Global J. Environ. Sci. Manage. 11(1): 21-38, Winter 2025, Serial #41

Global Journal of Environmental Science and Management
(GJESM)

Homepage: https://www.gjesm.net/

Podcasts

ORIGINAL RESEARCH ARTICLE

# Machine learning model for random forest acute oral toxicity prediction

**A.M. Elsayad[1,*], M. Zeghid[2], K.A. Elsayad[3], A.N. Khan[4], A.K.M. Baareh[5], A. Sadiq[4], S.A. Mukhtar[6], H.F. Ali[6], S. Abd El-kader[6]**

[1] *Department of Electrical Engineering, College of Engineering in Wadi Alddawasir, Prince Sattam Bin Abdulaziz University, Wadi Alddawasir 11991, Saudi Arabia*

[2] *Computer Engineering and Networks Department, College of Engineering, PSAU, Saudi Arabia*

[3] *Pharmacy Department, Cairo University Hospitals, Cairo University, Cairo 11662, Egypt*

[4] *Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan*

[5] *Department of Applied Science, Ajloun University College, Al-Balqa Applied University, Ajloun, Jordan*

[6] *Computers and Systems Department, Electronics Research Institute, Cairo, Egypt*

**ARTICLE INFO**

***Article History:***
Received 06 August 2024
Revised 11 October 2024
Accepted 19 November 2024

***Keywords:***
Explainable artificial intelligence
Machine Learning (ML)
Random forest (RF)
Rational discovery kit (RDKit)
Sustainable development goals (SDGs)
Toxicity

DOI: 10.22034/gjesm.2025.01.02

**ABSTRACT**

**BACKGROUND AND OBJECTIVES:** The focus of this study is on the importance of reliable and precise forecasting of acute oral toxicity to bolster chemical safety and advance sustainable development goals, particularly sustainable development goals-3 (good health and well-being), sustainable development goals-6 (clean water and sanitation), and sustainable development goals-12 (responsible consumption and production). Traditional toxicity assessments are often time-consuming and costly, necessitating the exploration of more efficient approaches. The focus of this study is to establish the most efficient method for constructing reliable and precise models for toxicity prediction.

**METHODS:** The random forests were evaluated, a robust ensemble method, for predicting acute oral toxicity using a comprehensive dataset from National Toxicology Program/Interagency Center for the Evaluation of Alternative Toxicological Methods and Environmental Protection Agency/ National Center for Competency Testing, which presented significant class imbalance, 8 percent very toxic 92 percent not very toxic. To address this imbalance, strategies such as cost-sensitive learning and data resampling techniques, including both under sampling and oversampling, were utilized. A diverse set of two-dimensional molecular descriptors generated via rational discovery kit were used as input features, and model preprocessing involved normalization, validation, and feature selection. Hyper-parameter tuning was conducted using Bayesian optimization and cross-validation, while the performance of random forests was evaluated in comparison to gradient boosting, extreme gradient boosting, artificial neural networks, and the generalized linear model.

**FINDINGS:** The random forests models, particularly those utilizing under sampling and cost-sensitive learning, demonstrated superior performance, achieving sensitivity of 0.81, Specificity of 0.85, accuracy of 0.85, and an area under the receiver operating characteristic curve of 0.89 on an independent test set. An examination of feature importance has shown that the primary molecular descriptors are those related to the Van der waals surface area and molecular quantum numbers. A surrogate decision tree developed from random forests predictions reached an area under the curve of 0.929.

**CONCLUSION:** Random forest models effectively predicted acute oral toxicity, particularly when addressing class imbalance through cost-sensitive learning and resampling. leveraging explainable artificial intelligence techniques, including permutation feature importance, surrogate decision tree analysis and local interpretable model-agnostic explanations, this study identified key molecular descriptors driving toxicity. This advancement improves model interpretability and represents a significant step toward enhancing chemical safety while supporting sustainable development goals.

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

NUMBER OF REFERENCES
31

NUMBER OF FIGURES
5

NUMBER OF TABLES
2

*Corresponding Author:
Email: *a.elsayyad@psau.edu.sa*
Phone: +9665 6520 4379
ORCID: 0000-0001-8053-9759

Note: Discussion period for this manuscript open until April 1, 2025 on GJESM website at the "Show Article".

## INTRODUCTION

Acute toxicity testing evaluates the toxic effects of chemicals or products through predefined exposure routes (oral, dermal, inhalation) over a 21-day period, essential for regulatory classification and labeling of substances like industrial chemicals, biocides, pesticides, and cosmetics (Creton et al., 2010). Established in vivo methodologies, including the lethal dose 50 (LD50) test, present challenges such as substantial financial expenditure, prolonged duration, and ethical issues stemming from the use of animal subjects (Chinedu et al., 2013). Consequently, there is a growing initiative to advance computational strategies that leverage current LD50 data, in alignment with the Sustainable Development Goals (SDGs), particularly goal 3 and goal 12, aimed at promoting safer and more sustainable chemical practices (Sachs et al., 2019). In 2017, the acute toxicity workgroup (ATWG) within Interagency Coordinating Committee on the Validation of alternative methods (ICCVAM), led by National Toxicology Program (NTP) and Environmental Protection Agency (EPA), initiated a project to create in silico models for predicting acute oral toxicity using a comprehensive rodent dataset (Kleinstreuer et al., 2018). The project aimed to identify very toxic (VT) as LD50<50 milligram per kilogram (mg/kg) and nontoxic as LD50 > 2,000 mg/kg chemicals, estimate rodent LD50 values, and categorize toxicity hazards and globally harmonized system (GHS) standards. The project endeavored to increase regulatory acceptance of computational predictions by synthesizing multiple models.. The NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) and National Center for Competency Testing (NCCT) curated an extensive rat acute oral lethality dataset, inviting international research groups to develop predictive models, with a portion of the data reserved for evaluation to ensure robust, consensus-based predictions for regulatory use (Mansouri et al., 2021). The present study explores the performance of several machine learning models in predicting acute oral toxicity, drawing on a publicly available dataset characterized by imbalance, obtained from NTP NICEATM and EPA NCCT. Focusing on compounds with LD50 values below 50 mg/kg, the research evaluates five algorithms (RF (More et al., 2017), gradient boosting (GB) (Bentéjac et al., 2021), extreme gradient boosting (XGB) (Mateo et al., 2021), artificial neural networks (ANN) (Zhang et al., 2016), and generalized linear model (GLM) (Durrant et al., 2009) and employs (RDKit)-derived two dimensional (2D) molecular descriptors to characterize chemical structures (Landrum, 2013). This study aims to confront the problem of class imbalance, where the minority class constitutes only 8 percent (%) of the dataset, by thoroughly analyzing three different strategies: cost-sensitive learning, which focuses on enhancing the accuracy of the minority class through the application of weighted penalties; under sampling, which attempts to create balance between classes but may inadvertently lead to the loss of critical information; and oversampling, which retains all data but could increase the likelihood of overfitting issues (Hasib *et al.,* 2022). By systematically evaluating these techniques, this study aims to identify the most effective strategy for building accurate and reliable toxicity prediction models. The analysis leverages RDKit's comprehensive 2D molecular descriptors, encompassing topological, physicochemical, and surface area properties, to capture crucial chemical information for predicting acute oral toxicity. The approach encompassed several key components, which comprised data preprocessing, the identification of significant features, hyperparameter tuning, and the interpretation of the resulting model. After normalizing data and selecting predictive features, Bayesian optimization (Wu *et al.,* 2019) and cross-validation (Tougui *et al.,* 2021) fine-tuned the model. The performance of RFs surpassed that of other algorithms. Additionally, the integration of explainable artificial intelligence (AI) techniques, including permutation feature importance (PFI), surrogate decision trees (SDT), and local interpretable model-agnostic explanations (LIME), contributed to enhanced model transparency and fostered greater trust in the results (Muschalik *et al.,* 2022; Wilhelm *et al.,* 2024). PFI is a technique in explainable AI (XAI) that measures the impact of each feature on the model's performance by shuffling feature values and observing changes in the model's accuracy. This helps identify which features are most influential. SDTs are used to approximate complex models with simpler, interpretable decision trees. By training a decision tree on the predictions of a complex model, insights can be gained into the decision-making process, making it easier to understand and trust the model's outputs. Furthermore, LIME offers localized

explanations for specific predictions by modifying input features and analyzing the subsequent alterations in the predictions. This allows for identifying key features influencing specific predictions, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This literature review examines recent advancements and challenges in utilizing machine learning (ML) for predicting chemical toxicity (Born *et al.,* 2023) (Tran *et al.,* 2023), with a focus on improving traditional methods that are often constrained by cost, time, and ethical concerns (Cavasotto *et al.,* 2022) (Pérez *et al.,* 2021). The study by Wu *et al*. (2019) presents five critical barriers that impede the effective application of ML, highlighting concerns such as the quality of data, the intricacies of model design, interpretability, and the availability of necessary tools. They suggest strategies such as enhancing data integration, leveraging XAI, and exploring hybrid ML approaches to address these challenges, ultimately leads for safer and faster drug development. The research presented by Tougui *et al*. (2021) involves the development of machine learning models designed to forecast acute dermal toxicity (ADT), leveraging a dataset that includes over 1,600 different compounds. They compare seven algorithms and achieve, the area under the curve (AUC) values of 78% and 82% for rats and rabbits, respectively. Their study highlights specific chemical structures linked to toxicity and introduces "DermalPred," a user-friendly tool that offers an alternative to animal testing. The research conducted by Muschalik *et al*. (2022) delves into the role of equivariant graph neural networks (EGNNs) for predicting toxicity. The results indicate that the addition of three-dimensional structural information contributes positively to the accuracy of predictions, while the incorporation of total molecular energy as a feature fails to provide any enhancement. The study conducted by Wilhelm *et al.* (2024) encapsulates the promise of ML and deep learning (DL) models, highlighting their potential as cost-efficient alternatives to traditional toxicity testing methods. Their analysis encompasses 82 publications that explore diverse toxicity endpoints. It emphasizes the importance of high-quality datasets and calls for benchmark datasets to improve prediction reliability. The study by Abou Hajal *et al*. (2024) centers on the advancement of interpretable models by assessing various chemical representations and incorporating uncertainty estimation. Their results demonstrate that the combination of augmented simplified molecular input line entry system (SMILES) with convolutional attention networks yielded superior performance compared to alternative approaches. Lou *et al*. (2024) provide an analysis of trends in artificial intelligence applications aimed at predicting the toxicity of environmental chemicals. This review encompasses 96 studies and underscores the importance of combining big data with toxicity mechanisms. The study by Cremer *et al.* (2023) encapsulates the advancements made ML models focused on various toxicity endpoints highlighting the challenges related to the interpretability of deep learning models and the critical need for appropriate evaluation metrics in this context. Lastly, Guo *et al.* (2023) provides a multidisciplinary overview of AI applications in toxicity prediction, advocating for the shift from in vivo animal testing to in Silico models and underscoring the potential of AI to enhance accuracy and reduce reliance on animal studies. Overall, these studies collectively highlight the transformative potential of ML and AI in the field of toxicology, despite existing challenges in data quality, model interpretability, and ethical considerations. The current study was conducted at the College of Engineering, Prince Sattam bin Abdulaziz University, situated in the southern region of Riyadh province, Saudi Arabia, in 2024.

## MATERIALS AND MethodS

This study utilizes a machine learning workflow in Konstanz Information Miner (KNIME) to develop and evaluate predictive models for acute oral toxicity, addressing challenges like class imbalance and model interpretability. he procedure commences with the preparation of data, which entails segmenting a dataset that includes chemical abstracts service registry numbers (CAS-RN), SMILES strings, and binary toxicity labels into distinct calibration and external validation groups. SMILES strings are converted into 2D molecular structures using RDKit. Data is standardized using z-score normalization, highly correlated descriptors are removed, and feature selection is performed using information gain to retain the most informative features. These preprocessing steps are illustrated in Fig. 1.

This study distinguishes itself from many toxicity prediction investigations that typically use molecular

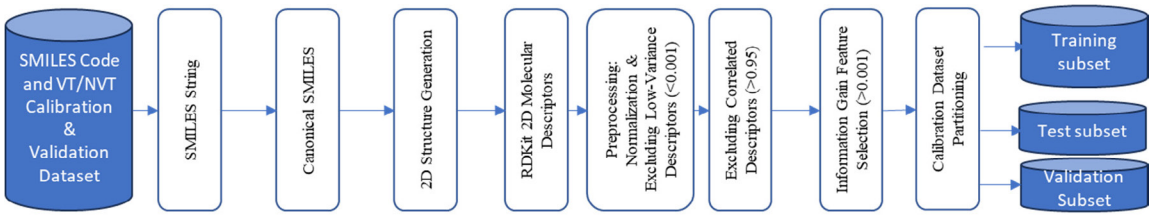


Fig. 1: Data acquisition, descriptor generation, feature selection, and data partitioning steps

fingerprints by employing RDKit 2D descriptors, which facilitate a more comprehensive analysis of molecular properties, including topological, physicochemical, and surface area characteristics. A rigorous preprocessing pipeline was implemented, including the removal of low-variance and highly correlated descriptors, followed by feature selection using information gain, enhancing model efficiency and interpretability by reducing redundancy and focusing on the most predictive features to optimize hyper-parameters, Bayesian optimization was utilized alongside five-fold cross-validation, thereby maximizing the performance of the model while also guaranteeing strong generalization capabilities. Five diverse algorithms were trained and evaluated: Random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), artificial neural networks (ANN), and generalized linear model (GLM). Both GB and XGB are esteemed for their superior predictive performance and their proficiency in capturing complex non-linear relationships. Nonetheless, they have a tendency to overfit, a concern that can be alleviated by employing cross-validation methods. ANNs excel at modeling complex patterns, providing powerful insights despite their lack of interpretability, making them useful for balancing predictive power and explain ability. GLM serves as a simple, interpretable baseline, highlighting the value of more complex models. RF was chosen for its robustness to high-dimensional data, ability to handle class imbalance, and feature importance evaluation, aligning with goals for accurate and interpretable predictions. To achieve efficient hyper-parameter optimization and maximize the area under AUC, a strong indicator for imbalanced datasets, Bayesian optimization was applied alongside five-fold cross-validation. Recognizing the significant class imbalance (8% VT), the modeling process was repeated with three different approaches to address this challenge: under sampling using random subsampling, oversampling using the synthetic minority over-sampling technique (SMOTE), and cost-sensitive learning. This extensive evaluation reveals the optimal strategy for handling this dataset. To enhance transparency and trust, explainable AI (XAI) techniques, including permutation feature importance (PFI), surrogate decision tree (SDT) and LIME, were employed to elucidate the model's decision-making process, highlighting influential features and their impact on predictions. The findings of this study outline a systematic and effective framework for constructing accurate and interpretable toxicity prediction models. It underscores the significance of data preprocessing, the selection of appropriate models, hyper-parameter optimization, and the resolution of class imbalance to enhance overall performance.

*Data acquisition and partitioning*

A comprehensive database of rat acute oral toxicity data, focused on LD50 values for systemic toxicity, was compiled by the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) and the Environmental Protection Agency National Center for Competency Testing (NCCT) as part of a collaborative effort to develop predictive models for acute oral toxicity. This publicly available dataset promotes transparency and reproducibility. It contains five endpoints pertinent to acute oral systemic toxicity, grounded in regulatory guidelines. Molecular structures were curated using the RDKit normalizer node to ensure accurate SMILES representation. The present study examines two primary endpoints: very toxic (VT) substances, defined by LD50 values that fall below 50 mg/kg, and not very toxic (NVT) compounds, which encompass all other substances.

Regarding ethical considerations, the data is derived from existing animal studies conducted according to regulatory guidelines. While the original animal studies raise ethical concerns inherent in animal testing, this research aims to reduce future reliance on such testing by developing predictive models. This dataset mainly emphasizes systemic toxicity findings from rat studies, potentially limiting its applicability to different species or other toxicity endpoints. In addition, despite the dataset's extensive coverage of different chemical types, it may not completely capture the diversity of chemicals that exist across various environmental contexts or in specific geographic locations. The dataset was initially divided into a calibration set (75%) and an external validation set (25%). The calibration set was further split into a training set (75%) and a test set (25%), resulting in 6,744 records for training, 2,248 records for testing, and 2,894 records for validation. Notably, the dataset shows a significant class imbalance, with only 8% of the records classified as VT and 92% as NVT.

*RDKit 2D descriptors computation*

The RDKit cheminformatics library is utilized to convert chemical information from SMILES strings, which represent molecular structures, into data that can be easily analyzed. This procedure commences with the validation of SMILES strings, which is essential for ensuring their correctness and alignment with the conventions governing chemical structures. Once validated, RDKit efficiently transforms the SMILES strings into 2D structural representations, visually depicting the molecule's connectivity and arrangement of atoms. Following this, the visual representations are translated into RDKit molecule objects, which are optimized for computational efficiency, thus permitting more in-depth analysis. A key strength of RDKit is its comprehensive suite of 119 2D descriptors, which include topological, constitutional, and electronic types. This broad spectrum of descriptors provides the means for in-depth molecular characterization and enhances the accuracy and reliability of predictions related to acute oral toxicity.

*Preprocessing and feature selection*

To improve the accuracy and efficiency of the ML models for acute oral toxicity prediction, a series of preprocessing steps were implemented:

*Low-variance descriptor removal*

Descriptors exhibiting minimal variance throughout the dataset are eliminated, as they contribute negligible information for differentiating between molecules. Descriptors with variance less than 0.001 were removed, eliminating four descriptors: slogp_VSA9, smr_VSA8, MQN18, and MQN39.

*Correlated descriptor removal*

Highly correlated descriptors (correlation > 95%) are identified and removed to avoid redundancy in the feature space. This step removed 14 descriptors: SMR, ExactMW, NumHeavyAtoms, NumAtoms, Chi3v, Chi4v, Chi2n, Chi3n, MQN1, MQN12, MQN17, MQN19, MQN22, and MQN23.

*Normalization*

The remaining descriptors are subjected to normalization to align the scales of all features. This process of normalization mitigates the risk of features with larger scales efrom having an excessive impact on the model's performance. This step normalized the training, test, and external validation subsets using the same mean and standard deviation from the training subset.

*Feature selection*

Information gain (IG) is used to select predictive descriptors and remove irrelevant ones, enhancing model performance by focusing on the most informative features. Such an approach yields greater accuracy and superior generalization capabilities, minimizes the complexity of the model, and facilitates easier interpretation. Additionally, it accelerates the training process, especially in the context of large datasets. Table 1 shows the top 20 predictive descriptors with their definitions.

The top five molecular descriptors (peoe_VSA14, slogp_VSA12, MQN7, slogp_VSA3, and NumHeteroAtoms) for predicting acute oral toxicity suggest that electrostatic interactions, lipophilicity, molecular complexity, and heteroatom content are key factors. The likelihood of toxicity is heightened in molecules that possess extensive polar surface areas, elevated lipophilicity, particular structural characteristics, and an increased count of heteroatoms. This finding emphasizes the model's proficiency in understanding the complex interplay between molecular structure and its toxic effects.

Table 1: The top 20 predictive descriptors for acute oral toxicity prediction and their definitions

| Serial | RDKit 2D descriptores | Definition | IG |
|---|---|---|---|
| 1 | Peoe_VSA14 | Represents the van der Waals surface area of atoms with partial charges in the range of 0.30 to infinity, capturing electrostatic properties. | 0.0396 |
| 2 | Slogp_VSA12 | Represents the van der Waals surface area of atoms with SlogP values from 0.60 to infinity, indicating hydrophobic regions. | 0.0304 |
| 3 | MQN7 | Counts the number of nitrogen atoms in a molecule, providing structural information related to nitrogen content. | 0.0292 |
| 4 | Slogp_VSA3 | Represents the van der Waals surface area of atoms with SlogP values from -0.20 to 0.001, capturing hydrophobic properties. | 0.0283 |
| 5 | NumHeteroAtoms | Counts the number of heteroatoms (non-carbon and non-hydrogen atoms) in a molecule, indicating structural diversity. | 0.0231 |
| 6 | Slogp_VSA8 | Represents the sum of van der Waals surface areas of atoms with a specific range of SlogP values (0.25 <= x < 0.30). | 0.0185 |
| 7 | Peoe_VSA1 | Represents the sum of van der Waals surface areas of atoms with partial charges in the range (-inf < x < -0.30). | 0.0181 |
| 8 | SMR_VSA10 | Represents the sum of van der Waals surface areas of atoms with a specific range of molar refractivity values (4.00 <= x < inf) | 0.0179 |
| 9 | TPSA | Topological Polar Surface Area (TPSA) is the surface sum over all polar atoms or molecules, primarily oxygen and nitrogen, including their attached hydrogen atoms. It is used to predict drug absorption and permeability. | 0.0175 |
| 10 | MQN29 | It is part of the Molecular Quantum Numbers (MQN) set, which includes various counts related to molecular structure, such as atom counts, bond counts, and polarity counts. | 0.0164 |
| 11 | Chi4n | It is a topological descriptor representing the fourth-order valence connectivity index, which is used to describe molecular branching . | 0.0154 |
| 12 | MQN16 | It is another descriptor from the MQN set, typically related to specific structural features or counts within the molecule. | 0.0146 |
| 13 | Peoe_VSA4 | Represents the sum of van der Waals surface areas of atoms with partial charges in the range (-0.20 <= x < -0.15). | 0.0132 |
| 14 | Chi2n | It is a second-order valence connectivity index, used to describe molecular branching. | 0.0131 |
| 15 | SMR_VSA1 | Represents the sum of van der Waals surface areas of atoms with molar refractivity values in the range (-inf < x < 1.29). | 0.0128 |
| 16 | Chi3n | It is a third-order valence connectivity index, used to describe molecular branching. | 0.0125 |
| 17 | HallKierAlpha | Represents the Hall-Kier alpha value for a molecule, which is related to molecular shape and branching. | 0.0122 |
| 18 | Peoe_VSA10 | Represents the sum of van der Waals surface areas of atoms with partial charges in the range (0.10 <= x < 0.15). | 0.0120 |
| 19 | Peoe_VSA2 | Represents the sum of van der Waals surface areas of atoms with partial charges in the range (-0.30 <= x < -0.25). | 0.0103 |
| 20 | MQN6 | It is another descriptor from the MQN set, typically related to specific structural features or counts within the molecule. | 0.0102 |

However, a comprehensive interpretation requires considering all features and their interplay.

*Machine learning models*

*Random forest*

During the training phase, Random Forest builds numerous decision trees and produces an output based on the predominant class among the trees for classification tasks, or the average prediction for regression tasks. Key hyper-parameters include the number of trees, the maximum depth of each tree, and the number of features to consider at each split. Its main strength lies in its ability to handle high-dimensional data and its robustness to outliers. RF is relatively easy to tune and often provides good performance without extensive hyper-parameter optimization. The training process can require significant computational resources, especially with extensive datasets, and it might not provide the same level of interpretability as simpler models. For

imbalanced datasets, RF can be beneficial as the random subspace sampling method used during tree construction can help to mitigate class bias. Recently, the RF regression model has been successfully developed as a quantitative structure-activity relationship (QSAR) model for predicting acute oral toxicity (LD50) in rats (Xiao *et al.,* 2022). In the study by Feng *et al*. (2021), the objective was to assess the reproductive toxicity of chemicals. The findings reveal that the amalgamation of Random Forest models, which are based on different types of fingerprints, results in superior predictive capabilities relative individual models.

*GB classifier*

GB sequentially builds an ensemble of weak learners, typically decision trees, with each tree attempting to correct the errors made by the previous trees. Important hyper-parameters encompass the learning rate, the count of trees, the maximum allowable depth for each tree, and the ratio of data samples employed for each tree. This iterative process leads to a strong learner with high predictive accuracy. GB excels at handling complex relationships within data and can achieve high performance on a variety of tasks. Nonetheless, this approach is vulnerable to overfitting, especially in the presence of noisy data, and it may require significant computational resources for training. For imbalanced datasets, specialized techniques like cost-sensitive learning or adjusting class weights can be used in conjunction with GB to address the bias towards the majority class. GB classification models have been successfully used to predict chemical toxicity in Jiang *et al.* (2021) and acute oral toxicity in (Ryu *et al.,* 2023).

*XGB classifier*

XGB is a specific implementation of GB that utilizes regularization techniques and parallel processing to improve performance and speed. Important hyper-parameters include the learning rate, the number of trees, the maximum depth of each tree, L1 and L2 regularization parameters, and the subsample ratio of data and features. XGB is known for its high accuracy, efficiency, and flexibility. This approach integrates methods including tree pruning and the identification of splits that account for sparsity, aiming to reduce overfitting and address the challenges posed by sparse data. Similar to GB, XGB can benefit from methods like cost-sensitive learning or adjusting class weights to improve performance on imbalanced datasets. XGBoost has been applied in various toxicity prediction tasks due to its high accuracy and efficiency (Wu *et al.,* 2024).

*ANN classifier*

ANNs are inspired by the structure and function of the human brain. They consist of interconnected nodes (neurons) organized in layers that process and transform data through weighted connections. Critical hyper-parameters that govern the network's architecture and learning dynamics include the number of hidden layers, the count of neurons in each layer, the activation functions assigned to the neurons, the learning rate, and the batch size utilized in training. ANNs excel at modeling complex non-linear relationships and can learn intricate patterns from large datasets. However, they can be computationally expensive to train, susceptible to overfitting, and often lack interpretability. In the context of imbalanced datasets, several approaches may be utilized to alleviate the imbalance, such as increasing the representation of the minority class through oversampling, reducing the majority class via under sampling, or employing cost-sensitive learning techniques. ANNs have a proven track record in toxin detection and classification research (Fan *et al.,* 2018).

*GLM classifier*

*GLMs* extend traditional linear regression models to accommodate a wider range of data distributions and relationships between variables. 2. Critical hyper-parameters involve determining the appropriate distribution for the response variable and the link function that establishes the relationship between the linear predictor and the response. GLMs are relatively simple to interpret, computationally efficient, and provide insights into feature importance. However, they may struggle to model complex non-linear relationships and can be sensitive to outliers. To address the challenges posed by imbalanced datasets, it is possible to adapt GLMs by modifying the loss function to increase the penalties associated with misclassifying instances from the minority class, which helps to mitigate the effects of class imbalance. The GLM model has been used extensively in modelling QSAR systems (Bharti *et al.,* 2017)

*Modelling and evaluation*

This study utilizes Bayesian Optimization within KNIME to efficiently tune hyper-parameters for imbalanced datasets, specifically targeting maximized AUC through five-fold cross-validation. By employing a probabilistic model, like a Gaussian Process, Bayesian Optimization systematically investigates the hyper-parameter space, establishing a relationship between hyper-parameters and AUC, which helps in steering the search towards configurations with greater potential. KNIME facilitates this with dedicated nodes, streamlining the process from defining the search space to outputting the optimal hyper-parameter set. To tackle the 8% toxicity class imbalance, the optimization is repeated using three strategies: under sampling, oversampling SMOTE, and cost-sensitive learning. Each approach has trade-offs:

*Under-sampling*

The goal of under-sampling is to create a more balanced class distribution by lowering the count of records in the majority class. This is achieved by randomly removing records from the majority class until a desired class ratio is achieved. Although this is the easiest method to implement, it could lead to under fitting due to the loss of significant information pertaining to the majority class.

*SMOTE*

oversampling balances class distribution by increasing the number of records in the minority class. This is frequently done by producing synthetic samples that closely resemble the existing instances of the minority class. By producing synthetic samples for the minority group, it helps prevent information loss, but it also poses the risk of overfitting and extending the training time.

*Cost-sensitive learning*

cost-sensitive learning directly addresses class imbalance by assigning different misclassification costs to each class. Misclassifications of the minority class incur higher costs, which incentivizes the model to prioritize accurate predictions for that class.

This systematic comparison under Bayesian optimization provides a comprehensive evaluation of these strategies, guiding the selection of the most effective approach for a well-performing, generalized model on this imbalanced dataset. Various performance metrics have been employed to assess the effectiveness of the optimized models across training, testing, and external validation subsets. They are including Recall (sensitivity), Precision, F1-score, specificity, F1-score Modified, Accuracy, and AUC. Most of them are computed using confusion matrix parameters like true positive (TP), true negative (TN), false positive (FP), false negative (FN):

*Recall (Sensitivity)*

This metric, calculated as the number of true positives divided by the sum of true positives and false negatives (TP / (TP + FN), reflects the model's ability to detect all positive cases.

*Precision*

This metric represents the fraction of true positive predictions relative to the overall number of positive predictions, expressed as TP / (TP + FP), reflecting the reliability of the positive outcomes.

*Specificity*

The ratio of true negatives that are correctly detected is represented by TN over the total of TN plus FP. This reflects the model's ability to capture all negative cases effectively.

*F1-score (Harmonic mean of recall and precision)*

This metric, which harmonizes recall and precision through the formula 2 * (Precision * Recall) / (Precision + Recall), serves as an effective tool when the objective is to maximize the identification of positive cases while simultaneously minimizing false positive rates.

*Adjusted F1-score (Harmonic mean of sensitivity and specificity)*

A balanced metric that combines sensitivity and specificity. Similar to F1-score but focuses on the correct classification of both positive and negative cases: (2 * (Sensitivity * Specificity)/ (Sensitivity + Specificity).

*Accuracy*

The overall ratio of correct predictions, defined as (TP + TN) / (TP + TN + FP + FN), can lead to misleading interpretations when applied to imbalanced datasets.

Table 2: Performance metrics on the training, test and external validation datasets using under sampling, oversampling and cos-sensitive learning

| Data subset | Balancing method | Model | Recall/sen | Precision | Spec. | F1 | ACC | AUC |
|---|---|---|---|---|---|---|---|---|
| Training Subset | Under sampling | RF | 1.00 | 0.39 | 0.86 | 0.57 | 0.87 | 0.99 |
| | | GB | 1.00 | 0.35 | 0.84 | 0.52 | 0.85 | 0.98 |
| | | XGB | 0.89 | 0.31 | 0.82 | 0.46 | 0.83 | 0.92 |
| | | ANN | 0.61 | 0.25 | 0.84 | 0.36 | 0.82 | 0.76 |
| | | GLM | 0.79 | 0.17 | 0.64 | 0.27 | 0.66 | 0.78 |
| | Over sampling | RF | 1.00 | 0.99 | 1.00 | 0.99 | 1.00 | 1.00 |
| | | GB | 1.00 | 0.97 | 1.00 | 0.98 | 1.00 | 1.00 |
| | | XGB | 0.69 | 0.54 | 0.95 | 0.61 | 0.93 | 0.93 |
| | | ANN | 0.68 | 0.20 | 0.76 | 0.31 | 0.76 | 0.75 |
| | | GLM | 0.79 | 0.18 | 0.68 | 0.29 | 0.69 | 0.80 |
| | Cost-sensitive learning | RF | 1.00 | 0.60 | 0.94 | 0.75 | 0.94 | 1.00 |
| | | GB | 0.99 | 0.88 | 0.99 | 0.93 | 0.99 | 1.00 |
| | | XGB | 0.87 | 0.48 | 0.91 | 0.62 | 0.91 | 0.96 |
| | | ANN | 0.92 | 0.43 | 0.89 | 0.59 | 0.89 | 0.96 |
| | | GLM | 0.81 | 0.17 | 0.65 | 0.29 | 0.67 | 0.80 |
| Test subset | Under sampling | RF | 0.81 | 0.32 | 0.85 | 0.46 | 0.85 | 0.89 |
| | | GB | 0.81 | 0.28 | 0.82 | 0.42 | 0.82 | 0.88 |
| | | XGB | 0.75 | 0.26 | 0.80 | 0.38 | 0.80 | 0.86 |
| | | ANN | 0.61 | 0.25 | 0.84 | 0.36 | 0.82 | 0.76 |
| | | GLM | 0.81 | 0.16 | 0.63 | 0.27 | 0.65 | 0.79 |
| | Over sampling | RF | 0.50 | 0.63 | 0.97 | 0.56 | 0.94 | 0.91 |
| | | GB | 0.54 | 0.63 | 0.97 | 0.58 | 0.94 | 0.90 |
| | | XGB | 0.57 | 0.42 | 0.93 | 0.49 | 0.90 | 0.86 |
| | | ANN | 0.67 | 0.21 | 0.77 | 0.32 | 0.76 | 0.75 |
| | | GLM | 0.77 | 0.17 | 0.66 | 0.28 | 0.67 | 0.79 |
| | Cost-sensitive learning | RF | 0.82 | 0.31 | 0.84 | 0.46 | 0.84 | 0.90 |
| | | GB | 0.57 | 0.63 | 0.97 | 0.60 | 0.94 | 0.90 |
| | | XGB | 0.72 | 0.39 | 0.90 | 0.51 | 0.88 | 0.88 |
| | | ANN | 0.75 | 0.34 | 0.87 | 0.47 | 0.86 | 0.86 |
| | | GLM | 0.82 | 0.17 | 0.64 | 0.28 | 0.65 | 0.79 |
| External validation subset | Under sampling | RF | 0.80 | 0.32 | 0.84 | 0.46 | 0.84 | 0.89 |
| | | GB | 0.83 | 0.30 | 0.82 | 0.44 | 0.82 | 0.89 |
| | | XGB | 0.78 | 0.27 | 0.81 | 0.40 | 0.81 | 0.87 |
| | | ANN | 0.58 | 0.24 | 0.83 | 0.34 | 0.81 | 0.75 |
| | | GLM | 0.75 | 0.16 | 0.63 | 0.26 | 0.64 | 0.77 |
| | Over sampling | RF | 0.53 | 0.66 | 0.98 | 0.59 | 0.94 | 0.91 |
| | | GB | 0.55 | 0.62 | 0.97 | 0.58 | 0.93 | 0.90 |
| | | XGB | 0.60 | 0.45 | 0.93 | 0.51 | 0.90 | 0.87 |
| | | ANN | 0.67 | 0.20 | 0.75 | 0.31 | 0.75 | 0.73 |
| | | GLM | 0.82 | 0.19 | 0.68 | 0.31 | 0.69 | 0.81 |
| | Cost-sensitive learning | RF | 0.84 | 0.32 | 0.83 | 0.46 | 0.83 | 0.91 |
| | | GB | 0.56 | 0.60 | 0.97 | 0.58 | 0.93 | 0.90 |
| | | XGB | 0.74 | 0.38 | 0.89 | 0.51 | 0.88 | 0.88 |
| | | ANN | 0.73 | 0.33 | 0.86 | 0.45 | 0.85 | 0.87 |
| | | GLM | 0.81 | 0.18 | 0.65 | 0.29 | 0.67 | 0.80 |

*AUC*

A measure of the model's ability to distinguish between positive and negative classes across different thresholds. A value of 1 indicates perfect discrimination, 0.5 indicates random guessing.

## results AND DISCUSSION

In this study, all operations were conducted using the KNIME platform and the RDKit library. The workflow encompassed data acquisition, segmentation, molecular descriptor calculation,

preprocessing, feature selection, and the tuning and initialization of five machine learning models through Bayesian optimization. Following the tuning process, the models were employed on the training, testing, and external validation datasets. In the preprocessing stage, descriptors with fixed values and low variability were discarded, and this was succeeded by the removal of descriptors that exhibited high correlation with one another. Ultimately, 47 features with Information Gain greater than 0.001 were selected for model training. For model tuning, Bayesian optimization was utilized to adjust the hyper-parameters of each machine learning model, including RF, GB, XGB, ANN, and GLM. Several performance metrics were computed for the training, test, and external validation subsets to evaluate model efficacy. The process was reiterated three times to mitigate class imbalance through the application of under sampling, oversampling, and cost-sensitive learning strategies. Within the Bayesian optimization framework, the maximum number of iterations was set to 100, the number of warm-up rounds was established, Gamma was configured to 0.25, and the number of candidate evaluations per round was set to 25.

*Model evaluation and comparison*

Table 2 presents the results of performance metrics—Recall/sensitivity, precision, specificity, F1 score, accuracy, and AUC—across three subsets (training, test, external validation) using three approaches (under sampling, oversampling, and cost-sensitive learning). The analysis commences by prioritizing accuracy, subsequently assessing the equilibrium between VT and NVT predictions. For the training subset, both the RF and GB models achieved near-perfect accuracy (100%) when utilizing the oversampling method. In contrast, while implementing cost-sensitive learning, the sensitivity—representing the accuracy of predictions for the minority class—was maintained at over 80% for all models. This implies that cost-sensitive learning offered a more uniform performance when contrasted with the other two techniques. When evaluating the test and external validation subsets, the oversampling method performed the least effectively, with sensitivity rarely exceeding 80%. Conversely, the under sampling approach exceeded 80% sensitivity three times for the test subsets and twice for external validation. Similarly, cost-sensitive learning achieved over 80% sensitivity twice for both the test and external validation subsets. As shown in Table 2, the RF model demonstrated consistent and robust performance, whether employing under sampling or cost-sensitive learning. In particular, RF achieved the highest accuracy (85%) for the test subset with under sampling, yielding a Recall/Sensitivity of 81%, Precision of 32%, Specificity of 85%, F1 Score of 0.46, and AUC of 0.89. The performance of the RF algorithm with cost-sensitive learning, yielding an accuracy of 84%, was comparable to the results obtained through the under-sampling technique, which achieved an accuracy of 85%.

Fig. 2 presents the adjusted F1-score, calculated as the harmonic mean of sensitivity and specificity, for all evaluated models, data subsets, and learning approaches. This metric focuses on the precise classification of predictions, ensuring accuracy for both majority and minority categories. As depicted in Fig. 2, RF model consistently demonstrates superior performance across most scenarios. While GB model exhibits competitive performance on the training data, RF maintains a distinct advantage when applied to the test and external validation sets. With under sampling, the RF achieved an adjusted F1 score of 0.83, while it recorded 0.66 with oversampling and 0.84 when employing angle-sensitive learning.

*Feature importance and explanation*

RF models possess an inherent capability for assessing feature importance, providing valuable insights into the underlying factors driving predictions. As each decision tree is constructed within the forest, the algorithm diligently records the reduction in Gini impurity associated with each feature when it is utilized for node splitting. This impurity reduction serves as a proxy for a feature's contribution to improving the model's ability to differentiate between classes. By averaging these impurity reduction values across all trees in the forest, the RF model derives a robust measure of each feature's overall importance. Features that reliably exhibit a marked decrease in impurity are viewed as particularly significant, as they directly enhance the predictive strength of the model. This intrinsic feature importance evaluation mechanism is a key advantage of RF models, offering valuable insights into the data without the need for additional computational steps or external methods. In Fig. 3, a visual depiction of the inherent feature

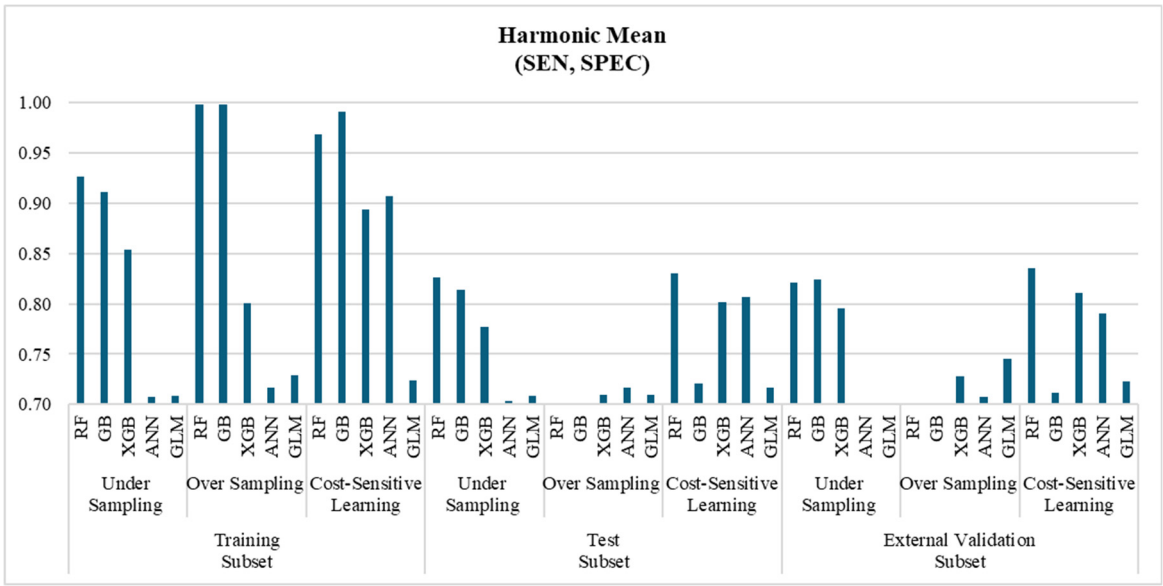


Fig. 2: Adjusted F1-score (harmonic mean of sensitivity and specificity) of all models

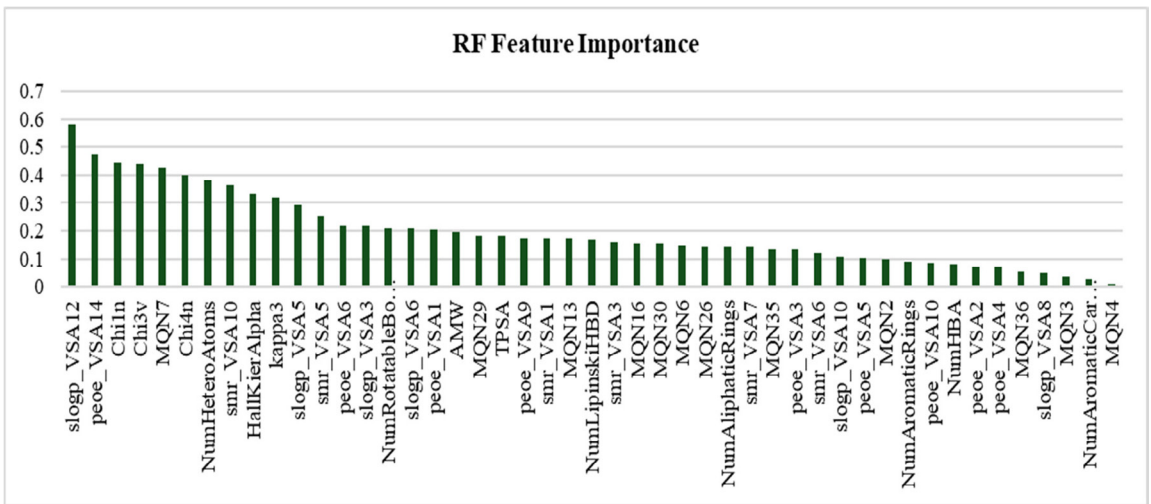


Fig. 3: All predictive descriptors sorted by their splitting frequencies at levels 0, 1, and 2 across the 370 trees comprising the RF model. The top ten molecular descriptors are slogp_VSA12, peoe_VSA14, Chi1n, Chi3v, MQN7, Chi4n, NumHeteroAtoms, smr_VSA10, HallKierAlpha, and kappa3

ranking is presented, with all descriptors arranged according to their splitting frequencies at levels 0, 1, and 2 across the 370 trees comprising the RF model. This hierarchical view illuminates not only the most influential features but also their relative importance at different depths of the decision trees, offering a nuanced understanding of their roles in the model's decision-making process. To elucidate the intricate mechanisms of the Random Forest (RF) model, a combined analysis utilizing PFI, SDT, and LIME was conducted. PFI highlights feature significance by evaluating the extent to which a model's performance deteriorates when the values of a feature are randomly shuffled. This interruption disconnects the

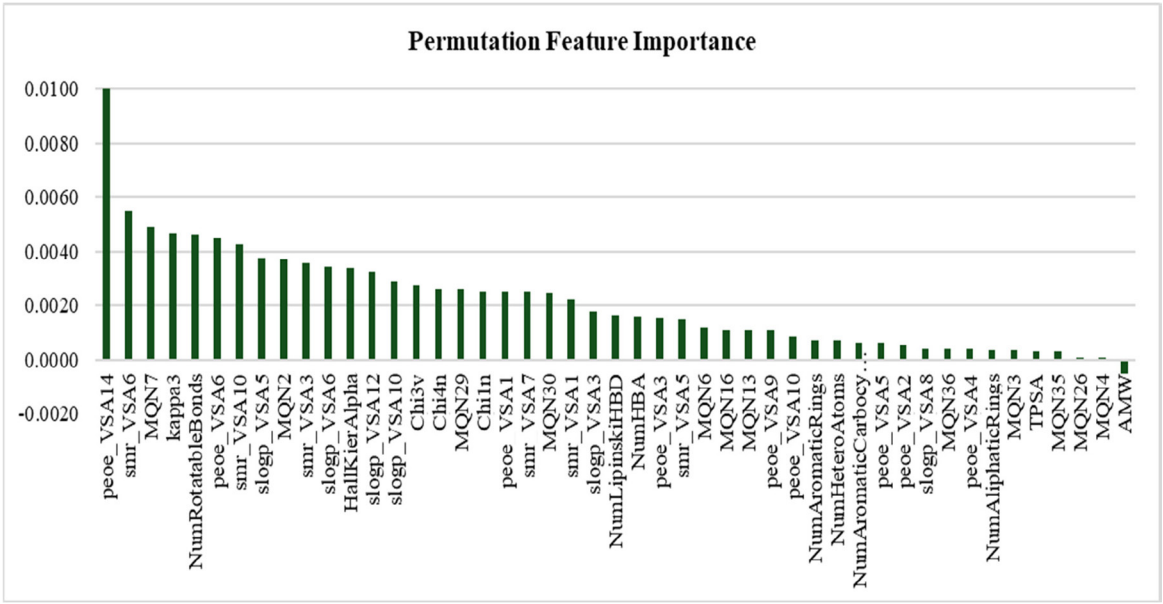


Fig. 4. All predictive descriptors sorted by PFI mean AUC difference of the RF model. The top ten descriptors are peoe_VSA14, smr_VSA6, MQN7, kappa3, NumRotatableBonds, peoe_VSA6, smr_VSA10, slogp_VSA5, MQN2, and smr_VSA3

feature from the target variable, facilitating an analysis of its effect. The PFI process involves calculating the difference in model performance between using all original features and using a dataset where a single feature is randomly permuted. In order to maintain robustness, this method is executed repeatedly for each feature, and the average score variation is determined. In this study, each feature was permuted 20 times, and the PFI was calculated using the AUC, focusing on the "very toxic: VT" class.

A considerable difference in scores signifies that the feature is vital for precise predictions. Shuffling the values of this feature disrupts the model's ability to harness this information, which ultimately results in reduced performance. Conversely, a minimal difference suggests that the feature has little impact, as its permutation does not affect the model's predictive power. A negative difference, though less common, implies that shuffling the feature actually enhances performance. This paradoxical result may point to the possibility that the feature includes redundant or noisy elements, and its removal through permutation contributes to a more optimized model. Fig. 4 presents the PFI results, computed using AUC and focusing on the "VT" class for the RF model with under sampling.

SDT is a simpler Decision Tree that emulates the RF's behavior, was constructed. By visualizing the SDT's decision logic, the insights are gained into how the top PFI-identified features contribute to predictions. This study involved training the SDT on preprocessed data, with an optimization focus on the "Minimum number of records per node." It successfully approximated the original RF model, achieving an AUC of 0.929 for the "VT" class. Fig. 5 illustrates the top six levels of the SDT model, highlighting its impressive AUC of 0.929 in approximating the original RF model. This visualization illustrates the primary molecular descriptors that are pivotal in predicting highly toxic compounds. It is noteworthy that the descriptors slogp_VSA12, peoe_VSA14, and MQN7 rank highly across different evaluation methods: slogp_VSA12 leads in the RF ranking, peoe_VSA14 tops the PFI ranking, and MQN7 is prominent in the SDT ranking. Each descriptor also holds a significant position in the other rankings.

Slogp_VSA12: Indicates the contribution of surface area to the log P (partition coefficient), providing insights into hydrophobicity and lipophilicity. High values suggest significant interaction with biological membranes, potentially leading to toxicity. Peoe_ VSA14: Measures the van der Waals surface area

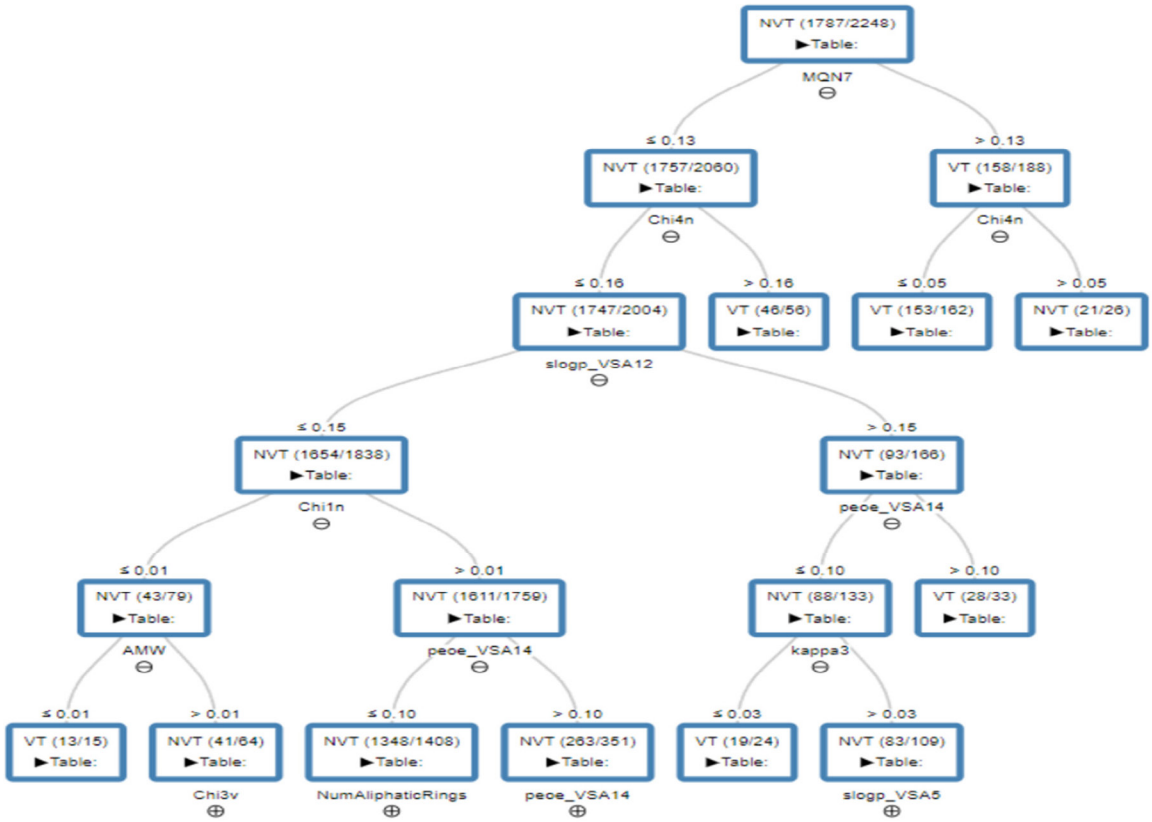


Fig. 5: The top six levels Global SDT model trained successfully with AUC equal to 0.929 to the original RF model. This visualization reveals that the molecular descriptors MQN7, Chi14n, Slogp_VSA12, Chi1n, AMW, poeoe_VSA14, kappa3, Chi3v, NumAliphaticRings, peoe, VSA14, and Slogp_VSA5 are the most influential descriptors for predicting VT compounds

of atoms with specific partial charges, reflecting the molecule's electrostatic potential. Grasping the interactions with biological targets is important, as these interactions play a significant role in determining toxicity via receptor binding and enzyme function. MQN7: Represents the number of nitrogen atoms in a molecule. Nitrogen-containing substances are known for their specific biological activities and can be related to various toxicological impacts. This relationship can alter their molecular structure and reactivity, potentially heightening their toxicity.

LIME is the third XAI method used in this study. It serves as a method to elucidate the predictions of sophisticated machine learning models on a case-by-case basis. This approach functions by modifying the input characteristics of a given instance and monitoring how the predictions of the model are affected. This allows LIME to identify the features that most influenced the prediction for that particular instance, providing local explanations. In the context of acute oral toxicity prediction, LIME can reveal which molecular descriptors are driving the model's assessment of a chemical's toxicity. Understanding this information is vital for grasping the model's decision-making process, enhancing trust in its predictions, and revealing insights into the molecular attributes related to toxicity. Fig. 6 shows the top ten importance descriptors for single true positive, true negative, false positive and false negative test records.

Fig. 6 show that four commonly important descriptors:

1. **MQN7**: The standout feature here is MQN7 (Number of nitrogen atoms), appearing in the top ten for all four prediction categories (TP, TN, FP, FN). This implies that nitrogen atoms significantly

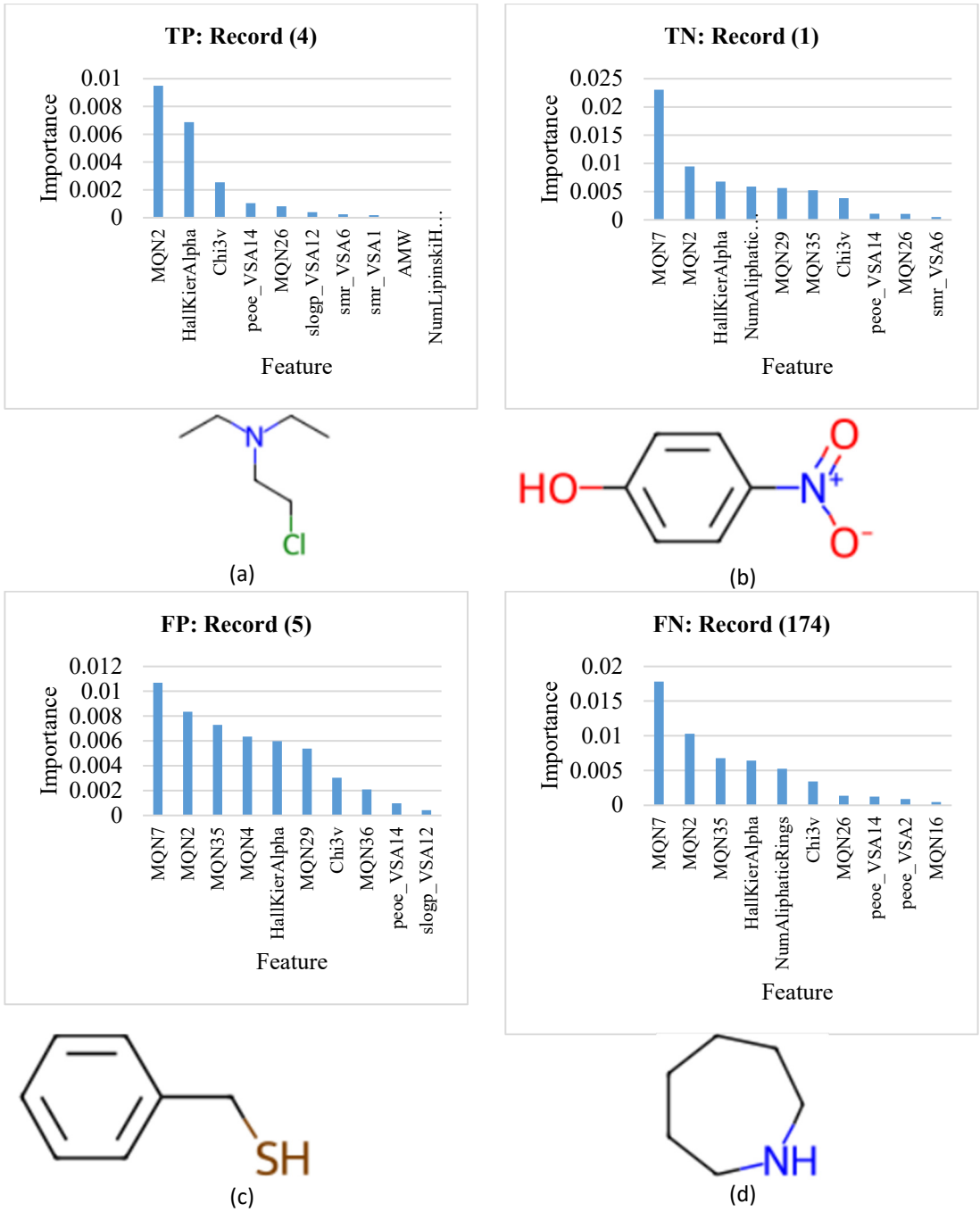


Fig. 6: LIME explanations for four individual records, including (a) true positive (TP), (b) true negative (TN), (c) false positive (FP), and (d) false negative (FN), along with the 2D structures of each record

influence the model's toxicity assessment, regardless of the correctness of the prediction. It's a feature worth investigating further in terms of its chemical significance in relation to acute oral toxicity.

2. **MQN2**: Also consistently important is MQN2, present in TP, TN, and FP. This suggests a

more general role in the model's decision-making, potentially related to structural characteristics influenced by nitrogen presence.

3. **HallKierAlpha**: Appearing in TP, TN, and FN, this descriptor related to molecular shape and branching seems to be another consistent factor influencing predictions.

4. **Chi3v**: This valence connectivity index appears in TP, TN, and FP. Like HallKierAlpha, it suggests a structural influence on toxicity assessment.

Analysis of discrepancies between correct (TP/TN) and incorrect (FP/FN) predictions using LIME reveals potential areas for model improvement. False positives (FP) may arise from overemphasis of polar (peoe_VSA14) or lipophilic (slogp_VSA12) surface areas, while false negatives (FN) might be due to underestimation of toxicity in compounds with specific aliphatic ring structures (NumAliphaticRings) or branching patterns (HallKierAlpha). A deeper exploration of the chemical importance of essential descriptors, along with enhancements to the model and the addition of data, can enhance both the accuracy of the model and the comprehension of toxicity prediction. This knowledge can inform chemical design by guiding the creation of safer compounds and enhance regulatory assessments by providing tools for evaluating new chemicals before market entry.

## Conclusion

This study successfully demonstrates the potential of machine learning, specifically the RF model, for accurately predicting acute oral toxicity, even in the presence of imbalanced datasets. The comprehensive approach taken in this study, which integrated data preprocessing, feature selection, hyper-parameter optimization using Bayesian Optimization with cross-validation, and an analysis of class balancing methods, facilitated the development of a predictive model that is both balanced and resilient. The best-performing RF model, coupled with under sampling, achieved an accuracy of 0.85, a sensitivity of 0.81, and a specificity of 0.85 on the independent test set, demonstrating its ability to reliably identify both VT and NVT compounds. The RF models' inherent capability to evaluate feature importance was utilized, offering valuable insights into the key factors influencing the predictions. Furthermore, the use of explainable AI techniques, particularly PFI and SDT, shed light on the underlying decision-making mechanisms of the model. The top-ranked descriptors, slogp_VSA12, peoe_VSA14, and MQN7, consistently appear across all feature importance analyses, highlighting their crucial roles in predicting toxicity. The Slogp_VSA12 value, representing a molecule's hydrophobic nature, suggests that greater engagement with biological membranes may be a factor in its toxic effects. Peoe_VSA14, capturing electrostatic potential, highlights the importance of interactions with biological targets like receptors and enzymes. Lastly, MQN7, representing nitrogen atom count, emphasizes the potential role of nitrogen-containing functional groups in toxicological effects. This research contributes to the growing body of evidence supporting the use of machine learning for accelerating and enhancing chemical safety assessment. Harnessing machine learning technology can facilitate a shift towards more efficient, cost-effective, and ethical practices, minimizing the need for traditional animal testing. The insights gleaned from this study, particularly the identification of key molecular descriptors and the interpretability offered by XAI techniques, can inform safer chemical design strategies and enhance regulatory decision-making processes, ultimately contributing to the realization of SDGs 3, 6, and 12. To advance the field of predictive toxicology and ensure a safer, more sustainable future for chemical development and application, it is vital to further investigate these findings. This includes expanding the model's applicability domain, utilizing 3D structural information, and integrating an understanding of toxicity mechanisms.

## AUTHOR CONTRIBUTIONS

A.M. Elsayad was responsible for conceptualization, methodology, writing the original draft, and supervision. M. Zeghid handled data curation, formal analysis, and writing - review and editing. K.A. Elsayad conducted investigation, software development, and visualization. A. N. Khan contributed resources, validation, and writing - review and editing. A.M. Baareh managed project administration, funding acquisition, and writing - review and editing. A. Sadiq performed data curation and formal analysis. S.A. Mukhtar worked on software development and visualization. H. F. Ali provided resources and validation. S. Abd El-kader carried out investigation and writing - review and editing.

## ACKNOWLEDGEMENT

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number [PSAU-2023/SDG/28].

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

## OPEN ACCESS

©2025 The author(s). This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third-party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit: http://creativecommons.org/licenses/by/4.0/

## PUBLISHER'S NOTE

GJESM Publisher remains neutral with regard to jurisdictional claims in published maps and institutional afflictions.

## ABBREVIATIONS

| | |
|---|---|
| *%* | Percent |
| *2D* | Two dimensional |
| *ADT* | Acute dermal toxicity |
| *AI* | Artificial intelligence |
| *AUC* | The area under the curve |
| *ATWG* | Acute toxicity workgroup |
| *ANN* | Artificial neural networks |
| *CAS-RN* | Chemical abstracts service registry number |
| *DL* | Deep learning |
| *EPA* | Environmental Protection Agency |
| *EGNNs* | Equivariant graph neural networks |
| *FP* | False positive |
| *FN* | False negative |
| *GHS* | Globally harmonized system |
| *GB* | Gradient boosting |
| *GLM* | Generalized linear model |
| *ICCVAM* | Interagency coordinating committee on the validation of alternative methods |
| *IG* | Information gain |
| *KNIME* | Konstanz information miner |
| *LD50* | Lethal dose 50 |
| *NTP* | National toxicology program |
| *NICEATM* | NTP Interagency Center for the Evaluation of Alternative Toxicological Methods |
| *NCCT* | National Center for Competency Testing |
| *NVT* | Not very toxic |
| *ML* | Machine learning |
| *PFI* | Permutation feature importance |
| *RDKit* | Rational discovery kit |
| *RF* | Random forests |
| *SDGs* | Sustainable development goals |
| *SDT* | Surrogate decision tree |
| *SMOTE* | Synthetic minority over-sampling technique |
| *TP* | True positive |
| *TN* | True negative |
| *VT* | Very toxic |
| *XGB* | Extreme gradient boost |
| *SMILES* | Simplified molecular input line entry system |

## REFERENCES


Abou Hajal, A.; Al Meslamani, A.Z., (2024). Overcoming barriers to machine learning applications in toxicity prediction. Expert Opin. Drug Metab. Toxicol., 20(7): 549-553 (**5 pages**).

Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G., (2021). A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev., 54: 1937-1967 (**41 pages**).

Born, J.; Markert, G.; Janakarajan, N.; Kimber, T.B.; Volkamer, A.; Rodríguez Martínez, M.; Manica, M., (2023). Chemical representation learning for toxicity prediction. Digital Discovery. 2(3): 674-691 (**34 pages**).

Bharti, D.R.; Lynn, A.M., (2017). QSAR based predictive modeling for anti-malarial molecules. Bioinformation. 13(5): 154 (**6 pages**).

Creton, S.; Dewhurst, I.C.; Earl, L. K.; Gehen, S.C.; Guest, R. L.; Hotchkiss, J.A.; Indans, I.; Woolhiser, M.R.; Billington, R., (2010). Acute toxicity testing of chemicals—opportunities to avoid redundant testing and use alternative approaches. Crit. Rev. Toxicol., 40(1): 50-83 (**34 pages**).

Cavasotto, C.N.; Scardino, V., (2022). Machine learning toxicity prediction: latest advances by toxicity end point. ACS Omega. 7(51): 47536-47546 (**11 pages**).

Chinedu, E.; Arome, D.; Ameh, F.S., (2013). A new method for determining acute toxicity in animal models. Toxicol. Int., 20(3): 224-226 (**3 pages**).

Cremer, J.; Medrano Sandonas, L.; Tkatchenko, A.; Clevert, D.-A.; De Fabritiis, G., (2023). Equivariant graph neural networks for toxicity prediction. Chem. Res. Toxicol., 36(10): 1561-1573 (**13 pages**).

Durrant, S.; Hardoon, D.R.; Brechmann, A.; Shawe-Taylor, J.; Miranda, E.R.; Scheich, H., (2009). GLM and SVM analyses of neural response to tonal and atonal stimuli: new techniques and a comparison. Connect. Sci., 21(2-3): 161-175 (**15 pages**).

Feng, H.; Zhang, L.; Li, S.; Liu, L.; Yang, T.; Yang, P.; Zhao, J.; Arkin, I.T.; Liu, H., (2021). Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints. Toxicol. Lett., 340: 4-14 (**10 pages**).

Fan, T.; Sun, G.; Zhao, L.; Cui, X.; Zhong, R., (2018). QSAR and classification study on prediction of acute oral toxicity of N-nitroso compounds. Int. J. Mol. Sci., 19(10): 3015 (**22 pages**).

Guo, W.; Liu, J.; Dong, F.; Song, M.; Li, Z.; Khan, M.K.H.; Patterson, T. A.; Hong, H., (2023). Review of machine learning and deep learning models for toxicity prediction. Exp. Biol. Med., 248(21): 1952-1973 (**22 pages**).

Hasib, K.M.; Showrov, M.I.H.; Mahmud, J.A.; Mithu, K., (2022). Imbalanced data classification using hybrid under-sampling with cost-sensitive learning method. In: Edge Analytics: Select Proceedings of 26th International Conference—ADCOM, Singapore. 423-435 (**13 pages**).

Jiang, J.; Wang, R.; Wei, G.-W., (2021). GGL-Tox: geometric graph learning for toxicity prediction. J. Chem. Inf. Model., 61(4): 1691-1700 (**10 pages**).

Kleinstreuer, N.C.; Karmaus, A.L.; Mansouri, K.; Allen, D.G.; Fitzpatrick, J.M.; Patlewicz, G., (2018). Predictive models for acute oral systemic toxicity: a workshop to bridge the gap from research to regulation. Comput. Toxicol., 8: 21-24 (**4 pages**).

Lou, S.; Yu, Z.; Huang, Z.; Wang, H.; Pan, F.; Li, W.; Liu, G.; Tang, Y., (2024). In silico prediction of chemical acute dermal toxicity using explainable machine learning methods. Chem. Res. Toxicol., 37(3): 513-524 (**11 pages**).

Landrum, G., (2013). Rdkit documentation. Release 2019.09.1 : 1-151 (**151 pages**).

Mansouri, K.; Karmaus, A. L.; Fitzpatrick, J.; Patlewicz, G.; Pradeep, P.; Alberga, D.; Alepee, N., et al., (2021). CATMoS: collaborative acute toxicity modeling suite. Environ. Health Perspect., 129(4): 047013 (**18 pages**).

More, A.S.; Rana, D.P., (2017). Review of random forest classification techniques to resolve data imbalance. In: 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM): 72-78 (**7 pages**).

Muschalik, M.; Fumagalli, F.; Hammer, B.; Hüllermeier, E., (2022). Agnostic explanation of model change based on feature importance. KI-Künstliche Intelligenz., 36(3): 211-224 (**14 pages**).

Mateo, J.; Rius-Peris, J.M.; Maraña-Pérez, A.I.; Valiente-Armero, A.; Torres, A. M., (2021). Extreme gradient boosting machine learning method for predicting medical treatment in patients with acute bronchiolitis. Biocybern. Biomed. Eng., 41(2): 792-801 (**10 pages**).

Pérez Santín, E.; Rodríguez Solana, R.; González García, M.; García Suárez, M. D. M.; Blanco Díaz, G. D.; Cima Cabal, M. D.; Moreno Rojas, J. M.; López Sánchez, J. I., (2021). Toxicity prediction based on artificial intelligence: A multidisciplinary overview. Wiley Interdisciplinary Reviews: Comput. Mol. Sci., 11(5): (**32 pages**).

Ryu, J.Y.; Jang, W.D.; Jang, J.; Oh, K.-S., (2023). PredAOT: a computational framework for prediction of acute oral toxicity based on multiple random forest models. BMC Bioinf., 24(1): 66 (**10 pages**).

Sachs, J.D.; Schmidt-Traub, G.; Mazzucato, M.; Messner, D.; Nakicenovic, N.; Rockström, J., (2019). Six transformations to achieve the sustainable development goals. Nat. Sustainability. 2(9): 805-814 (**10 pages**).

Tougui, I.; Jilbab, A.; El Mhamdi, J., (2021). Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. Healthcare Inf. Res., 27(3): 189-199 (**11 pages**).

Tran, T.T.V.; Wibowo, A.S.; Tayara, H.; Chong, K.T., (2023). Artificial intelligence in drug toxicity prediction: recent advances, challenges, and future perspectives. J. Chem. Inf. Model., 63(9): 2628-2643 (**16 pages**).

Wu, S.; Li, S.-X.; Qiu, J.; Zhao, H.-M.; Li, Y.-W.; Feng, N.-X.; Liu, B.-L., et al., (2024). Accurate prediction of rat acute oral toxicity and reference dose for thousands of polycyclic aromatic hydrocarbon derivatives based on chemometric QSAR and machine learning. Environ. Sci. Technol., 58(34): 15100-15110 (**11 pages**).

Wilhelm, A.; Zweig, K. A., (2024). Hacking a surrogate model approach to XAI. ArXiv preprint ArXiv:2406.16626.:1-24 (**24 pages**).

Wu, J.; Chen, X.Y.; Zhang, H.; Xiong, L.D.; Lei, H.; Deng, S.H., (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. J. Electron. Sci. Technol., 17(1): 26-40 (**15 pages**).

Xiao, L.; Deng, J.; Yang, L.; Huang, X.; Yu, X., (2022). Random forest algorithm-based accurate prediction of rat acute oral toxicity. Mol. Phys., 120(24): e2140083 (**7 pages**).

Zhang, Z., (2016). A gentle introduction to artificial neural networks. Ann. Transl. Med., 4(19):370-375 (**6 pages**).

**AUTHOR (S) BIOSKETCHES**

**Elsayad, A.M.**, Ph.D., Professor, Department of Electrical Engineering, College of Engineering in Wadi Alddawasir, Prince Sattam Bin Abdulaziz University, Wadi Alddawasir 11991, Saudi Arabia.
- Email: *a.elsayyad@psau.edu.sa*
- ORCID: 0000-0001-8053-9759
- Web of Science ResearcherID: ABI-1217-2020
- Scopus Author ID: 6508294022
- Homepage: https://faculty.psau.edu.sa/en/psau/facultymember/a.elsayyad

**Zeghid, M.**, Ph.D., Associate Professor, Computer Engineering and Networks Department, College of Engineering, Psau, Saudi Arabia.
- Email: *m.zeghid@psau.edu.sa*
- ORCID: 0000-0001-8217-3455
- Web of Science ResearcherID: HKV-1851-2023
- Scopus Author ID: 19934668700
- Homepage: https://faculty.psau.edu.sa/en/psau/facultymember/m.zeghid

**Elsayad, K.A**., M.Sc. Student, Pharmacy Department, Cairo University Hospitals, Cairo University, Cairo 11662, Egypt.
- Email: *khaled.al.elsayad@std.pharma.cu.edu.eg*
- ORCID: 000-0002-9179-0460
- Web of Science ResearcherID: NA
- Scopus Author ID: 57220804962
- Homepage: https://scholar.google.com/citations?user=y1zj8bEAAAAJ&hl=en

**Khan, A.N.,** Ph.D., Associate Professor, Department of Electrical Engineering, Jalozai Campus, UET, Peshawar, Pakistan.
- Email: *akhtar.nawaz@uetpeshawar.edu.pk*
- ORCID: 0000-0002-6323-5638
- Web of Science ResearcherID: HRC-5787-2023
- Scopus Author ID: 55744978100
- Homepage: https://scholar.google.com.br/citations?user=aNlC-CgAAAAJ&hl=th

**Baareh, A.K.M.,** Ph.D., Department of Applied Science, Ajloun University College, Al-Balqa Applied University, Ajloun, Jordan.
- Email: *baareh@bau.edu.jo*
- ORCID: 0000-0002-5562-4043
- Web of Science ResearcherID: IQT-4005-2023
- Scopus Author ID: 15077827900
- Homepage: https://scholar.google.com/citations?user=R2no5bMAAAAJ&hl=ar

**Sadiq, A.**, Ph.D., Associate Professor, Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan.
- Email: *sadiqali@uetpeshawar.edu.pk*
- ORCID: 0000-0002-2329-3649
- Web of Science ResearcherID: NA
- Scopus Author ID: 55861525000
- Homepage: https://scholar.google.co.uk/citations?user=YkSI7kMAAAAJ&hl=en

**Mukhtar, S.A.,** Ph.D., Associate Professor Computers and Systems Department, Electronics Research Institute, Cairo, Egypt.
- Email: *sahar@eri.sci.eg*
- ORCID:0000-0002-9560-5525
- Web of Science ResearcherID: NA
- Scopus Author ID: 55480537100
- Homepage: https://scholar.google.com.eg/citations?user=1Jt0UDIAAAAJ&hl=en

**Ali, H.F.,** Ph.D., Associate Professor**,** Computers and Systems Department, Electronics Research Institute, Cairo, Egypt.
- Email: *Hesham.Farouk.ali@gmail.com*
- ORCID: 0000-0002-7132-3881
- Web of Science ResearcherID: NA
- Scopus Author ID: 56658455800
- Homepage: https://www.itu.int/en/ITU-D/Regional-Presence/ArabStates/Documents/events/2017/IoTSMW/Bios/Dr%20Hesham%20farouk%20resume.pdf

**Abd El-kader, S.,** Ph.D., Professor**,** Computers and Systems Department, Electronics Research Institute, Cairo, Egypt.
- Email: *sherine@eri.sci.eg*
- ORCID: 0000-0001-5561-2424
- Web of Science ResearcherID: W-8095-2019
- Scopus Author ID: 46560948500
- Homepage: https://eri.sci.eg/staff/prof-dr-sherine-mohamed-abd-el-kader-abd-el-raoof/

**HOW TO CITE THIS ARTICLE**

*Elsayad, A.M.; Zeghid, M.; Elsayad, K.A.; Khan, A. N.; Baareh, A.M.; Sadiq, A.; Mukhtar, S.A.; Ali, H.F.; Abd El-kader, S., (2025). Machine learning model for random forest acute oral toxicity prediction. Global J. Environ. Sci. Manage., 11(1): 21-38.*

**DOI:** 10.22034/gjesm.2025.01.02

**URL:** https://www.gjesm.net/article_718277.html