This was followed by distinguishing toxic substances handling imbalanced datasets, and scalability being the main determinants of effectiveness. The combined model, which uses four algorithms like the former models, was established and was implemented in Python using the scikit-learn package for modeling and the Matplotlib and Seaborn packages for visualizing results. RDKit has provided means for the creation of writings that involve the depiction of molecules. The predictive models have the know-how of how to use in environmental applications. They are easy to use for policy- and decision-making that deals with the elimination of toxic pollutants from biodiversity and ecosystems. Thus, they are supportive as some means of conservation and management.

CHAPTER 6

RESULTS AND DISCUSSIONS

The Machine Learning Algorithms designed to predict toxicity provided useful insights into classifying chemical molecules as harmful or safe. Comparing the models tested, Random Forest and Support Vector Machines (SVM) had the best levels of accuracy and robustness, with precision and recall scores indicating accurate predictions in both categories. The ROC-AUC scores demonstrated their ability to successfully differentiate between dangerous drugs. On the other hand, k-Nearest Neighbors (KNN) performed modestly, with only minor class imbalances, whereas Self-Organizing Neural Networks (SNN) suffered severely, possibly due to the dataset's complexity and the algorithm's unsupervised nature. According to the comparison [Table 1] and comparative study demonstrates the potential of machine learning in the field of computational toxicology, allowing for faster and more cost-effective toxicity studies than previous methods. The Random Forest model was found to be the most effective because of its capacity to manage complicated, non-linear interactions and provide insights into feature relevance. These findings highlight the importance of incorporating machine learning technology into environmental monitoring and risk assessment, paving the way for proactive efforts to reduce the effects of pollutants on ecosystems and biodiversity.

Metric	**SVM**	**KNN**	**Random Forest**	**SNN**	**Hybrid Model**
Precision (Toxic)	0.56	0.52	0.53	0.45	1.00
Precision (Non-Toxic)	0.50	0.46	0.47	0.41	0.94
Recall (Toxic)	0.51	0.43	0.45	0.36	0.03
Recall (Non-Toxic)	0.55	0.55	0.55	0.51	1.00
F1-Score (Toxic)	0.53	0.47	0.49	0.40	0.06
F1-Score (Non-Toxic)	0.53	0.50	0.51	0.46	0.97
Accuracy	0.53	0.49	0.50	0.43	0.94

Table 6.1: Comparison of Performance Metrics for Individual and Hybrid Models.

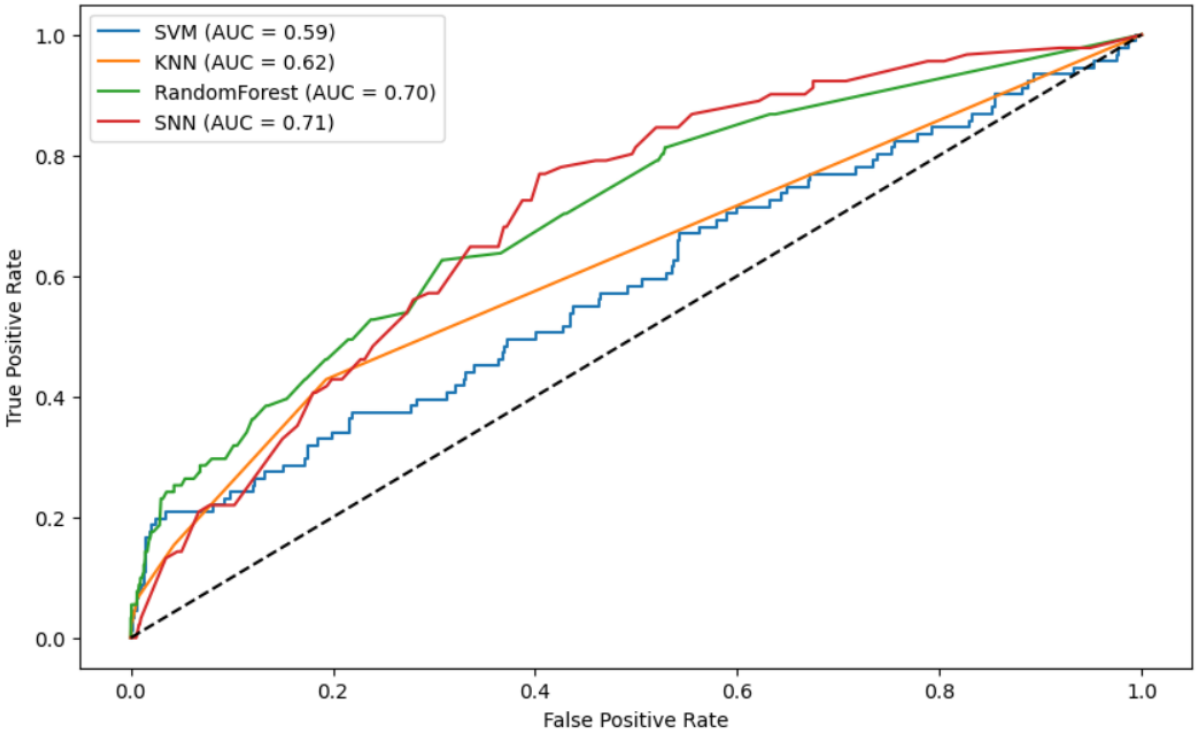

Figure 6.1: ROC Curve of the Models Used

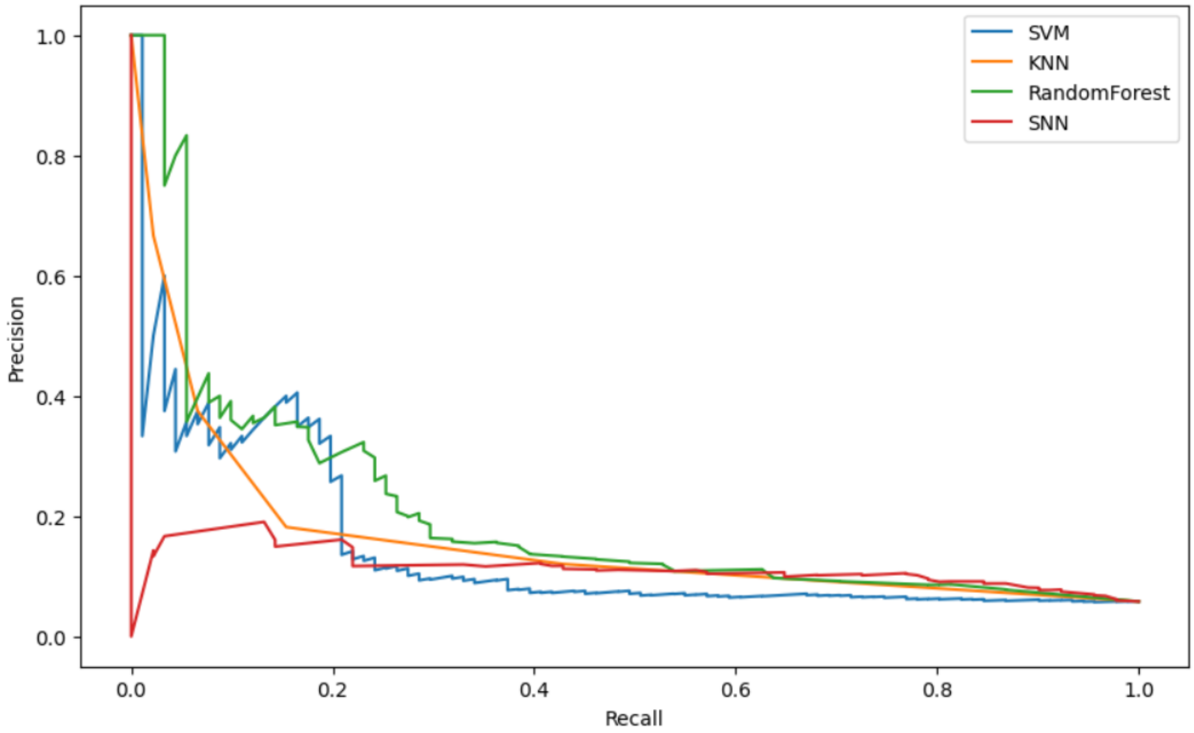

Figure 6.2: Precision - Recall Curve

The ROC curve [Fig.2] illustrates the trade-off between the true positive rate and false positive rate, showing that SNN achieves the highest AUC, indicating better overall performance.

The Precision-Recall curve [Fig.3] highlights the balance between precision and recall, demonstrating that RandomForest and SNN perform better in distinguishing toxic and non-toxic samples, especially in imbalanced datasets.

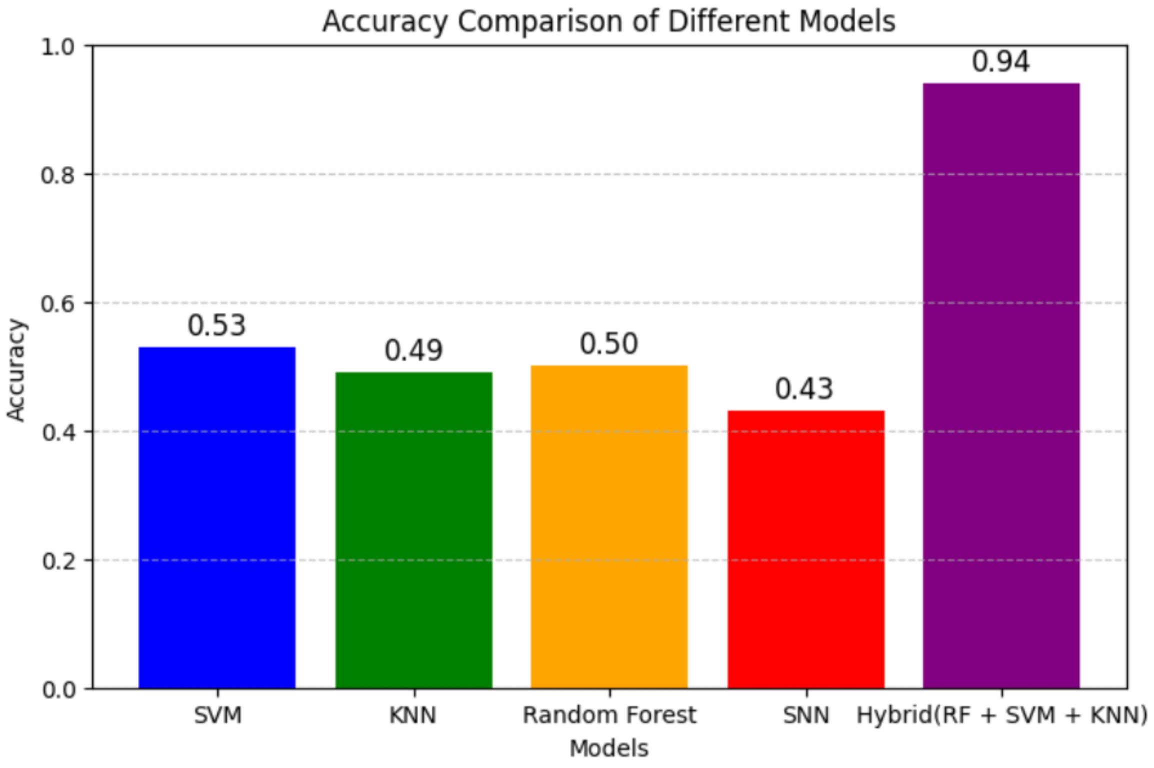

Figure 6.3: Comparison of Accuracies of Individual and Hybrid Models

The bar chart [Fig.4] compares the accuracy of different models, showing that the Hybrid (Random Forest + KNN + SVM) model achieves the highest accuracy (0.94), outperforming individual models like Random Forest (0.50), SVM (0.53), KNN (0.49), and SNN (0.43).

CHAPTER 7

CONCLUSION

The results of the present study throw light on the great promise held by machine learning models in the field of computational toxicology concerning determining the toxicity of chemical pollutants depending on their structure at the molecular level. Comparing Support Vector Machines (SVM), k-nearest Neighbors (KNN), Random Forest (RF), and Self-Organizing Neural Networks (SNN) revealed that ensemble-based methods promise greater accuracy and reliability compared to conventional methods used for toxicity screening. Among the models compared, Random Forest demonstrated better performance in dealing with high-dimensional datasets and identifying complicated molecular interactions. SVM provided good classification power but was confronted with lower-bound decision areas, while KNN signified difficulties in the handling of large data sets due to its reliance on local feature similarity. However, the integration of models in hybrid learning frameworks, such as soft voting classifiers, improved predictive performance to a lesser extent, reducing classification errors while increasing generalizability across broad sets of chemical collections.

Beyond the comparisons in performance, the work also underlines the necessity of using feature selection methods and molecular descriptors such as LogP, rotatable bonds, and atom composition to enhance toxicity prediction. Having the Tox21 dataset as a benchmark allowed for better organized and reproducible testing of these computational models. In addition, the exploration of the effects of individual molecular descriptors on toxicity classification revealed deeper insights into the chemical properties affecting levels of toxicity. Such insights can prove critical in regulatory decision-making, allowing policymakers to make more confident and efficient judgments about environmental hazards. Also, the strength of AI-based toxicology testing transcends chemical pollution management, providing avenues for pharmaceutical safety testing, agrochemical assessments, and industrial chemical risk evaluations.

CHAPTER 8

FUTURE SCOPE

This Research work demonstrates substantial computational toxicology opportunities, specifically in using Machine Learning models like SVM and KNN to process spatial and topological molecular data. Like Random Forest, Ensemble learning techniques can be used to improve toxicity predictions, and increasing datasets to include a wider variety of pollutants from databases such as PubChem and Tox21 can facilitate model generalization. With real-time toxicity evaluation tools in the form of software and APIs with IoT sensors, on-site measurements in aquatic and agricultural environments can become more robust and applicable.

SHAP contributes considerably to interpreting the results, permitting researchers to make deeper interpretations about how different molecular properties contribute to toxicity prediction and how toxicity prediction can be targeted to allow for better decision-making. By providing empirical evidence regarding the relationship between Environmental Sciences, Chemistry, and Biology and the work done in this area, the research lends credibility to certain model predictions. Machine Learning predictions can eventually complement regulatory policies that will be highly useful in the development of chemical safety regulations and sustainable production processes. The application of these models for exploring pollutants' impacts on the ecosystem, food web, and biodiversity levels may provide insight into environmental changes. By including climate and ecological parameters in their work, computational toxicologists can cultivate sustainability practices, safeguard the environment, and alleviate pollution's long-term effects.

BIBLIOGRAPHY

[1] Kavlock R, Dix D. Computational toxicology as implemented by the U.S. EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. J Toxicol Environ Health B Crit Rev. 2010 Feb;13(2-4):197-217. doi: 10.1080/10937404.2010.483935. PMID: 20574897.

[2] Jeong J, Choi J. Artificial Intelligence-Based Toxicity Prediction of Environmental Chemicals: Future Directions for Chemical Management Applications. Environ Sci Technol. 2022 Jun 21;56(12):7532-7543. doi: 10.1021/acs.est.1c07413. Epub 2022 Jun 6. PMID: 35666838.

[3] Chattopadhyay, Arnab , Banerjee, Swarnendu , Samadder, Amit ,Bhattacharya, Sabyasachi. (2022). Environmental toxicity influences disease spread in consumer population. 10.48550/arXiv.2211.08558.

[4] M. Chatterjee and K. Roy, ”Recent Advances on Modelling the Toxicity of Environmental Pollutants for Risk Assessment: from Single Pollutants to Mixtures,” *Curr. Pollut. Rep.*, vol. 8, no. 2, pp. 81–97, Jun. 2022, doi: 10.1007/s40726-022-00219-6.

[5] Jeong J, Choi J. Artificial Intelligence-Based Toxicity Prediction of Environmental Chemicals: Future Directions for Chemical Management Applications. Environ Sci Technol. 2022 Jun 21;56(12):7532-7543. doi: 10.1021/acs.est.1c07413. Epub 2022 Jun 6. PMID: 35666838.

[6] M. Banaee, A. Zeidi, R. Sinha, and C. Faggio, ”Individual and Combined Toxic Effects of Nano-ZnO and Polyethylene Microplastics on Mosquito Fish (*Gambusia holbrooki*),” *Water*, vol. 15, no. 9, Art. no. 1660, 2023, doi: 10.3390/w15091660.

[7] Li L, Dong Y, Chen Y, Jiao J, Zou X. A New Method for Environmental Risk Assessment of Pollutants Based on Multi-Dimensional Risk Factors. Toxics. 2022 Oct 30;10(11):659. doi: 10.3390/toxics10110659. PMID: 36355950; PMCID: PMC9697580.

[8] Feng H, Zhang L, Li S, Liu L, Yang T, Yang P, Zhao J, Arkin IT, Liu H. Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints. Toxicol Lett. 2021 Apr 1;340:4-14. doi: 10.1016/j.toxlet.2021.01.002. Epub 2021 Jan 6. PMID: 33421549.

[9] A. M. Smaldone and V. S. Batista, ”Quantum-to-Classical Neural Network Transfer Learning Applied to Drug Toxicity Prediction,” *J. Chem. Theory Comput.*, vol. 20, no. 11, pp. 4901–4908, Jun. 2024, doi: 10.1021/acs.jctc.4c00432.

[10] B. Sharma, V. Chenthamarakshan, A. Dhurandhar, et al., ”Accurate Clinical Toxicity Prediction Using Multi-Task Deep Neural Nets and Contrastive Molecular Explanations,” Sci. Rep., vol. 13, Art. no. 4908, 2023, doi: 10.1038/s41598-023-31169-8.

[11] M. Da Silveira, L. Deladiennee, K. Acem and O. Freudenthal, ”Combining knowledge graphs and LLMs for hazardous chemical information management and reuse,” 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 2024, pp. 6766-6773, doi: 10.1109/BIBM62325.2024.10821991.

[12] Q. Yu, Z. Zhang, G. Liu, W. Li, and Y. Tang, ”ToxGIN: An *In Silico* Prediction Model for Peptide Toxicity via Graph Isomorphism Networks Integrating Peptide Sequence and Structure Information,” *Brief. Bioinform.*, vol. 25, no. 6, Art. no. bbae583, Nov. 2024, doi: 10.1093/bib/bbae583.

[13] Y. Myung, A. G. C. de Sá, and D. B. Ascher, ”Deep-PK: Deep Learning for Small Molecule Pharmacokinetic and Toxicity Prediction,” *Nucleic Acids Res.*, vol. 52, no. W1, pp. W469–W475, Jul. 2024, doi: 10.1093/nar/gkae254.

[14] J. Jiang, R. Wang, and G.-W. Wei, ”GGL-Tox: Geometric Graph Learning for Toxicity Prediction,” *J. Chem. Inf. Model.*, vol. 61, no. 4, pp. 1691–1700, Apr. 2021, doi: 10.1021/acs.jcim.0c01294.

[15] W. Tang, X. Zhang, H. Hong, J. Chen, Q. Zhao, and F. Wu, ”Computational Nanotoxicology Models for Environmental Risk Assessment of Engineered Nanomaterials,” *Nanomaterials (Basel)*, vol. 14, no. 2, Art. no. 155, Jan. 2024, doi: 10.3390/nano14020155.

[16] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, ”DeepTox: Toxicity Prediction Using Deep Learning,” *Front. Environ. Sci.*, vol. 3, 2016, doi: 10.3389/fenvs.2015.00080.

[17] R. Huang, M. Xia, D.-T. Nguyen, T. Zhao, S. Sakamuru, J. Zhao, S. A. Shahane, A. Rossoshek, and A. Simeonov, ”Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure

to Environmental Chemicals and Drugs," *Front. Environ. Sci.*, vol. 3, 2016, doi: 10.3389/fenvs.2015.00085.

[18] N. Archanaa, V. V. J. Daniel, S. Divya, K. Raja, and I. R. Oviya, "Tomato Disease Classification Using CNN," in *Smart Systems: Innovations in Computing*, A. K. Somani, A. Mundra, R. K. Gupta, S. Bhattacharya, and A. P. Mazumdar, Eds. Singapore: Springer Nature Singapore, 2024, pp. 259–272, doi: 10.1007/978-981-97-3690-4-20.

[19] S. S. Rohit Mamidi, C. Akhil Munaganuri, T. Gollapalli, A. T. V. S. Aditya and R. C. B, "Implementation of Machine learning Algorithms to Identify Freshness of Fruits," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 1395-1399, doi: 10.1109/ICICICT54557.2022.9917989

[20] S. N, S. Nema, B. K. R, P. Seethapathy and K. Pant, "The Plant Disease Detection Using CNN and Deep Learning Techniques Merged with the Concepts of Machine Learning," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1547-1551, doi: 10.1109/ICACITE53722.2022.9823921.

[21] M. Sathya, K. Pavithra, and V. Poojasree, "Stroke Prediction Using Machine Learning," IARJSET, vol. 9, Jun. 2022, doi: 10.17148/IARJSET.2022.9620.