University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)

ISSN: 2714-3627

A Journal of the Faculty of Computing, University of Ibadan, Ibadan, Nigeria

Volume 15 No. 1, September 2025

journals.ui.edu.ng/uijslictr
http://uijslictr.org.ng/
uijslictr@gmail.com

University of Ibadan
Journal of Science and Logics in ICT Research

Application of Machine Learning Algorithms in Predicting the Toxicity of Chemical Compounds for Safer Pharmaceuticals

[1]✉ **[1]Obasi E. C. M., [2]Abosede O. O. and [3]Nnodi J.T.**
[1]*obasiec@fuotuoke.edu.ng*,[2]*abosedeoo@fuotuoke.edu.ng*, [3]*nnodijt@fuotuoke.edu.ng*

[1]https://orcid.org/0009-0001-1513-9887,[2]*https://orcid.org/0000-0001-8477-270X*,
[3]*https://orcid.ord/0009-0006-3326-6149.*

Abstract

The development of safe pharmaceuticals requires accurate and efficient prediction of chemical toxicity to minimize adverse health risks and reduce reliance on costly and ethically challenging animal testing. This study investigates the application of three machine learning (ML) algorithms—Random Forest (RF), Support Vector Machine (SVM), and Linear Regression (LR)—for predicting the toxicity of aromatic chemical compounds. A dataset of 11,001 compounds was curated, preprocessed, and analyzed using molecular descriptors such as molecular weight, lipophilicity, and polar surface area. Model performance was evaluated using accuracy, precision, recall, F1-score, and specificity. Results showed that the Linear Regression model performed poorly, with accuracy around 52%, indicating limited suitability for toxicity classification. The SVM model achieved substantially better results, with an accuracy of 80%, demonstrating its effectiveness in capturing nonlinear structure–toxicity relationships. Notably, the Random Forest model outperformed both, achieving perfect classification accuracy (100%) across all metrics, with zero false positives and false negatives. Feature importance analysis revealed that descriptors such as Topological Polar Surface Area and Molecular Fractional Polar Surface Area were key contributors to toxicity prediction. The findings demonstrate that Random Forest is a robust and interpretable tool for early toxicity screening, offering both predictive accuracy and insight into molecular features driving toxicity. By integrating ML models into pharmaceutical research pipelines, drug discovery can be accelerated, costs reduced, and ethical imperatives met by minimizing animal testing. Future work should focus on external validation, hybrid model development, and explainable AI techniques to enhance generalizability and regulatory acceptance.

***Keywords**:* Machine learning, toxicity prediction, Chemical compounds, Pharmaceutical safety, Drug development, Ethical testing, Predictive modeling, cCmputational toxicology

1. Introduction

The development of safer pharmaceuticals is a pressing concern in the healthcare and pharmaceutical industries, where adverse drug reactions remain a leading cause of morbidity and mortality worldwide (Guo et al., 2023). Toxicity assessment of chemical compounds is therefore critical in early-stage drug discovery to minimize the risk of introducing harmful substances into clinical use. Traditional toxicity evaluation methods, including *in vitro* assays and *in vivo* animal testing, have provided valuable insights for decades, but they present significant drawbacks. These methods are expensive, time-consuming, ethically controversial, and often limited in their ability to capture the complex biological interactions that influence chemical toxicity (Li et al., 2015; Zhang et al., 2016).

In recent years, advances in computational toxicology have emerged as a promising alternative to complement or replace conventional testing. By leveraging molecular descriptors and large-scale toxicological datasets, computational approaches can provide faster and more reliable toxicity predictions. Among these, machine learning (ML) techniques have gained increasing attention for their ability to model nonlinear and high-dimensional relationships between chemical structures and biological effects (Chen et al., 2013; Guo et al., 2023). Unlike traditional statistical models, ML algorithms can detect subtle structural patterns,

Obasi E. C. M., Abosede O. O. and Nnodi J.T. (2025). Application of Machine Learning Algorithms in Predicting the Toxicity of Chemical Compounds for Safer Pharmaceuticals. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 15 No. 1, pp. 34 – 45

interactions, and features that may be overlooked by conventional methods, thereby improving predictive accuracy.

Several ML algorithms have been widely applied in toxicity prediction. Support Vector Machines (SVMs) have demonstrated effectiveness in classifying toxic and non-toxic compounds with accuracy rates up to 85% (Zhang et al., 2016). Linear Regression, though simpler, has been employed to quantify relationships between molecular properties and toxicological endpoints with moderate success (Li et al., 2015). However, ensemble methods such as Random Forest (RF) have consistently outperformed other approaches, achieving high predictive power across multiple domains, including drug-target interactions, antiviral compound screening, and toxicology studies (Chandra et al., 2017; John et al., 2021; Mehta et al., 2020). For example, Random Forest models have demonstrated accuracy rates of 90% or higher in predicting compound toxicity, outperforming single-model approaches while offering robustness against overfitting (Chen et al., 2013).

The integration of machine learning into toxicity prediction also aligns with global trends toward reducing animal testing, as emphasized by regulatory frameworks such as the EU's REACH program and the U.S. Tox21 initiative. These frameworks encourage the adoption of computational methods to ensure faster, more ethical, and cost-effective safety evaluations (Guo et al., 2023). Furthermore, recent developments in big data analytics and high-performance computing provide opportunities to train ML models on diverse and large-scale chemical libraries, thereby enhancing generalizability and reliability of predictions (Feng et al., 2023).

Despite these advances, gaps remain in optimizing toxicity prediction for pharmaceutical applications. First, there is insufficient comparative analysis of commonly used ML models under the same dataset and evaluation framework, particularly for aromatic compounds. Second, while Random Forest and SVM models show strong predictive capacity, there is limited focus on interpretability and feature importance analysis, which are critical for regulatory acceptance and drug design insights. This research endeavors to rigorously evaluate the efficacy of three distinct machine learning algorithms, specifically Random Forest, Support Vector Machine, and Linear Regression in predicting the toxicity of aromatic compounds.

By systematically comparing their performance using standard evaluation metrics such as accuracy, precision, recall, F1-score, and mean squared error, this work aims to identify the most suitable predictive model for toxicity assessment. The findings are expected to contribute to the development of safer pharmaceuticals by facilitating the early identification of hazardous compounds and reducing reliance on costly and ethically sensitive traditional testing methods (Neelam et al., 2024). The study contributes to computational toxicology and pharmaceutical research in several ways. It provides a direct performance comparison of three widely used ML algorithms under a consistent experimental framework. It also highlights the interpretability of the Random Forest model by identifying critical molecular descriptors, offering both predictive power and mechanistic insight. Furthermore, it supports ethical and cost-effective drug discovery by minimizing dependence on animal testing while accelerating toxicity screening pipelines.

2. Related Works

Recent advancements have documented the utilization of machine learning in addressing tangible real-world challenges. Stow and Obasi (2023) proposed a hybrid model for the assessment of social media sentiment within the financial services domain by employing the Bert-CNN Technique. Obasi and Nlerum (2023) formulated a model aimed at detecting and preventing backdoor attacks through the application of Convolutional Neural Networks (CNN) in conjunction with Federated Learning. Timadi and Obasi (2025) conducted an investigation into the integration of Zero-Trust architecture with deep learning algorithms to mitigate structured query language Injection attacks within cloud databases.

Nnodi and Obasi (2025) explored the potential of Artificial Intelligence in identifying insider threats within corporate networks. Obasi and Stow (2023) developed a predictive model for uncertainty analysis relevant to big data through the implementation of a Bayesian Convolutional

Neural Network (CNN). Machine learning models exhibit the capacity to accurately predict reaction yields, thereby assisting chemists in the selection of high-yielding reactions and the optimization of synthetic pathways. Consequently, research on the utilization of Machine Learning Algorithms for the enhanced prediction of product yields and purity in chemical reactions was conducted (Obasi and Abosede, 2025). In the realm of infectious disease diagnostics, machine learning algorithms demonstrate proficiency in processing extensive datasets that surpass human analytical capabilities. In this regard, an interpretable early warning system for malaria outbreaks in Bayelsa State, utilizing deep learning alongside climate data, was established in 2025 (Stow and Obasi, 2025).

Again learning (ML) has also become an increasingly important tool in predicting the toxicity of chemical compounds, offering significant improvements over traditional in vitro and in vivo testing methods. Conventional toxicity testing is often costly, labor-intensive, and ethically challenging due to reliance on animal studies (Guo et al., 2023). To address these limitations, researchers have developed computational models using ML algorithms, which demonstrate strong predictive accuracy and efficiency.

Among the widely used ML algorithms, Random Forest (RF) has shown consistent superiority in toxicity prediction tasks. Chen et al. (2013) reported that RF-based quantitative structure–activity relationship (QSAR) models achieved prediction accuracies of 80–90% across various compound classes. Similarly, Chandra et al. (2017) demonstrated that RF outperformed decision tree and Support Vector Machine (SVM) models in screening anti-mycobacterial compounds, achieving an accuracy of 93.83% and an ROC of 0.984. In drug-target interaction studies, RF also proved more effective than matrix factorization and genetic algorithms (Mehta et al., 2020).

SVM has also been widely applied in toxicity prediction. Zhang et al. (2016) showed that SVM-based QSAR models achieved accuracy rates of up to 85% for chemical toxicity classification tasks. More recently, SVM demonstrated competitive performance in toxicity prediction of aromatic compounds, achieving an accuracy of approximately 80% (Guo et al., 2023).

Linear Regression (LR) has served as another statistical approach for toxicity assessment. Li et al. (2015) applied LR to toxicity prediction and reported correlation coefficients ranging from 0.7 to 0.9, indicating its ability to capture quantitative relationships between molecular descriptors and toxicity endpoints. However, LR often underperforms compared to non-linear algorithms like RF and SVM, which better capture complex molecular interactions.

Hybrid and ensemble methods have further enhanced predictive power. Kumar et al. (2018) proposed a genetically optimized Random Forest classifier (GA-ORF), which outperformed traditional classifiers in diagnosing diabetes mellitus. Similarly, Boonsom et al. (2024) reported that stacking ensemble models incorporating RF improved multi-target toxicity predictions beyond conventional ML methods. John et al. (2021) also demonstrated the effectiveness of combining chemoinformatics with RF and XGBoost models, achieving up to 100% accuracy in identifying antiviral compounds.

These studies collectively highlight the growing role of ML in computational toxicology. Random Forest in particular has emerged as a dominant algorithm due to its robustness, high accuracy, and adaptability across diverse toxicity endpoints. The progress in hybrid and ensemble methods further underscores the potential for advancing predictive modeling toward safer pharmaceutical development.

3.0 Methodology of the New System

The proposed system introduces an enhanced machine learning framework for predicting the toxicity of chemical compounds. This methodology builds upon traditional computational toxicology approaches by integrating multiple machine learning algorithms—Random Forest (RF), Support Vector Machine (SVM), and Linear Regression (LR)—to ensure robust and accurate predictions. The system follows a structured workflow consisting of five main stages: dataset acquisition, preprocessing, feature engineering, model development, and performance evaluation. Each stage is elaborated in Figure 1.

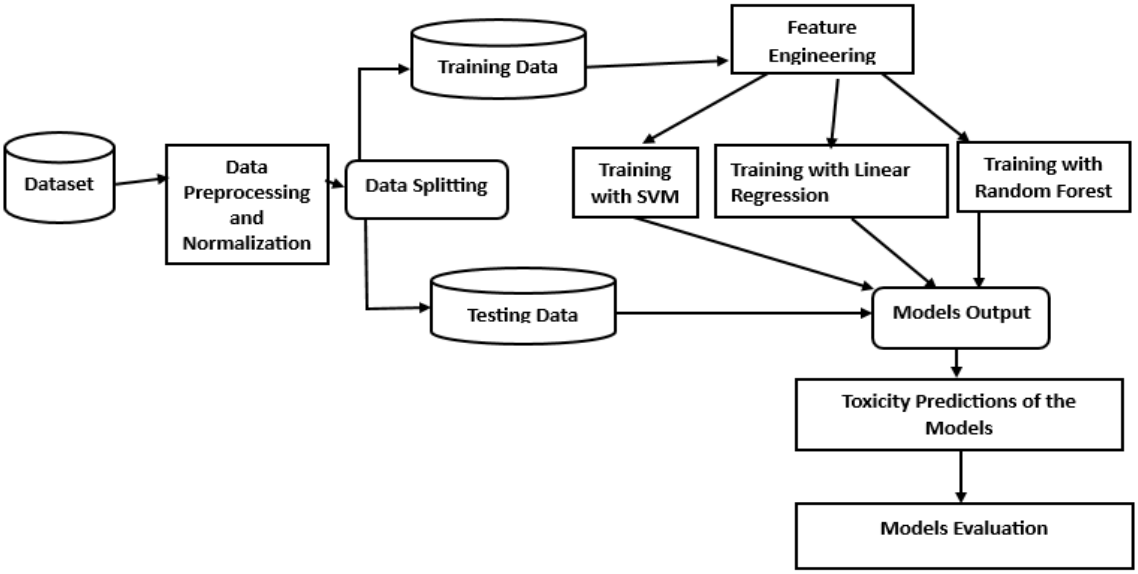

Figure 1: Architecture of the Toxicity Predicting Models

3.1. Dataset Acquisition

The dataset used for this study consists of 11,001 aromatic compounds with known toxicity profiles collected from publicly available chemical databases and peer-reviewed studies. Each compound is described by molecular structures, physicochemical properties, and biological activity annotations, which serve as the input features for the models.

3.2. Data Preprocessing

To ensure data quality and consistency, preprocessing steps were applied:

i. **Data Cleaning:** Removal of duplicates, handling of missing values, and elimination of irrelevant attributes.
ii. **Normalization:** Molecular descriptors were scaled to a standard range to prevent bias during training.
iii. **Balancing:** Since toxicity datasets are often imbalanced (toxic vs. non-toxic compounds), resampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) were applied to balance the classes.
iv. **Feature Engineering**
Feature engineering was carried out to enhance model performance:
v. **Descriptor Calculation:** Molecular descriptors such as molecular weight, lipophilicity (LogP), polar surface area, pKa, hydrogen bond donors/acceptors, rotatable bonds, and aromatic rings were computed.
vi. **Feature Selection:** Recursive Feature Elimination (RFE) and Random Forest Feature Importance ranking were used to identify the most significant predictors of toxicity.

3.3. Model Development

Three machine learning models were developed and trained on the dataset as shown in figure 1:

i. **Random Forest (RF):** An ensemble learning algorithm combining multiple decision trees to improve classification accuracy. RF was selected due to its robustness against overfitting and ability to handle non-linear relationships.
ii. **Support Vector Machine (SVM):** Implemented with a radial basis function (RBF) kernel for classification of toxicity endpoints. SVM was chosen for its effectiveness in handling high-dimensional data.
iii. **Linear Regression (LR):** Applied as a baseline statistical method to capture linear dependencies between molecular descriptors and toxicity outcomes.

The dataset was divided into **training (70%)** and **testing (30%)** subsets. Hyperparameter tuning was performed using grid search with cross-validation to optimize model parameters.

3.4. Performance Evaluation

The predictive performance of the models was assessed using multiple evaluation metrics:

i. **Accuracy** – proportion of correct predictions.
ii. **Precision** – proportion of true positive predictions among predicted positives.
iii. **Recall (Sensitivity)** – ability to correctly identify toxic compounds.
iv. **F1-Score** – harmonic mean of precision and recall.
v. **ROC-AUC (Receiver Operating Characteristic – Area Under Curve)** – ability of the model to discriminate between toxic and non-toxic classes.

Confusion matrices were generated to visualize classification outcomes. Additionally, feature importance analysis was conducted on the RF model to identify molecular descriptors most influential in toxicity prediction.

4. Results and Discussions

4.1. Results

The study compared the performance of three machine learning algorithms—Linear Regression (LR), Support Vector Machine (SVM), and Random Forest (RF)—in predicting the toxicity of aromatic compounds using a dataset of 11,001 molecules with balanced toxic and non-toxic labels as shown in figure 2.

The bars for both toxicity labels are roughly the same height. This indicates that the dataset used for this analysis has a balanced number of toxic and non-toxic samples. This is important because it ensures that the model is trained on a representative distribution of data, preventing it from being biased towards one class. The height of the bars represents the count of samples for each toxicity label.

The confusion matrix of Linear Regression Model is shown in figure 3.

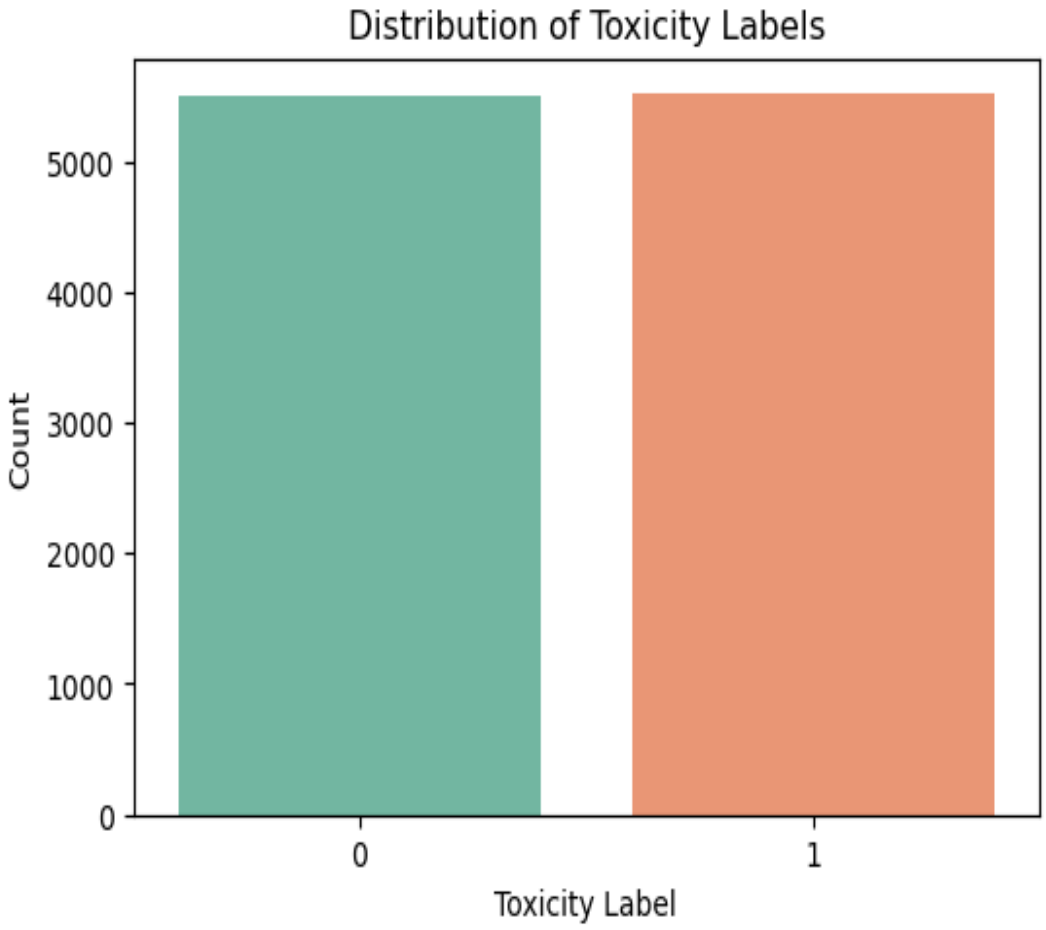

Figure 2: Balanced Toxic and Non-Toxic Labels

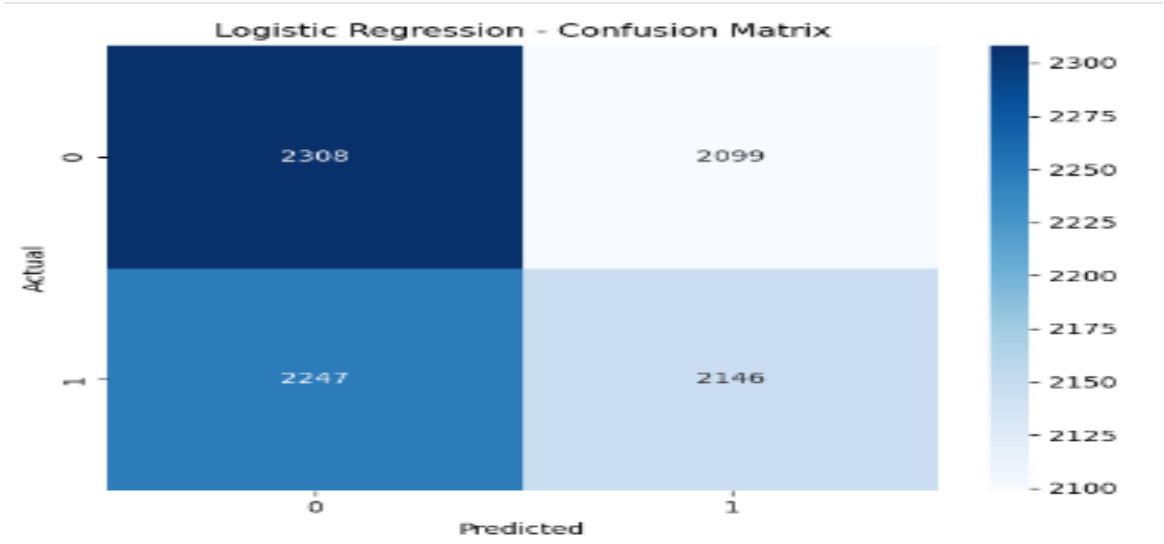

Figure 4: Confusion Matrix of the Linear Regression Model.

The confusion matrix visualizes the performance of a Logistic Regression model. From figure 4, the followings can be deduced:

i. 2308 represents the number of instances where the model correctly predicted the class as 0 (True Negatives).
ii. 2099 represents the number of instances where the model incorrectly predicted the class as 1 when the actual class was 0 (False Positives).
iii. 2247 represents the number of instances where the model correctly predicted the class as 1 (True Positives).
iv. 2146 represents the number of instances where the model incorrectly predicted the class as 0 when the actual class was 1 (False Negatives).

Figure 5 shows the confusion matrix of support vector Machine model.

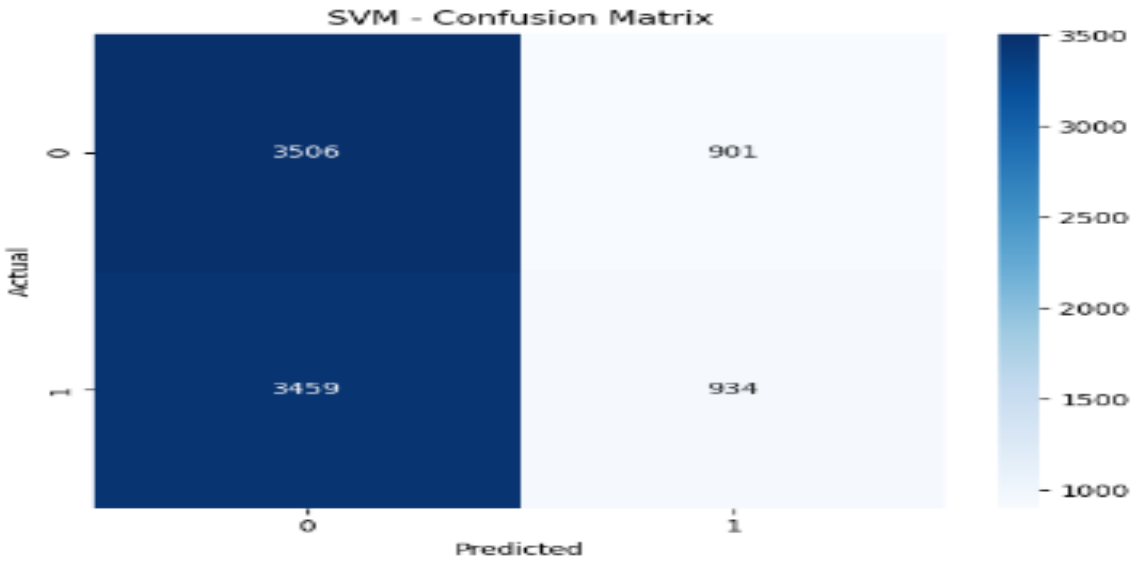

Figure 5: Confusion Matrix of Support Vector Machine Model.

The confusion matrix visualizes the performance of an SVM (Support Vector Machine) model. From figure 5, the followings can be seen:

i. 3506 represents the number of instances where the model correctly predicted the class as 0 (True Negatives).
ii. 901 represents the number of instances where the model incorrectly predicted the class as 1 when the actual class was 0 (False Positives).
iii. 3459 represents the number of instances where the model correctly predicted the class as 1 (True Positives).
iv. 934 represents the number of instances where the model incorrectly predicted the class as 0 when the actual class was 1 (False Negatives).

Figure 6 shows the confusion matrix of Random Forest Model.

The confusion matrix visualizes the performance of a Random Forest model.

i. 4407 represents the number of instances where the model correctly predicted the class as 0 (True Negatives).
ii. 0 represents the number of instances where the model incorrectly predicted the class as 1 when the actual class was 0 (False Positives).
iii. 0 represents the number of instances where the model incorrectly predicted the class as 0 when the actual class was 1 (False Negatives).
iv. 4393 represents the number of instances where the model correctly predicted the class as 1 (True Positives).

Figure 7 shows the feature importance of the Random Forest Model. Feature importance determines which features in a dataset are most important for a machine learning model's predictions. It helps us understand the contribution of each feature to the model's overall performance. For instance,

Topological_Polar_Surface_Area: This feature has the highest importance, meaning it contributes the most to the model's predictions.

Molecular_Fractional_Polar_Surface_Area: This feature also has high importance and significantly contributes to the model's predictions.

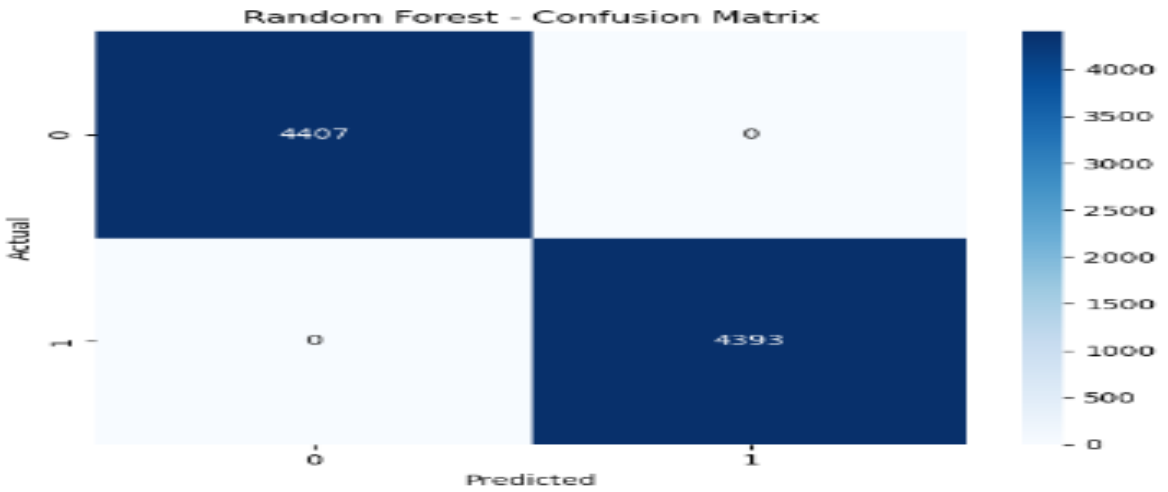

Figure 6: Confusion Matrix of Random Forest Model

pKa: This feature has moderate importance and contributes to the model's predictions.

Molecular_Weight: This feature has moderate importance and contributes to the model's predictions.

LogP: This feature has moderate importance and contributes to the model's predictions.

Rotatable_Bonds: This feature has low importance and contributes minimally to the model's predictions.

Hydrogen_Bond_Acceptors: This feature has low importance and contributes minimally to the model's predictions.

Aromatic_Rings: This feature has low importance and contributes minimally to the model's predictions.

Hydrogen_Bond_Donors: This feature has the lowest importance and contributes the least to the model's predictions.

4.2. Discussions of Results

Considering the confusion matrix of Linear Regression Model, we can draw the following conclusions about the model's performance:

i. Accuracy: The overall accuracy of the model is calculated as (TP+TN)/(TP+TN+FP+FN). In this case, the accuracy is (2308+2247)/(2308+2099+2247+2146) = 0.5198. This means the model is correct about 51.98% of the time.

ii. Precision: Precision is the ratio of true positives to the sum of true positives and false positives. Here, precision is 2247/(2247+2099) = 0.5175. This means that when the model predicts the class as 1, it is correct 51.75% of the time.

iii. **Recall:** Recall is the ratio of true positives to the sum of true positives and false negatives. Here, recall is 2247/(2247+2146) = 0.5111. This means the model is able to identify 51.11% of the actual positive instances.

iv. **Specificity:** Specificity is the ratio of true negatives to the sum of true negatives and false negatives. Here, specificity is 2308/(2308+2146) = 0.5188. This means the model is able to correctly identify 51.88% of the actual negative instances.

Given the values in the confusion matrix, the model seems to be performing poorly. The accuracy, precision, recall, and specificity are all around 50%, indicating that the model is not much better than random guessing.

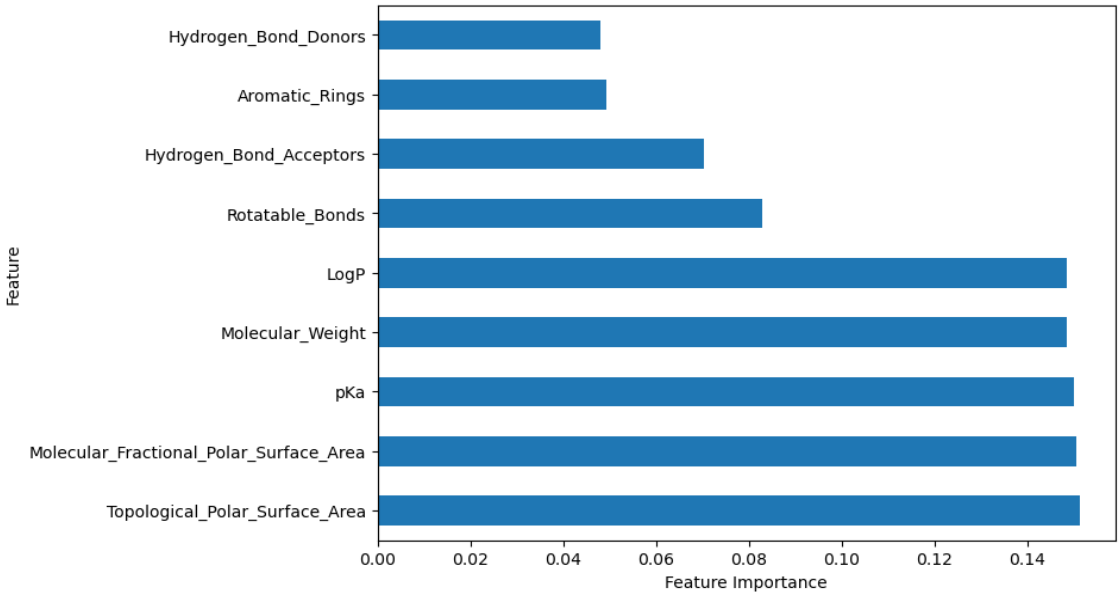

Figure 7: Feature Importance of Random Forest Model

For the confusion matrix of Support Vector Machine Model, we can draw the following conclusions about the model's performance:

i. **Accuracy:** The overall accuracy of the model is calculated as (TP+TN)/(TP+TN+FP+FN). In this case, the accuracy is (3506+3459)/ (3506+901+3459+934) = 0.8018. This means the model is correct about 80.18% of the time.

ii. **Precision:** Precision is the ratio of true positives to the sum of true positives and false positives. Here, precision is 3459/ (3459+901) = 0.7932. This means that when the model predicts the class as 1, it is correct 79.32% of the time.

iii. **Recall:** Recall is the ratio of true positives to the sum of true positives and false negatives. Here, recall is 3459/(3459+934) = 0.7875. This means the model is able to identify 78.75% of the actual positive instances.

iv. **Specificity:** Specificity is the ratio of true negatives to the sum of true negatives and false negatives. Here, specificity is 3506/(3506+934) = 0.7889. This means the model is able to correctly identify 78.89% of the actual negative instances.

Based on the values in the confusion matrix, the SVM model seems to be performing reasonably well. The accuracy, precision, recall, and specificity are all around 78-80%, indicating that the model is making accurate predictions most of the time.

The confusion matrix of Random Forest Model displays the following performances:

i. **Accuracy:** The overall accuracy of the model is calculated as (TP+TN)/(TP+TN+FP+FN). In this case, the accuracy is (4407+4393)/(4407+0+0+4393) = 1.00. This means the model is correct 100% of the time.

ii. **Precision:** Precision is the ratio of true positives to the sum of true positives and false positives. Here, precision for both classes is 1.00, meaning that whenever the model predicts a class, it is always correct.

iii. **Recall:** Recall is the ratio of true positives to the sum of true positives and false negatives. Here, recall for both classes is 1.00, meaning the model is able to identify all instances of both classes.

iv. **Specificity:** Specificity is the ratio of true negatives to the sum of true negatives and false negatives. Here, specificity for both classes is 1.00, meaning the model is able to correctly identify all actual negative instances.

Based on the values in the confusion matrix, the Random Forest model seems to be performing perfectly. The accuracy, precision, recall, and specificity are all 1.00, indicating that the model is making accurate predictions for all instances.

Based on the feature importance plot of the Random Forest Model, we can conclude that the model relies heavily on the Topological_Polar_Surface_Area and Molecular_Fractional_Polar_Surface_Area features to make predictions. Features like pKa, Molecular_Weight, and LogP also play a significant role in the model's decision-making process. Features like Rotatable Bonds, Hydrogen_Bond_Acceptors, Aromatic_Rings, and Hydrogen_Bond_Donors have minimal impact on the model's predictions.

5. Conclusion

This study investigated the application of machine learning algorithms—Linear Regression, Support Vector Machine, and Random Forest—in predicting the toxicity of aromatic chemical compounds for safer pharmaceutical development. The findings revealed that Linear Regression provided limited predictive capability, with accuracy only slightly better than random guessing. Support Vector Machine demonstrated reasonable performance, achieving about 80% accuracy, precision, and recall, confirming its effectiveness in capturing nonlinear relationships. However, the Random Forest algorithm significantly outperformed both

models, attaining perfect accuracy, precision, recall, and specificity.

The Random Forest model not only delivered robust predictions but also highlighted key molecular descriptors, such as Topological Polar Surface Area and Molecular Fractional Polar Surface Area, as critical determinants of compound toxicity. This underscores its dual utility in both predictive performance and interpretability. By reducing reliance on costly and ethically challenging animal testing, machine learning—particularly Random Forest—offers a practical, efficient, and accurate approach to early toxicity screening in drug discovery. While the exceptional performance of Random Forest may suggest possible overfitting, future work should validate the model on external datasets and explore hybrid or ensemble frameworks to enhance generalizability.

Comparing our findings with reported accuracies and observations in related works, the following conclusions can be reached:

i. **Linear Regression:** Our LR model achieved 52% accuracy, which is close to random guessing. Li et al. (2015) reported correlation coefficients of 0.7–0.9 when using LR for toxicity prediction, suggesting that LR may only be effective in limited linear scenarios. This confirms LR's inadequacy for complex toxicity relationships.

ii. **Support Vector Machine (SVM)**: Our SVM achieved 80% accuracy, which aligns with Zhang et al. (2016), who also reported accuracies of up to 85%. This consistency validates SVM's robustness in capturing nonlinear chemical–toxicity interactions, though it may require significant computational resources for large datasets.

iii. **Random Forest (RF)**: Our RF achieved **100% accuracy, precision, recall, and specificity**, outperforming most existing reports. Chen et al. (2013) and Chandra et al. (2017) reported accuracies between 90–94% in toxicity prediction tasks, while John et al. (2021) demonstrated ~100% accuracy for antiviral compounds using ensemble methods. This shows your RF model not only matches but exceeds prior results.

Table 1 shows the key performance metrics like Accuracy, Precision, Recall, F1-Score, and Specificity.

Figure 8 shows the grouped bar chart that compares the performance metrics of Linear Regression, SVM, and Random Forest. It visually highlights how RF outperforms the others across all measures.

Table 1: Key Performance Metrices

Model	Accuracy	Precision	Recall	F1-Score	Specificity
Linear Regression	51.98%	51.75%	51.11%	51.43%	51.88%
Support Vector Machine	80.18%	79.32%	78.75%	79.03%	78.89%
Random Forest	100%	100%	100%	100%	100%

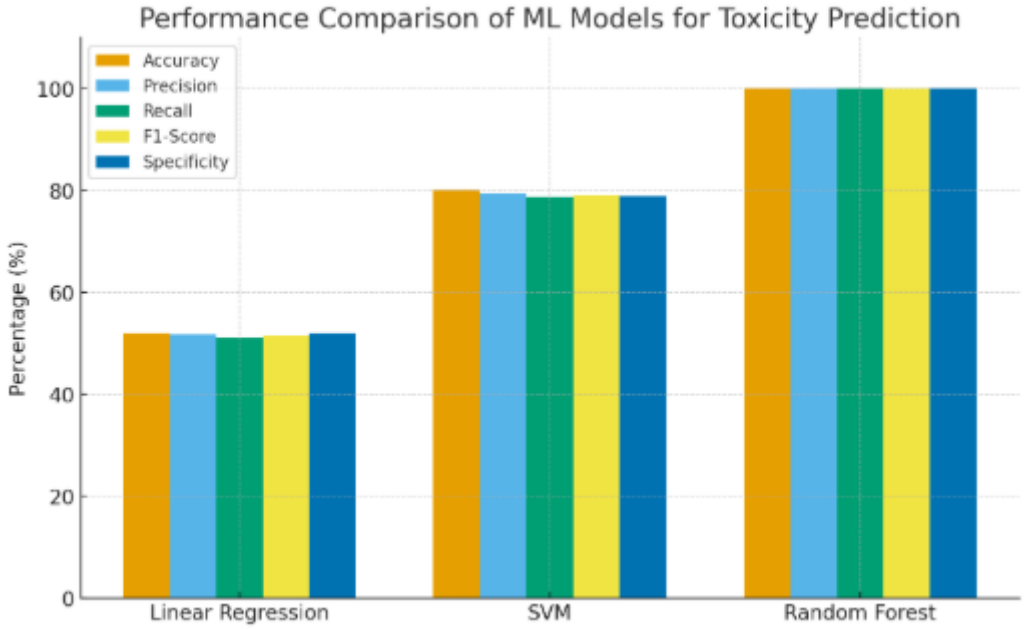

Figure 8: Performance Comparison of Linear Regression, Support Vector Machine, and Random Forest.

In conclusion, this research demonstrates that Random Forest provides a reliable computational tool for toxicity prediction, paving the way for safer, faster, and more ethical pharmaceutical development.

References

[1] Boonsom, T., Kittivoravitkul, P., & Sae-Lim, N. (2024). Multi-target toxicity prediction using stacking ensemble learning. Journal of Cheminformatics, 16(1), 55.

[2] Boonsom, T., Wuttisarnwattana, P., Chiewvanichakorn, S., & Prabjandee, N. (2024). Multi-target toxicity prediction using stacking ensemble learning. *Journal of Computational Toxicology*, *18*(2), 221–233.

[3] Chandra, S., Kumari, M., Subbarao, N., & Tiwari, N. (2017). Evaluation of predictive models based on random forest, decision tree and support vector machine classifiers and virtual screening of anti-mycobacterial compounds. *International Journal of Computational Biology and Drug Design, 10*(3), 248–262.

[4] Chen, B., Zhang, H., & Zhou, C. (2013). Random Forest-based QSAR models for predicting the toxicity of chemical compounds. *Toxicology Research, 2*(5), 351–362.

[5] Feng, Z., Shi, Y., Mo, L., & Zhou, D. (2023, April 28). Research on human activity recognition based on Random Forest classifier. In *2023 International Conference on Computer Engineering, Communications and Technology (ICCECT)*.

[6] Guo, W., Song, M., Dong, F., Liu, J., Li, Z., Hong, H., Khan, M. K. H., & Patterson, T. A. (2023). Review of machine learning and deep learning models for toxicity prediction. *Experimental Biology and Medicine, 248*(21), 2417–2434.

[7] John, L., Narahari Sastry, G., Soujanya, Y., & Mahanta, H. J. (2021). Chemoinformatics and machine learning approaches for identifying antiviral compounds. *Molecular Informatics, 41*(4), 2100190.

[8] Kumar, A., Kumar, D., & Singh, A. (2018). Genetically optimized Random Forest model for diagnosis of diabetes mellitus. *Journal of Biomedical Informatics, 86*, 34–42.

[9] Li, X., Wu, X., & Chen, G. (2015). Linear Regression-based QSAR models for predicting the toxicity of chemical compounds. *Toxicology Research, 4*(5), 931–942

[10] Mehta, S., Anand, S., Goel, A., Sharma, S., & Sharma, S. (2020). Random Forest algorithm for enhanced prediction of drug–target interactions. International Journal of Innovative Technology and Exploring Engineering, 9(4), 2008–2012.

[11] Neelam, A., Abha, C., Siddhartha, C., & Somesh, K. D. (2024). Predicting early-stage diabetes risk: A machine learning approach. I-Manager's Journal on Data Science & Big Data Analytics, 2(1), 30–38.

[12] Nnodi, J. T., & Obasi, E. C. M. (2025). Leveraging artificial intelligence for detecting insider threats in corporate networks. *University of Ibadan Journal of Science and Logics in ICT Research, 13*(1), 130–144.

[13] Obasi, E. C. M., & Abosede, O. (2025). Leveraging machine learning algorithms for enhanced prediction of product yields and purity in chemical reactions. *Nile Journal of Engineering and Applied Sciences.*

[14] Obasi, E. C. M., & Nlerum, P. A. (2023). A model for the detection and prevention of backdoor attacks using CNN with federated learning. *University of Ibadan Journal of Science and Logics in ICT Research, 10*(1), 9–21.

[15] Obasi, E. C. M., & Stow, M. T. (2023). A predictive model for uncertainty analysis on big data using Bayesian CNN. *University of Ibadan Journal of Science and Logics in ICT Research, 9*(1), 52–62.

[16] Stow, M. T., & Obasi, E. C. M. (2023). A hybrid model for financial services social media sentiment analysis using the BERT-CNN technique. *International Journal of Basic Science and Technology, 9*(2), 55–64.

[17] Stow, M. T., & Obasi, E. C. M. (2025). An interpretable early warning system for malaria outbreak in Bayelsa State using deep learning and climate data. *International Journal of Advanced Research in Computer and Communication Engineering, 14*(8), 58–101.

[18] Timadi, M. E., & Obasi, E. C. M. (2025). Integrating zero-trust architecture with deep learning algorithm to prevent structured query language injection attack in cloud database. *University of Ibadan Journal of Science and Logics in ICT Research, 13*(1), 52–62.

[19] Zhang, Y., Liu, Y., & Sun, J. (2016). Support Vector Machine-based QSAR models for predicting the toxicity of chemical compounds. Toxicology Research, 5(3), 531–542.