Journal of Research in Pharmacy

Research Article

www.jrespharm.com

An explainable prediction model for drug-induced interstitial pneumonitis

Feyza KELLECİ ÇELİK[1]*, Sezen YILMAZ SARIALTIN[2]

1 Vocational School of Health Services, Karamanoglu Mehmetbey University, Karaman, Türkiye.
2 Department of Pharmaceutical Toxicology, Faculty of Pharmacy, Ankara University, Ankara, Türkiye.

* Corresponding Author. E-mail: feyza-kelleci@hotmail.com (F.K.Ç); Tel. +90-0338-226 21 92.

Received: 1 December 2024 / Revised: 14 December 2024 / Accepted: 17 December 2024

ABSTRACT:

Drug-induced interstitial pneumonitis (DIP) is an inflammation of the lung interstitium, emerging due to the pneumotoxic effects of pharmaceuticals. The diagnosis is challenging due to nonspecific clinical presentations and limited testing. Therefore, identifying the risk of drug-related pneumonitis is required during the early phases of drug development. This study aims to estimate DIP using binary quantitative structure-toxicity relationship (QSTR) models. The dataset was composed of 468 active pharmaceutical ingredients (APIs). Five critical modeling descriptors were chosen. Then, four machine-learning (ML) algorithms were conducted to build prediction models with the selected molecular identifiers. The developed models were validated using the internal 10-fold cross-validation and external test set. The Logistic Regression (LR) algorithm outperformed all other models, achieving 95.72% and 94.68% accuracy in internal and external validation, respectively. Additionally, the individual effect of each descriptor on the model output was determined using the SHapley Additive exPlanations (SHAP) approach. This analysis indicated that the pneumonitis effects of drugs might predominantly be attributed to their atomic masses, polarizabilities, van der Waals volumes, surface areas, and electronegativities. Apart from the strong model performance, the SHAP local explanations can assist molecular modifications to reduce or avoid the risk of pneumonitis for each molecule in the test set. Contributing to the drug safety profile, the current classification model can guide advanced pneumotoxicity testing and reduce late-stage failures in drug development.

KEYWORDS: Pulmonary toxicity; computational toxicology; QSTR; QSAR; machine learning; SHapley Additive exPlanations.

1. INTRODUCTION

Drug-induced interstitial lung disease (DI-ILD), a subclass of diffuse parenchymal lung disease [1], is caused by the pulmonary side effects of pharmaceuticals [2]. This pharmacotherapy-related complication begins with inflammation of the lung interstitium (pneumonitis) and progresses to fibrosis with prolonged use of the causative drug [3]. DI-ILD can present with a wide range of clinical signs from mild respiratory complications to rapidly progressive respiratory failure and death [4]. Without early intervention, complications such as pneumothorax, pulmonary hypertension, lung cancer, and respiratory failure may occur [5].

Over 350 drugs have been reported to cause pneumonitis [3], with the most common being amiodarone, gefitinib, lenvatinib, nivolumab, and nitrofurantoin [6]. Approximately 70% of drug-induced interstitial pneumonitis (DIP) cases are associated with systemic cancer treatment, and the early mortality rate is high in these patients. Among 770 patients diagnosed with advanced-stage lung cancer, pneumonitis developed in 6% of cases during systemic chemotherapy with an associated mortality rate of 36% [7]. The clinical, laboratory, histological, and radiological findings of acute/chronic DIP are variable and non-specific [3, 4]. Prompt diagnosis followed by cessation of the causative drug may improve the prognosis by suppressing inflammation in the lungs; however, once fibrosis begins to develop, this process becomes irreversible [3].

How to cite this article: Kelleci Çelik F, Yılmaz Sarıaltın S. An explainable prediction model for drug-induced interstitial pneumonitis. J Res Pharm. 2025; 29(1): 322-334.

ISSN: 2630-6344

https://doi.org/10.12991/jrespharm.1644357

The pneumotoxicity risk of drugs is typically detected at later stages of drug development or during post-launch safety monitoring [1]. Thus, various open-source web-based platforms like the Pneumotox [6] and the SIDER 4.5 [8], were designed to provide healthcare professionals with comprehensive information on drug-related lung injury. Although these databases are regularly updated using case reports and current literature data [1, 8], the risk of new drug candidates remains. Considering its fulminant progression, timely diagnosis of DIP is critical for the safety of pharmacotherapy [6].

Computational approaches such as quantitative structure-toxicity relationships (QSTRs) are currently adopted in pharmaceutical toxicology research due to their advantages in saving time, reducing costs, and avoiding ethical concerns. Machine learning (ML) algorithms constructed in QSTR models offer fast and robust approaches to toxicity prediction [9]. In the literature, the available prediction models have typically focused on the comprehensive respiratory toxicity profile of chemicals [2, 9, 10]. Contrary to these studies, this research focused solely on pharmaceuticals and a specific endpoint, such as pneumonitis.

In this study, binary QSTR models for DIP were built using 468 active pharmaceutical ingredients (APIs) provided by the Pneumotox database [6], the Food and Drug Administration (FDA) official website [11], and the SIDER 4.5 website [8]. Classification models using different ML algorithms were built based on five selected features related to the target toxic effect. They can serve as an early detection tool in drug development or before clinical trials to evaluate the pneumonitis risk of drug candidates or available pharmaceuticals. Additionally, a game theory-based SHapley Additive exPlanations (SHAP) method [12] was applied to interpret the highest-performing model and to prioritize the importance of key descriptors influencing the prediction outcomes. The direction of the identifiers for each specific sample in the test set was also determined. These mechanistic insights into molecules can guide the molecular optimization aimed at reducing or avoiding DIP. Our computational approaches like ours for predicting drug toxicity profiles and optimizing drug development processes can enhance healthcare quality while minimizing economic costs.

2. RESULTS

2.1. Feature extraction

In this study, PaDEL software [13] was used as the descriptor calculation tool. The calculated 1444 2D descriptors were decreased to 1222 using the filters Waikato Environment for Knowledge Analysis (WEKA 3.9.5) to remove ineffective characteristics on the model performance [14]. After this process, the two-stage feature elimination strategy was performed to determine the optimal subset of descriptors responsible for the target toxic effect. Initially, the best five 2D descriptors were selected from the remaining 1222 descriptors using the CfsSubsetEval filter+BestFirst search method of WEKA 3.9.5 [14]. In the next step, the Correlation Heatmap was utilized to analyze the inter-correlation matrix of molecular descriptors (Figure 1). A strong correlation criterion of 0.7 or higher is often used in this analysis [15, 16]. As illustrated in Figure 1, since the correlation between the descriptors in our study is quite weak, no descriptors were removed. As a result, we developed the prediction model using the following 5 optimal descriptors AATS0s, ETA_Shape_X, AATSC0p, AVP-3, and SssssNp to maximize the modeling success (Table 1).

Table 1. The molecular descriptors of the high-performance model.

Descriptor	Description	Descriptor Class
AATS0s	Average Broto-Moreau autocorrelation - lag 0 / weighted by I-state	Autocorrelation
ETA_Shape_X	Shape index X	Extended Topochemical Atom
AATSC0p	Average centered Broto-Moreau autocorrelation - lag 0 / weighted by polarizabilities	Autocorrelation
AVP-3	Average valence path, order 3	Chi Path
SssssNp	Sum of atom-type E-State: >N<+	Electrotopological State Atom Type

2.2. Performances of the models

In the present study, Logistic Regression (LR) [17], Naïve Bayes (NB) [18], k-Nearest Neighbor with Stochastic Search (KStar) [19], and Instance-Based Learning with k-Nearest Neighbors (IBk) [20] algorithms were used to construct QSTR models. The confusion matrix for the model is shown in Table 2. Internal and

external validation results were analyzed to assess the performance of this model by calculating accuracy (ACC), specificity (SP), sensitivity (SE), F-measure, Matthews correlation coefficient (MCC), and area under the receiver operating characteristic curve (ROC) metrics.

Table 2. Confusion matrix for the model

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

As indicated in Table 3, the LR model showed the best performance among other models in terms of internal and external validation (95.72% and 94.68%, respectively). Furthermore, this model outperformed the other models across all metrics on the training set, while providing superiority on all metrics except the ROC metric (0.953) in the test set. Besides, the Topliss ratio was calculated as 74.8 (374 compounds/5 descriptors), which supported the validity of the models.

Table 3. Performance of QSTR models after eliminating attributes (n_T=468)

Classifiers	Validation Sets	ACC %	SP	SE	F-Measure	MCC	ROC
LR	Training set	**95.72**	**0.959**	**0.957**	**0.957**	**0.914**	**0.988**
	Test set	**94.68**	**0.947**	**0.947**	**0.947**	**0.893**	0.953
NB	Training set	94.92	0.950	0.949	0.949	0.897	0.985
	Test set	85.11	0.859	0.851	0.849	0.704	**0.955**
KStar	Training set	93.85	0.942	0.939	0.938	0.878	0.953
	Test set	89.37	0.897	0.894	0.894	0.790	0.952
IBk	Training set	91.98	0.920	0.920	0.920	0.836	0.921
	Test set	82.98	0.842	0.830	0.826	0.664	0.816

LR: Logistic Regression; NB: Naïve Bayes: KStar: k-Nearest Neighbors with Star Schema, IBk: Instance-Based Learning with k-Nearest Neighbors, ACC: Classifier accuracy, SP: Specificity, SE: Sensitivity, MCC: Matthews correlation coefficient, ROC: Area under the receiver operating characteristic curve; n_T: total number of compounds

2.3. Applicability domain

The applicability domain (AD) indicates the chemical space within which the model's estimations are reliable [21]. To ensure the robustness of the model, we conducted the Tanimoto similarity index [22] and chemical space distribution analyses [21]. In the current study, the average Tanimoto scores were found 0.3302 for the training set and 0.3300 for the test set, indicating chemical diversity in the dataset and AD compatibility. Molecular weight (MW) and Ghose-Crippen LogKow (ALogP) values were used for chemical space distribution analysis (Figure 2). The MW values of the molecules in the dataset ranged from 74.9216 to 2678.4796 g/mol, while their ALogP values were from -27.4706 to 8.9350. This visualization verified that the chemical domain covered by the training set adequately included the components in the test set.

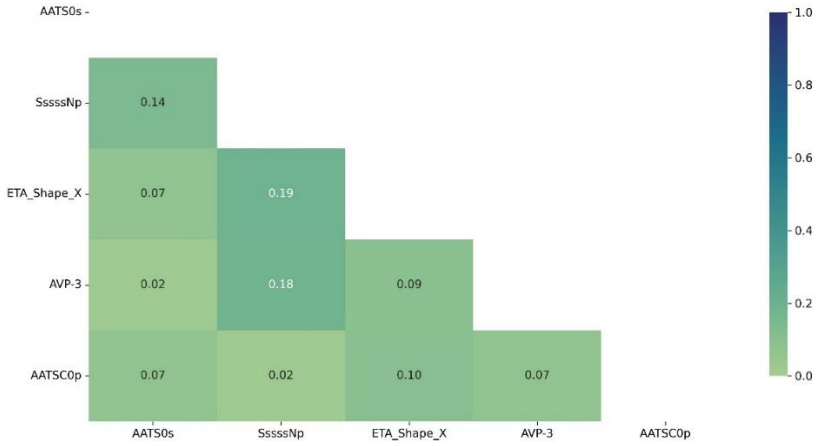

Figure 1. Correlation Heatmap of molecular descriptors

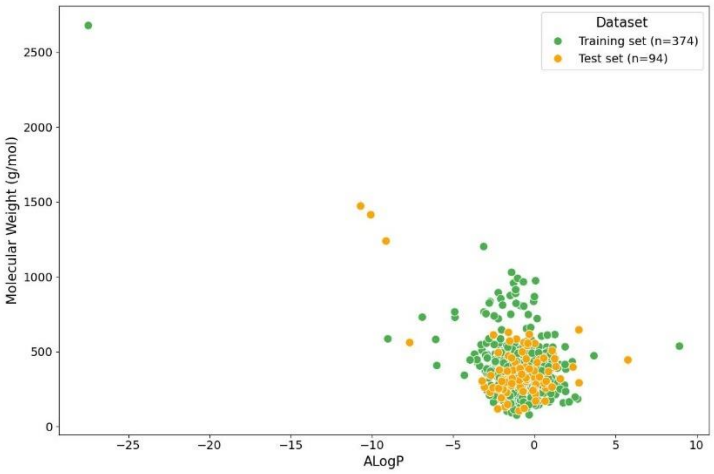

Figure 2. Exploring chemical diversity in the training and test sets using Molecular weight (MW) and Ghose-Crippen LogKow (AlogP) to define chemical space (n: number of compounds)

2.4. Explainability of the best-performing model

2.4.1. Explanation of features at the global level

SHAP summary plots were used for a global explanation of the best-performing LR model (Figure 3). Each point in the graph represents a compound from the external test set. Descriptors are ranked in descending order of overall effect size on the vertical (Y) axis. The horizontal (X) axis exhibits the values indicating the effect of the identifiers on the model outcomes (toxic or non-toxic class). Positive SHAP values increase the probability of a molecule being assigned to the toxic class, while negative ones reduce it. The color scale stands for value range of the descriptor (red for high values - blue for low values) [2].

The most dominant identifier for the LR model is AATS0s followed by ETA_Shape_X, contributing slightly less. According to the global SHAP graph, AATS0s and ETA_Shape_X features showed anti-correlation with the target toxic effect. As the values of these descriptors decrease, they correspond to a higher positive SHAP value and ultimately, a greater toxicity potential. Increasing values of AVP-3 contribute to the toxic effect, while SssssNp contributes less, although both have low effects on model performance. The effect of the AATSC0p descriptor remained undefined under the global explanations. These results are a general evaluation of the model and are mainly used to determine the importance of the descriptors. To reach a detailed explanation for each molecule, the local SHAP results are examined separately.

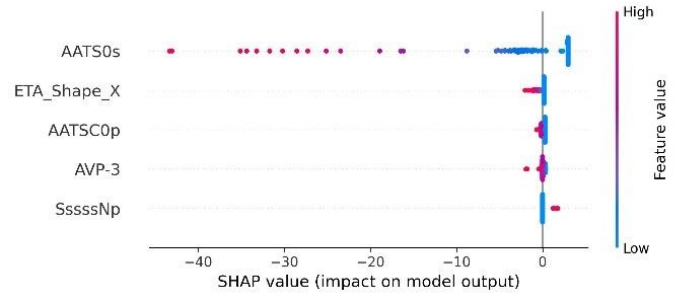

Figure 3. SHAP summary plots for the global explanation of the model

2.4.2. Explanation of features at the local level

The SHAP waterfall plots illustrated the relevance and direction of the descriptors for the entire test set (Supplementary File 2_Table S1). These graphs provided detailed insight into which features for each compound in the test set were critical for pneumonitis risk prediction. Figure 4 presents the SHAP waterfall plots of two specific molecules randomly chosen from the external set.

In SHAP waterfall graphs, the horizontal axis indicates the magnitude and direction which each identifier deviates from the expected value (E [f(x)]) to the final model forecast. The expected value at the bottom of the graph is the model's estimate when no descriptor value is available for a particular compound, also referred to as the baseline value. The descriptors, along with their original values, are arranged on the y-axis in decreasing order. The SHAP values shown above the horizontal bars in the graph represent the contribution of each variable to the prediction. The colours of the bars indicate the direction of this effect.

In the high-quality LR model, AATS0s played a strong role in the prediction, ranking first place for 90 molecules in the test set. For 51 of these molecules, increasing values of AATS0s were associated with an increased toxic effect, while the opposite was observed for 39 molecules. The second common descriptors for molecules were found in AATSC0p (43 compounds) and ETA_Shape_X (39 compounds). The effects and prevalence of SssssNp and AVP-3 descriptors are very low.

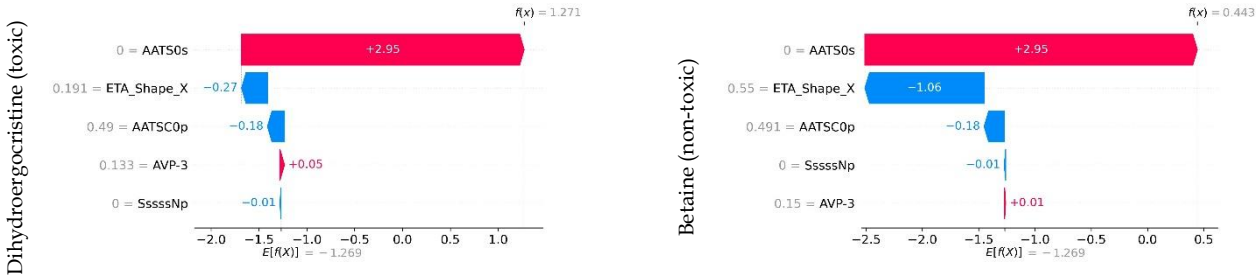

Figure 4. The SHAP waterfall plots of two randomly selected specific molecules of the test set

3. DISCUSSION

3.1. Model outputs and applicability domain

The highest performance was recorded in the LR model. This model shows the best accuracy in the test set (94.68%), with high SE (94.7%) and high SP (94.7%), indicating few false negatives and false positives, respectively. This model predicted the potential risk of pneumonitis with high accuracy based on the physicochemical properties of the drugs. The results obtained from the internal and external validation processes demonstrated the generalization capability and reliability of the LR model. In particular, the ACC rates for the training and test sets showed that the model provided consistent success (95.72% and 94.68%, respectively). This consistency supports the model's ability to avoid overfitting and adapt to new data in the real world.

In our model, AD was evaluated using two techniques to enhance reliability. One of the approaches used was the Tanimoto similarity score, which illuminates the internal diversity of the dataset [21]. The average Tanimoto scores were recorded as 0.3302 for the training set and 0.3300 for the test set. These outcomes show that both sets have molecular diversity and fit well with the AD of the proposed model. We also used Chemical Space Distribution Analysis to assess AD. This analysis visualized the distribution of training and test datasets within the chemical space, providing clearer insights into the chemical variety [23].

The calculation of the Topliss ratio is another critical step in showing the model's validity [21]. Our QSTR models satisfied the model validity criteria with a Topliss ratio of 74.8 (374 compounds/5 descriptors). This finding indicated that the proposed model avoids overfitting [24].

The k-fold cross-validation evaluates the training performance by splitting the data into different subgroups. This method guarantees the strong predictive performance of the model and its generalizability over different chemical domains [25]. All these applied techniques increased the validity of the existing model and identified well-defined AD boundaries. Overall, the superior performance of the LR model demonstrates its potential to obtain reliable and robust predictions in various chemical fields, providing both accuracy and stability. These results support the preference for LR, particularly for large data sets and applications where consistency between different classes is required.

3.2. Mechanistic explanations of the final selected molecular descriptors

ML models are complicated to explain due to their complex structure and are therefore called "black boxes". On the other hand, maximizing model explainability as much as possible increases model reliability. Post hoc methods are applied to increase explainability. One of these methods is to evaluate the contributions of independent features to the predictions [26].

In this study, SHAP analysis was used to provide in-depth insight into the underlying factors driving the predictions of the LR model. Thus, ML models that are described as "black boxes" can be converted into "grey boxes".

3.2.1. Global explanation

Among the selected descriptors, AATS0s was found to have a dominant effect in detecting DIP. Positive values of AATS0s from the Autocorrelation Class of descriptors introduced by Todeschini and Consonni (2009) [27] are associated with larger atomic masses, van der Waals volumes, electronegativities, surface areas, and polarizabilities [28]. AATS0s assesses the uniformity of the electronic environment throughout the molecule and reflects the consistency in the electronic state of atoms that are directly bonded. Higher AATS0s value indicates more consistent electron distribution, generally associated with more stable and less reactive molecular structure. Greater homogeneity in electron distribution may decrease the chance of the molecule damaging biological targets, thereby reducing its toxicity potential [29].

This data in the literature coincides with our results. Contrary, Khan and Roy (2017) associate high AATS0s values with high toxicity despite its low contribution to the model [30]. Although a negative correlation with toxic effects on a global scale was found for the AATS0s descriptor in our study, this is considered a general explanation of the model. The local scale provides more precise results about the contribution of descriptors to the toxic effect of the molecule. As a result, we suggest that the AATS0s descriptor may play a critical role in reducing or preventing DIP, increasing or decreasing its numerical value should be evaluated on a by-molecule basis.

Another significant identifier in this model is ETA_Shape_X. Roy and Ghosh (2004) introduced the ETA Descriptors [31] and are widely used in drug toxicity models [32, 33]. This descriptor class is key in identifying the contributions of branching, functionality, shape, and size to target activity [31]. There are QSTR models created with only ETA indices to evaluate the toxic effects of various chemicals [34, 35]. In a quantitative structure-property relationship (QSPR) study, ETA_Shape_X was found to be the strongest descriptor. This identifier provides structure information based on the number of nuclei related to the shape of the molecule, and a negative correlation was recorded between high ETA_Shape_X and the half-life of the molecule [36]. In the present research, an increase in the ETA_Shape_X value was found to reduce the likelihood of harmful effects; this might be related to a shorter medication half-life. On the other hand, in a QSTR study on the acute toxic effect of 1,2,4-triazole derivatives, ETA_Shape_X contributes positively to the intraperitoneal lethal dose 50 (LD_{50}) values of rodents [37]. As the LD_{50} value of a drug molecule increases, it becomes safer. That is, as the numerical value of the ETA_Shape_X identifier increases, the molecule becomes safer, which supports our global SHAP results.

Our other descriptors, namely AVP-3 and SssssNp have low contributions to model success. Nevertheless, the effect of the AATSC0p descriptor stays unidentified in the context of global explanation.

3.2.2. Local explanation

Global explanations are used to identify the average behavior of an ML model, whereas local explanations consider individual estimations in detail to explain why the model makes a particular prediction for a given descriptor [38]. Individual substructures may not work autonomously and may

influence each other's contribution [39]. Therefore, the SHAP values of the same identifier observed in different molecules may change [40].

In the individual evaluation, the AATS0s descriptor, the most critical in this prediction model, was found to contribute to toxicity in some molecules, while reducing it in others. There are different findings in the literature regarding the correlation of AATS0s with toxicity, supporting this duality [29, 30]. The effect of AATS0s on toxicity may be non-linear; that is, they may contribute to toxic effects within a certain range but show non-toxic effects outside this range. This non-linear relationship may explain the complex effects observed in SHAP analysis. Besides, a single descriptor alone may not be able to determine toxicity in all cases [39]. In this context, assessing a descriptor's contribution to the target effect at the molecular level is required.

AATSC0p and ETA_Shape_X are the most frequently encountered descriptors in the second row. Both descriptors mostly show anti-correlation with the pneumonitis effect. AATSC0p is a descriptor related to the polarizability of the molecule. Although AATSC0p has been reported to negatively correlate with toxicity [29], another study claims that it contributes positively to toxicity [41]. Based on the literature and our findings, we propose that increasing the value of ETA_Shape_X could contribute to a non-toxic effect, although results may vary at the molecular level.

3.3. The advantages and disadvantages of the model

The list of drugs associated with pneumonitis is increasing due to new monoclonal antibodies and biological agents, especially in the treatment of neoplastic and rheumatological diseases [4]. Determining the toxicity in molecules is challenging [2], and hazard reduction/avoidance processes are based on complex, costly, and time-consuming trial-and-error methods [42]. So, drug-related pneumotoxic effects have been determined relatively late in preclinical studies [43]. This uncertainty causes drug development studies to fail and some long-used drugs to still cause respiratory toxicity [2]. Based on this gap in the literature, animal-free prediction models such as the one developed in our study, can facilitate assessing DIP in the early stages of drug development or preclinical trials. Integration of QSTR-supported methods into drug risk management steps is cost- and time-effective and ethically acceptable. Although the reconstruction of the algorithms requires special expertise, all training and test sets are available for reuse. Also, mechanistic explanations of the model offered new perspectives on minimizing or preventing adverse effect potential of drugs. In preliminary toxicity assessment, structural elements of new drug candidates can be re-evaluated with global explanations regarding pneumonitis risk. On the other hand, our local explanations revealed which molecular identifiers should be modified in a particular drug known to cause pneumonitis.

Utilizing a dataset derived directly from case reports and literature, this model eliminates the risk of species-specific selective toxicity. Moreover, the inclusion of only pharmaceuticals in the current study can be interpreted as both a strength and a limitation. By focusing on drugs, a certain level of homogeneity was achieved in the dataset, which increased the success rate of the model. In addition to predicting DIP with high accuracy, another aim of this study was to contribute to the development of safer drugs.

The current model has not evaluated non-drug pneumotoxic substances due to its scope. As in most conventional QSTR models, inorganic chemicals and salts were excluded from the modeling dataset.

3.4. Comparison of our model and other prediction models for pulmonary toxicity

This research is unique since it concentrated on pneumonitis and the dataset contains only APIs. A limited number of QSTR models exist to predict human respiratory toxicity [44]. Most of the models deal with the adverse effects of various chemical substances on the respiratory system in a broader framework. By focusing on a specific toxicity endpoint, this study has contributed to higher accuracy and consistency of the model. Models built on general respiratory toxicity have the advantage of including larger datasets. On the other hand, high chemical diversity may adversely affect the model performance by reducing the homogeneity of the datasets. The support vector machines (SVM) algorithm achieved 86.2% accuracy and a MCC value of 0.722 in the test set in a respiratory toxicity prediction model developed using 2527 compounds [2], whereas 94.68% accuracy and a MCC value of 0.893 MCC were obtained with the LR algorithm in our model (n_T=468). Similarly, in another DI-ILD model using the SVM algorithm (n=2529), an 86.9% success rate was reported in the external set [9]. Besides, Zhang et al. (2018) recorded an ACC of 84.3% with the NB classifier method (n=1241) [45]. The high success rate of the suggested model can be attributed to the creation of a more homogeneous dataset by focusing on a single endpoint.

In addition to pneumonitis, there are a few prediction models that focus on a single endpoint related to respiratory toxicities. These models are mostly concerned with the respiratory sensitization of chemicals

with low MW and specifically examine their occupational exposure risks. Among these studies aiming to improve indoor air quality, Mekenyan et al. (2014) achieved 72% success in their respiratory sensitization model with 202 chemicals, while Jarvis et al. (2015) developed a QSTR model of chemical asthma hazard with 90% SE using 303 compounds [46, 47]. Models for pulmonary irritation with 1997 organic substances [44] and pulmonary inflammation prediction with 54 compounds related to engine exhaust have also been established [48].

The majority of pneumotoxicity models focused only on prediction success and explainable artificial intelligence methods, such as SHAP, were not included. The study of Jaganathan et al. (2022) is one of the few studies that use these methods and has only addressed SHAP global explanations in detail. Local explanations are kept brief [2]. Our research highlighted molecular features, contributing positively or negatively to pneumonitis as a side effect for each test molecule. Most other prediction models present only descriptors without detailing how the effects of these descriptors vary for each compound. In conclusion, this model provides an innovative approach for assessing DIP risk by identifying key molecular features along with their direction.

4. CONCLUSION

The developed binary QSTR model has predicted the risk of DIP and provided insights into the prevention of this adverse effect. Our computational-based model study has supported the design of safer and cost-effective medicines in the early phases of drug development. The given descriptors in the global explanations can guide the process of designing medications for pharmacological groups at high risk for pneumotoxicity. Using local analysis, identifiers for existing pharmaceuticals can be evaluated in detail and optimized concerning pneumonitis risk. Comprehending the molecular structures underlying toxicity is crucial for drug discovery. Our model aims to reduce the dependence on animal-based testing by providing insights into the molecular basis. This study advocates the adoption of reliable *in silico* techniques in the early phases of drug development before clinical trials. This approach supports a more thorough evaluation of the potential side effects of drugs, along with ethical, time, and cost benefits. However, as with other toxicology studies, model outputs need to be validated through further research to reach a definitive conclusion.

5. MATERIALS AND METHODS

5.1. Dataset preparation

Based on pneumonitis risk, the dataset contained 468 APIs divided into two classes, toxic (n=255) and non-toxic (n=213) (Supplementary File 1_Table S1). We gathered all APIs associated with pneumonitis from the Pneumotox database to create the toxic class [6]. The FDA database was used to collect safe drug molecules in terms of pneumonitis [11]. APIs from almost every pharmacological group were randomly collected from the FDA, keeping the number of molecules in two classes close to obtaining a balanced dataset. The pneumonitis risk profiles of the molecules were checked from the Pneumotox and FDA databases as well as the SIDER 4.1 website [8]. To characterize the compound's physico-chemical attributes, two-dimensional structural data files (2D SDF) were obtained from the PubChem database [49]. In addition, 1444 2D descriptors were computed for each molecule using the open-source PaDEL tool [13].

5.2. Data pre-processing

Raw data are often incomplete, erroneous, inconsistent, or unsuitable for modeling. Data pre-processing aims to address such issues and prepare data that are more suitable for computation. The steps of this method are as follows; data cleaning, data reduction, feature selection, data scaling, and data partitioning [50]. In this study, WEKA 3.9.5 software [14] and Python version 3.9.5 were used for basic data pre-processing on raw data [51].

Initially, raw 2D SDF files were examined, and then corrupted files were eliminated. Missing data were filled in using appropriate methods, noisy and duplicate data were removed, and outliers were detected.

Selecting the most significant features without losing information is a critical step [52]. In this study, the descriptive selection process consisted of two stages. Firstly, the CfsSubsetEval filter of the WEKA framework was used in combination with the BestFirst search method to select the best descriptors [4]. Secondly, the Correlation Heatmap was employed to evaluate the correlation matrix of molecular

descriptors with each other. In the presence of highly correlated characteristics, one is usually removed to mitigate multicollinearity and improve model performance [53, 54]. As a result, an optimal set of descriptors was obtained through a two-stage feature selection, thereby enhancing the quality and explainability of the model.

The data are scaled with a proper scale technique to avoid the effect of large values on small ones. In this study, we adopted the widely used Min-Max scaling method [50]. The dataset was randomly divided into training and test sets with an 80:20 ratio, in the current study. As a result of the split, the training set comprised 374 molecules and the test set 94 molecules (Figure 5).

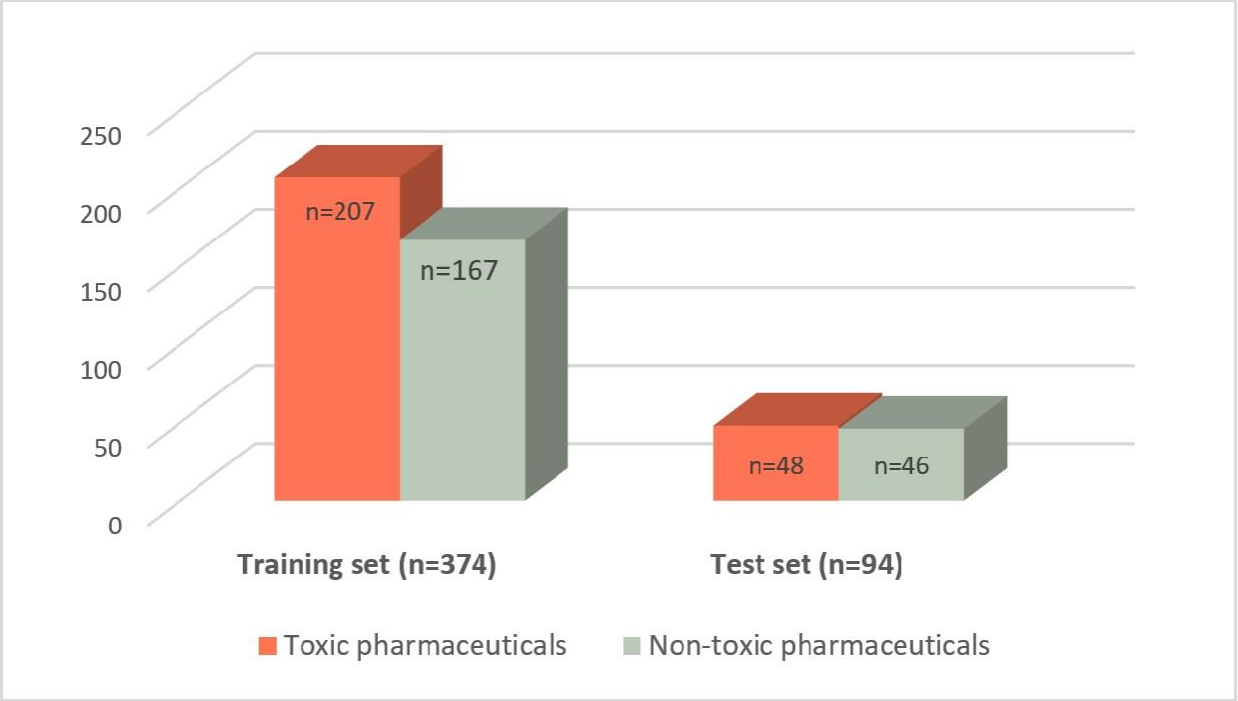

Figure 5. The distribution of the modeling dataset (n: number of compounds)

5.3. Machine learning algorithms

The four ML algorithms including LR [17], NB [18], KStar [19], and IBk [20], were applied to build binary QSTR models based on the extracted descriptors, by using WEKA 3.9.5. software [14].

5.4. Model validation

The combination of internal and external validations is a common technique employed to validate ML models. To evaluate model performance with the training set, k-fold cross-validation is used. Internal validation aids to prevent overfitting and optimize model parameters. External validation uses an independent test dataset and provides a realistic estimate of the model's performance in the real world [21]. In this research, the training set was validated with 10-fold cross-validation, and the model's generalization was determined using an external test set.

Another essential factor to examine when analyzing model validity is the link between the number of training set components and the selected descriptors, which is known as the Topliss ratio [21]. As a generally accepted rule, the Topliss ratio should exceed 5 [24].

5.5. Model evaluation

5.5.1. Performance metrics

The prediction performance of the model was determined using confusion matrix elements including true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Then, we calculated ACC, SP, SE, F-Measure and MCC metrics as shown in the following equations (1). The ROC shows the relationship between SP and SE.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}.$$

$$SP = \frac{TN}{TN + FP}.$$

$$SE = \frac{TP}{TP + FN}. \quad (1)$$

$$F - Measure = \frac{2 * TP}{2 * TP + FP + FN}.$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

5.5.2. Applicability domain

In QSTR modeling, strict rules have been set by the OECD to ensure the applicability of models. In this context, one of the main requirements is the creation of an AD which determines the boundaries of chemical structures within which the model can produce valid and reliable predictions [21]. In this study, the Tanimoto similarity index [22] and chemical space distribution analysis [21] were used to create a well-defined AD. The Tanimoto similarity index is used to measure chemical diversity, ranging from 0 to 1, with values near to 0 indicating wide diversity and values near to 1 indicating high similarity [22]. Chemical space distribution analysis visualizes the dispersion of training and test sets within the chemical space, revealing the extent of chemical diversity [21].

5.6. Model explainability

In the created model, the SHAP strategy was utilized to explain the importance of identifiers in predicting DIP. The SHAP approach provides global and local explanations [55]. Global annotations show the average effect of each feature. Local explainability identifies the particular features influencing the estimation of a single data point [56].

5.6.1. SHapley additive exPlanations

SHAP calculates the effect of identifiers on the prediction, stating positive or negative contributions [57]. The SHAP value was originally utilized in cooperative game theory to ensure a fair distribution of the game's winnings by accounting for the contributions of each participant [12]. The SHAP technique for use in ML models was developed by Lundberg and Lee (2017) to calculate the contributions of identifiers to prediction [55]. In the cooperative game, each participant's SHAP value:

$$\emptyset_i = \sum_{S \subseteq M \backslash i} \frac{|S|!(|M|-|S|-1)!}{|M|!}[f(S \cup i - f(S)] \quad (2)$$

In the formula, M indicates all input identifiers, and S indicates an identifier subset that not contain the identifier i whose marginal contribution is calculated. S⊆M\i points out all subsets S that are subsets of M, except identifier i. f(s) indicates the predicted value made by this coalition. Equation 2 computes the difference between the estimation with the i feature of the model and the estimation without the i identifier. After computing the marginal contributions, equation 3 may be used to get the total of the SHAP values.

$$g(z') = \emptyset_i + \sum_{i=1}^{M} \emptyset_i z_i' \quad (3)$$

Here, $z' \in \{0,1\}^M$ indicates the detected and unidentidied variables.

Acknowledgments: This study received no financial or institutional support.

Author contributions: Concept – F.K.Ç.; Design – F.K.Ç.; Supervision – F.K.Ç.; Resources – F.K.Ç.; Materials – F.K.Ç.; Data Collection and/or Processing – F.K.Ç.; Analysis and/or Interpretation – F.K.Ç.; Literature Search – F.K.Ç.; S.Y.S. Writing – F.K.Ç., S.Y.S.; Critical Reviews – F.K.Ç., S.Y.S.; Other - F.K.Ç., S.Y.S.

Conflict of interest statement: The authors declared no conflict of interest.

REFERENCES

[1] Skeoch S, Weatherley N, Swift AJ, Oldroyd A, Johns C, Hayton C, Giollo A, Wild JM, Waterton JC, Buch M, Linton K, Bruce IN, Leonard C, Bianchi S, Chaudhuri N. Drug-induced interstitial lung disease: a systematic review. J Clin Med. 2018; 7(10): 356. https://doi.org/10.3390/jcm7100356.

[2] Jaganathan K, Tayara H, Chong KT. An explainable supervised machine learning model for predicting respiratory toxicity of chemicals using optimal molecular descriptors. Pharmaceutics. 2022; 14(4): 832. https://doi.org/10.3390/pharmaceutics14040832.

[3] Antoine MH, Mlika M. Interstitial Lung Disease. Treasure Island (FL): StatPearls Publishing. 2024. https://www.ncbi.nlm.nih.gov/books/NBK541084/ (accessed August 30, 2024).

[4] Spagnolo P, Bonniaud P, Rossi G, Sverzellati N, Cottin V. Drug-induced interstitial lung disease. Eur Respir J. 2022; 60(4): 2102776. https://doi.org/10.1183/13993003.02776-2021.

[5] Interstitial Lung Diseases. https://www.nhlbi.nih.gov/health/interstitial-lung-diseases (accessed September 09, 2024).

[6] Pneumotox Website. https://www.pneumotox.com/drug/index/ (accessed August 30, 2024).

[7] Fujimoto D, Kato R, Morimoto T, Shimizu R, Sato Y, Kogo M, Ito J, Teraoka S, Nagata K, Nakagawa A, Otsuka K, Tomii K. Characteristics and prognostic impact of pneumonitis during systemic anti-cancer therapy in patients with advanced non-small cell lung cancer. PLoS One. 2016; 11(12): e016846. https://doi.org/10.1371/journal.pone.0168465.

[8] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016; 44(D1): D1075–D1079. https://doi.org/10.1093/nar/gkv1075. http://sideeffects.embl.de/ (accessed August 30, 2024).

[9] Wang Z, Zhao P, Zhang X, Xu X, Li W, Liu G, Tang Y. In silico prediction of chemical respiratory toxicity via machine learning. Comput Toxicol. 2021; 18: 100155. https://doi.org/10.1016/j.comtox.2021.100155.

[10] Lei T, Chen F, Liu H, Sun H, Kang Y, Li D, Li Y, Hou T. ADMET evaluation in drug discovery. Part 17: Development of quantitative and qualitative prediction models for chemical-induced respiratory toxicity. Mol Pharm. 2017; 14: 2407–2421. https://doi.org/10.1021/acs.molpharmaceut.7b00317.

[11] Food and Drug Administration (FDA). https://www.fda.gov/ (accessed August 30, 2024).

[12] Shapley LS. A value for n-person games. In: Kuhn HW, Tucker AW. (Eds). Contributions to the Theory of Games, 2; Princeton University Press: Princeton, NJ, USA, 1953, pp. 307-317. https://doi.org/10.1515/9781400881970-018.

[13] Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011; 32(7): 1466-1474. https://doi.org/10.1002/jcc.21707.

[14] Frank E, Hall MA, Witten IH. The WEKA Workbench. Online Appendix For "Data Mining: Practical Machine Learning Tools and Techniques", fourth ed., Morgan Kaufmann, Burlington, MA, USA 2016.

[15] Nantasenamat C, Li H, Mandi P, Worachartcheewan A, Monnor T, Isarankura-Na-Ayudhya C, Prachayasittikul V. Exploring the chemical space of aromatase inhibitors. Mol Divers. 2013; 17: 661-677. https://doi.org/10.1007/s11030-013-9462-x.

[16] Pradeep P, Judson R, DeMarini DM, Keshava N, Martin TM, Dean J, Gibbons CF, Simha A, Warren SH, Gwinn MR, Patlewicz G. Evaluation of existing QSAR models and structural alerts and development of new ensemble models for genotoxicity using a newly compiled experimental dataset. Comput Toxicol. 2021; 18: 100167. https://doi.org/10.1016/j.comtox.2021.100167.

[17] Le Cessie S, van Houwelingen JC. Ridge estimators in logistic regression. Appl Stat. 1992; 41(1): 191-201. https://doi.org/10.2307/2347628.

[18] John GH, Langley P. Estimating continuous distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo. 1995, pp. 338-345.

[19] Cleary JG, Trigg LE. K*: An instance-based learner using an entropic distance measure. In: 12th International Conference on Machine Learning. Mach Learn Proceed. 1995; 108-114. https://doi.org/10.1016/b978-1-55860-377-6.50022-0.

[20] Aha D, Kibler D. Instance-based learning algorithms. Mach Learn. 1991; 6: 37-66. https://doi.org/10.1007/BF00153759.

[21] Organisation for Economic Co-Operation and Development (OECD), 2014. Guidance document on the validation of (quantitative) structure-activity relationship [(Q) SAR] models, OECD Series on testing and assessment, No. 69, OECD Publishing, Paris. https://doi.org/10.1787/10.1787/9789264085442-en (accessed August 12, 2024).

[22] Vogt M, Bajorath J. Modeling tanimoto similarity value distributions and predicting search results. Mol Inform. 2017; 36(7): 1600131. https://doi.org/10.1002/minf.201600131.

[23] Medina-Franco JL, Martínez-Mayorga K, Giulianotti MA, Houghten RA, Pinilla C. Visualization of the chemical space in drug discovery. Curr Comput-Aided Drug Des. 2008; 4(4): 322-333. https://doi.org/10.2174/157340908786786010.

[24] Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014; 57: 4977-5010. https://doi.org/10.1021/jm4004285.

[25] Héberger K. Selection of optimal validation methods for quantitative structure–activity relationships and applicability domain. SAR QSAR Environ Res. 2023; 34(5): 415-434. https://doi.org/10.1080/1062936X.2023.2214871.

[26] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020; 58: 82-115. https://doi.org/10.1016/j.inffus.2019.12.012.

[27] Todeschini R, Consonni V. Molecular descriptors for chemoinformatics. In: Mannhold R, Kubinyi H, Folkers G. (Eds). Methods and Principles in Medicinal Chemistry. Wiley VCH, Weinheim, 2009, pp. 27-37. https://doi.org/10.1002/9783527628766.

[28] Moreira-Filho JT, Ranganath D, Conway M, Schmitt C, Kleinstreuer N, Mansouri K. Democratizing cheminformatics: interpretable chemical grouping using an automated KNIME workflow. J Cheminform. 2024; 16: 101. https://doi.org/10.1186/s13321-024-00894-1.

[29] Yang S, Kar S. How safe are wild-caught salmons exposed to various industrial chemicals? First ever in silico models for salmon toxicity data gaps filling. J Hazard Mater. 2024; 477: 135401. https://doi.org/10.1016/j.jhazmat.2024.135401.

[30] Khan K, Roy K. Ecotoxicological modelling of cosmetics for aquatic organisms: A QSTR approach. SAR QSAR Environ Res. 2017; 28(7): 567-594. https://doi.org/10.1080/1062936X.2017.1352621

[31] Roy K, Ghosh G. QSTR with extended topochemical atom indices. 2. fish toxicity of substituted benzenes. J Chem Inf Comput Sci. 2004; 44(2): 559-567. https://doi.org/10.1021/ci0342066.

[32] Roy K, Das RN. The "ETA" Indices in QSAR/QSPR/QSTR Research. In: Pharmaceutical Sciences: Breakthroughs in Research and Practice. IGI Global, Hershey, Pennsylvania, 2017, pp. 978-1011.

[33] Roy K, Ghosh G. Exploring QSARs with Extended Topochemical Atom (ETA) indices for modeling chemical and drug toxicity. Curr Pharm Des. 2010; 16(24): 2625-2639. https://doi.org/10.2174/138161210792389270.

[34] Seth A, Ojha PK, Roy, K. QSAR modeling with ETA indices for cytotoxicity and enzymatic activity of diverse chemicals. J Hazard Mater. 2020; 394: 122498. https://doi.org/10.1016/j.jhazmat.2020.122498.

[35] De P, Roy K. Greener chemicals for the future: QSAR modelling of the PBT index using ETA descriptors. SAR QSAR Environ Res. 2018; 29(4): 319-337. https://doi.org/10.1080/1062936X.2018.1436086.

[36] Khan PM, Lombardo A, Benfenati E, Roy K. First report on chemometric modeling of hydrolysis half-lives of organic chemicals. Environ Sci Pollut Res Int. 2021 Jan;28(2):1627-1642. https://doi.org/10.1007/s11356-020-10500-0.

[37] Liu Z, Dang K, Gao J, Fan P, Li C, Wang H, Huan L, Xiaoni D, Yongchao G, Qian A. Toxicity prediction of 1, 2, 4-triazoles compounds by QSTR and interspecies QSTTR models. Ecotoxicol Environ Saf. 2022; 242: 113839. https://doi.org/10.1016/j.ecoenv.2022.113839.

[38] Houdou A, El Badisy I, Khomsi K, Andrade S. Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review. Aerosol Air Qual Res. 2024; 24(1): 230151. https://doi.org/10.4209/aaqr.230151.

[39] Alves V, Muratov E, Capuzzi S, Politi R, Low Y, Braga R., Zakharov AV, Sedykh A, Mokshyna E, Farag S, Andrade C, Kuz'min C, Fourches D, Tropsha A. Alarms about structural alerts. Green Chem. 2016; 18(16): 4348-4360. https://doi.org/10.1039/C6GC01492E.

[40] Han M, Jin B, Liang J, Huang C, Arp HPH. Developing machine learning approaches to identify candidate persistent, mobile and toxic (PMT) and very persistent and very mobile (vPvM) substances based on molecular structure. Water Res. 2023; 244: 120470. https://doi.org/10.1016/j.watres.2023.120470.

[41] Shavalieva G, Papadokonstantakis S, Peters G. Prior knowledge for predictive modeling: The case of acute aquatic toxicity. J Chem Inf Model. 2022;62(17):4018-4031. https://doi.org/10.1021/acs.jcim.1c01079.

[42] Pal R, Patra SG, Chattaraj PK. Quantitative structure–toxicity relationship in bioactive molecules from a conceptual DFT perspective. Pharmaceuticals. 2022; 15(11): 1383. https://doi.org/10.3390/ph15111383.

[43] Bassan A, Alves VM, Amberg A, Anger LT, Beilke L, Bender A, Bernal A, Cronin MTD, Hsieh JH, Johnson C, Kemper R, Mumtaz M, Neilson L, Pavan M, Pointon A, Pletz J, Ruiz P, Russo DP, Sabnis Y, Sandhu R, Schaefer M, Stavitskaya L, Szabo DT, Valentin JP, Woolley D, Zwickl C, Myatt GJ. In silico approaches in organ toxicity hazard assessment: Current status and future needs for predicting heart, kidney and lung toxicities. Comput Toxicol. 2021; 20: 100188. https://doi.org/10.1016/j.comtox.2021.100188.

[44] Wehr MM, Sarang SS, Rooseboom M, Boogaard PJ, Karwath A, Escher SE. RespiraTox–development of a QSAR model to predict human respiratory irritants. Regul Toxicol Pharmacol. 2022; 128: 105089. https://doi.org/10.1016/j.yrtph.2021.105089.

[45] Zhang H, Ma JX, Liu CT, Ren JX, Ding L. Development and evaluation of in silico prediction model for drug-induced respiratory toxicity by using Naïve Bayes classifier method. Food Chem Toxicol. 2018; 121: 593–603. https://doi.org/10.1016/j.fct.2018.09.051.

[46] Mekenyan O, Patlewicz G, Kuseva C, Popova I, Mehmed A, Kotov S, Zhechev T, Pavlov T, Temelkov S, Roberts DW. A mechanistic approach to modeling respiratory sensitization. Chem Res Toxicol. 2014; 27(2): 219-239. https://doi.org/10.1021/tx400345b.

[47] Jarvis J, Seed MJ, Stocks SJ, Agius RM. A refined QSAR model for prediction of chemical asthma hazard. Occup Med (Lond). 2015; 65(8): 659–666. https://doi.org/10.1093/occmed/kqv105.

[48] Hosoya J, Tamura K, Muraki N, Okumura H, Ito T, Maeno M. A novel approach for a toxicity prediction model of environmental pollutants by using a quantitative structure- activity relationship method based on toxicogenomics. ISRN Toxicol. 2011; 2011(1): 515724. https://doi.org/10.5402/2011/515724.

[49] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoeaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem 2023 update. Nucleic Acids Res. 2023; 51(D1): D1373-D1380. https://doi.org/10.1093/nar/gkac956.

[50] Fan C, Chen M, Wang X, Wang J, Huang B. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. Front Energy Res. 2021; 9: 652801. https://doi.org/10.3389/fenrg.2021.652801.

[51] Python 3.9.5. Software. https://www.python.org/ (accessed August 30, 2024).

[52] Iranzad R, Liu X. A review of random forest-based feature selection methods for data science education and applications. Int J Data Sci Anal. 2024; 1-15. https://doi.org/10.1007/s41060-024-00509-w.

[53] Rodgers JL, Nicewander WA. Thirteen ways to look at the correlation coefficient. The American Statistician. 1988; 42(1): 59-66. https://doi.org/10.1080/00031305.1988.10475524.

[54] Freedman D, Pisani R, Purves R. Statistics (International Student Edition), fourth ed., WW Norton & Company: New York, USA 2007.

[55] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017; 30. https://doi.org/10.48550/arXiv.1705.07874.

[56] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE Trans. Neural Netw Learn Syst. 2020; 32(11): 4793-4813. https://doi.org/10.1109/TNNLS.2020.3027314.

[57] Sun J, Sun CK, Tang YX, Liu TC, Lu CJ. Application of SHAP for explainable machine learning on age-based subgrouping mammography questionnaire data for positive mammography prediction and risk factor identification. Healthcare (Basel). 2023; 11(14): 2000. https://doi.org/10.3390/healthcare11142000.