DOI: 10.32552/2025.ActaMedica.1142

acta medica

ORIGINAL ARTICLE

In silico prediction of rhabdomyolysis-inducing drugs utilizing a supervised machine learning model

Feyza Kelleci Çelik[1]
ORCID: 0000-0003-4874-6648

[1]Karamanoğlu Mehmetbey University, Vocational School of Health Services, Karaman, Türkiye

Corresponding Author: Feyza Kelleci Çelik
E-mail: feyza-kelleci@hotmail.com

*Received: 26 April 2025, Accepted: 7 July 2025,
Published online: 30 September 2025*

ABSTRACT

Objective: Rhabdomyolysis is a life-threatening syndrome characterized by the release of myocyte components into the bloodstream and can be induced by pharmaceutical agents. Although quantitative structure-activity relationship (QSAR) models are widely used for assessing adverse drug reactions, studies in computational toxicology focusing on rare and serious side effects, such as rhabdomyolysis, are still relatively limited. Due to this gap, this study aims to build an *in silico* QSAR model for early prediction of drugs at risk of rhabdomyolysis.

Materials and Methods: A binary dataset was developed by gathering 187 pharmaceutical compounds from the Drug-Induced Rhabdomyolysis Atlas (DIRA), and classification models were developed in the research. Machine learning (ML) algorithms, such as Instance-Based Learning with k-Nearest Neighbors (IBk), Simple Logistic (SL), and BayesNet (BN), were employed. Additionally, post-hoc model explanations and importance rankings of molecular descriptors were provided using Permutation Feature Importance (PFI).

Result: The performances of the ML classifiers ranged from 82.00% to 85.33% in the training set and from 75.67% to 81.08% in the test set. The highest success rate for the test set was achieved by the IBk model, with a rate of 81.08%. The most significant feature in the post-hoc IBk model explanation using PFI was highlighted as the JGI6 descriptor. The descriptor class with the most identifiers was the Electrotopological State Atom Type (E-State) Descriptors.

Conclusion: The physicochemical properties presented in this study regarding rhabdomyolysis and the developed models are anticipated to serve as effective tools for assessing the risk of rhabdomyolysis in drug molecules.

Keywords: machine learning, pharmaceutical toxicology, QSAR, rhabdomyolysis

INTRODUCTION

Rhabdomyolysis is a clinical syndrome characterized by the release of intracellular elements, including myoglobin, electrolytes, aldolase, and creatine kinase (CK), into the bloodstream as a result of acute or subacute damage to striated muscle [1,2]. Without prompt and aggressive interventions [3], fatal complications may occur, including acute renal failure [4], cardiac arrhythmia, intravascular coagulation [2], and acute muscle necrosis. Additionally, clinical signs such as limb weakness, myalgias, fever, leukocytosis, dark urine, and myoglobinuria may also develop [5]. The factors causing rhabdomyolysis are categorized into two main groups: hereditary and acquired. The acquired causes are further divided into traumatic and non-traumatic types. The most prevalent non-traumatic factors include the use of pharmaceuticals. Approximately 150 pharmaceutical compounds have been identified as causes of rhabdomyolysis [1], including various pharmacological groups such as psychoactive drugs [6], selective serotonin reuptake inhibitors [7], statins [8], antihistamines, and antidepressants [9]. In this regard, the Drug-Induced Rhabdomyolysis Atlas (DIRA), a web-based application, was developed to ensure safe drug use without causing rhabdomyolysis. The DIRA presents a classification framework based on drug labeling information provided by the Food and Drug Administration (FDA). This framework classifies drugs into four classes based on their risk of causing rhabdomyolysis (DIR) [8].

Currently, quantitative structure-activity relationship (QSAR), a computational method used in drug development, serves as an essential tool for evaluating potential drug side effects. These computational models can predict possible toxicity profiles by analyzing the structural features of compounds. Based on machine learning (ML) algorithms, QSAR models analyze the information obtained from the input dataset and provide faster, ethical, and cost-effective results compared to traditional laboratory tests [10,11].

Despite the numerous [ITALIC]in silico[/ITALIC] studies in the literature concerning the toxic effects of pharmaceuticals, predictive models assessing the rhabdomyolysis risk remain relatively limited [12-14]. This study focused on predicting rare but life-threatening DIR risk using QSAR models. The dataset was collected from the DIRA, which contains 187 active pharmaceutical ingredients presented by Wen et al. to support developing new methodologies for addressing the rhabdomyolysis side effect [8]. The present study focused on the potential of pharmaceuticals to induce rhabdomyolysis rather than assessing varying degrees of risk. For this purpose, drugs with varying hazard levels in DIRA were grouped together, whereas safe drugs with no rhabdomyolysis risk were classified separately. Thus, the drug status was evaluated as binary: "induces rhabdomyolysis" or "does not induce rhabdomyolysis". In the current study, classification-based QSAR models were created as binary, and the Permutation Feature Importance (PFI) method was employed to improve the model's explainability and prioritize the descriptors. These models enable the early identification of rhabdomyolysis risk for a molecule with an unclear side effect profile during the initial phases of drug development. Furthermore, the prioritized descriptors guide which physicochemical properties need to be modified. Thus, optimizing molecular descriptors may prevent or reduce the risk of rhabdomyolysis. These QSAR models can enhance the management of side effect profiles for pharmaceuticals, resulting in safer and more sustainable drug development.

MATERIALS AND METHODS

Data collection, curation, and preparation

The dataset comprised 187 orally or parenterally administered pharmaceuticals for human use, split into two groups: DIR-positive (n=147) and DIR-negative (n=40) (Table S1). All individual molecules were collected from the FDA-based DIRA website [8]. The molecular characteristics were gathered from two-dimensional structure data files (2D-SDFs) [15]. Then, the open-source PaDEL tool was used to generate the descriptors. The software currently calculates 1444 2D physicochemical properties of the molecules [16].

Data curation and preparation are crucial processes for converting raw data into a suitable format for modeling. The steps include cleaning, instance reduction, attribute selection, data transformation, and data partitioning [17]. This research utilized

Python 3.9.5 [18] and WEKA 3.9.5 [19] to prepare the raw data. WEKA 3.9.5 is an open source software package used for data mining and machine learning applications [19]. Raw 2D-SDFs were initially analyzed, and then corrupted data was removed. Noisy and duplicate data were removed.

The attribute selection process finds the optimal descriptor set with the highest correlation with the specific target variable [20]. This study utilized the CfsSubsetEval-Best First method in WEKA 3.9.5 [19] to select a relevant subset of features for model construction.

To mitigate the impact of large values on smaller ones, the data are scaled using a suitable scaling method during the data transformation step. We employed the popular Min-Max scaling approach [16]. After scaling the data, the analysis set was randomly divided into training (80%, n=150) and test sets (20%, n=37) (Table 1).

Table 1. Composition of the training and test sets

	Dataset (n=187)	
	Training set (80%, n=150)	**Test set (20%, n=37)**
DIR-positive	120	27
DIR-negative	30	10

n: number of molecules; DIR: drug-induced rhabdomyolysis

Development and validation of the models

The three ML algorithms—Instance-Based Learning with k-Nearest Neighbors (IBk) [21], Simple Logistic (SL) [22], and BayesNet (BN) [23]—were employed to construct binary-QSAR models based on the selected optimal identifiers. IBk is grounded in instance-based classification, relying on k-nearest neighbour methods for prediction. Instead of constructing a general model, it retains all training instances and classifies new inputs based on the majority class among the k nearest examples, determined by a distance metric [21]. SL is a classification algorithm that builds logistic regression models using the LogitBoost technique. It incrementally adds base learners to minimize logistic loss, resulting in a probabilistic model suitable for both binary and multi-class problems [22]. BN employs a directed acyclic graph to represent a probabilistic model, illustrating the relationships between random variables and their conditional dependencies. Its fundamental mathematical foundation is based on Bayes' theorem [23].

The k-fold cross-validation technique is used for validating the training set. This technique divides the data set into k equal parts, using each part as a validation set while the remaining parts are used to train the model. Repeating the process k times measures the model's generalization ability and reduces the risk of overfitting [24]. The 10-fold cross-validation method was employed in the study. Additionally, an independent test set was used for external validation to assess the model's performance.

This research employed the Topliss ratio, as recommended by the OECD, for validation in drug modeling studies. The ratio is key in assessing the model's reliability [24]. For a model to be considered validated under this criterion, the Topliss ratio must exceed 5 [25].

The model's performance was evaluated using the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) components of the confusion matrix. Performance metrics, including accuracy (ACC), specificity (SP), sensitivity (SE), F-score, and Matthews correlation coefficient (MCC), were computed using the confusion matrix elements. The metrics were calculated using equations (1) through (5), as shown below.

Accuracy (ACC)

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

Specificity (SP)

$$SP = \frac{TN}{TN + FP}. \tag{2}$$

Sensitivity (SE)

$$SE = \frac{TP}{TP + FN}. \tag{3}$$

F-score

$$F - score = \frac{2\,x\,TP}{2\,x\,TP + FP + FN}. \tag{4}$$

Matthews correlation coefficient (MCC)

$$MCC = \frac{TP\,x\,TN - FP\,x\,FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{5}$$

The Organisation for Economic Co-operation and Development (OECD) has established a framework of rules to ensure the applicability of models to different chemical structures. In this context, defining an applicability domain (AD) is essential. This domain sets the boundaries for the structures to which the model can provide accurate and reliable predictions. The Tanimoto index [26] and chemical space analysis [24] were employed to establish a well-defined AD in this research. The Tanimoto index measures chemical diversity, with values ranging from 0 to 1. An index closes to 0 indicates a high degree of diversity, while one close to 1 indicates a high degree of similarity [26]. The chemical space analysis shows the distribution of the test and training sets in chemical space [24].

Post-Hoc explainability of the models

The OECD guidance on QSAR models encourages improving model explainability to increase reliability [24]. The PFI analysis from the model-agnostic methods is widely employed to explain decisions made by black-box models. The PFI method quantifies the effect of each feature on the predictive accuracy. To understand the effect on the model's performance, the values of a specific feature are randomly mixed, and the change in the accuracy is measured. This analysis enhances the clarity of the complex estimation processes [27].

RESULTS

Selected molecular descriptors

1444 2D descriptors were calculated by the open-source PaDEL tool [16]. After this process, the best 24 2D descriptors were selected using the CfsSubsetEval filter+BestFirst search method. Consequently, we created prediction models utilizing the 24 optimal descriptors (Table 2) to enhance modeling success.

Table 2. The selected molecular descriptors

No	Descriptor	Description	Descriptor Class
1.	JGI6	Mean topological charge index of order 6	Topological Charge Descriptor
2.	GGI7	Topological charge index of order 7	
3.	maxsOH	Maximum atom-type E-State: -OH	Electrotopological State Atom Type (E-State) Descriptor
4.	hmin	Minimum H E-State	
5.	minHBint10	Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 10	
6.	minHBint5	Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 5	
7.	DELS	Sum of all atoms intrinsic state differences	
8.	maxHCsats	Maximum atom-type H E-State: H bonded to B, Si, P, Ge, As, Se, Sn or Pb	
9.	naaS	Count of atom-type E-State: aSa	
10.	MATS4s	Moran autocorrelation - lag 4 / weighted by I-state	Autocorrelation Descriptor
11.	MATS2c	Moran autocorrelation - lag 2 / weighted by charges	
12.	ATSC8p	Centered Broto-Moreau autocorrelation - lag 8 / weighted by polarizabilities	
13.	MIC3	Modified information content index (neighborhood symmetry of 3-order)	Information Content Descriptor
14.	MIC2	Modified information content index (neighborhood symmetry of 2-order)	
15.	VE3_D	Logarithmic coefficient sum of the last eigenvector from detour matrix	Detour Matrix Descriptor
16.	VE3_Dt	Logarithmic coefficient sum of the last eigenvector from detour matrix	
17.	VE2_Dzp	Average coefficient sum of the last eigenvector from Barysz matrix / weighted by polarizabilities	Barysz Matrix Descriptor
18.	VE3_DzZ	Logarithmic coefficient sum of the last eigenvector from Barysz matrix / weighted by atomic number	
19.	VC-3	Valence cluster, order 3	Chi Cluster Descriptor
20.	SCH-5	Simple chain, order 5	Chi Chain Descriptor
21.	nF9Ring	Number of 9-membered fused rings	Ring Count Descriptor
22.	nAtomLAC	Number of atoms in the longest aliphatic chain	Longest Aliphatic Chain Descriptor
23.	WTPT-5	Sum of path lengths starting from nitrogens	Weighted Path Descriptor
24.	SpMin7_Bhm	Smallest absolute eigenvalue of Burden modified matrix - n 7 / weighted by relative mass	Burden Modified Eigenvalues Descriptor

Table 3. Performance measurements of the classifiers

	IBk		SL		BN	
	Training	**Test**	**Training**	**Test**	**Training**	**Test**
ACC %	84.00	81.08	85.33	78.38	82.00	75.67
SP	0.833	0.806	0.844	0.773	0.855	0.732
SE	0.840	0.811	0.853	0.784	0.820	0.757
F-score	0.936	0.808	0.847	0.776	0.831	0.727
MCC	0.476	0.506	0.509	0.420	0.531	0.293

IBk: Instance-Based Learning with k-Nearest Neighbors; SL: Simple logistic; BN: BayesNet; ACC: Accuracy; SP: Specificity; SE: Sensitivity; MCC: Matthews correlation coefficient.

Evaluation of the model performance

In the current study, the IBk [21], SL [22], and BN [23] algorithms were used to construct QSAR models due to their highest performance. The ACC, SP, SE, F-score, and MCC metrics were computed for each algorithm. The internal and external validation outcomes were analyzed to assess the performance (Table 3).

The classifiers' performances were recorded with values ranging from 82.00% to 85.33% in the training set and from 75.67% to 81.08% in the test set. The model has been trained using the training dataset, and the observed performance reflects its capability and consistent learning from the training data. In this study, the 10-fold cross-validation technique was preferred for internal validation. The k-fold cross-validation assesses training performance by dividing the data into several subgroups. The approach guarantees strong predictive performance and versatility across different chemical domains for the model [28]. The test set is a dataset that is excluded from the training process, serving to evaluate the model's ability to generalize. The model's final performance evaluation is conducted on the test set. In the current model, the IBk algorithm demonstrated the highest success rate on the test set (81.08%), along with all other evaluated metrics. The values of 0.806 and 0.811 for SP and SE, respectively, indicate the model's high capability for true negative and true positive rates. The results from the internal and external validation processes demonstrated the IBk model's high generalization capability and reliability. Specifically, the ACC rates for the training and test sets indicated that the IBk model achieved consistent success, with rates of 84.00% and 81.08%, respectively. This consistency enhances the model's ability to prevent overfitting and adapt to new real-world data. The findings indicate that the IBk model effectively distinguishes between DIR-positive and DIR-negative compounds, demonstrating its robust predictive capabilities.

Calculating the Topliss ratio is crucial in validating the QSAR model [24]. The proposed models met the validity criteria, achieving a Topliss ratio of 7.8, based on 187 compounds and 24 descriptors. Based on this result, overfitting appears to be prevented.

We performed analyses of the Tanimoto similarity index [26] and chemical space distribution [24] to establish a reliable AD area for ensuring model robustness. The average Tanimoto scores for the training and test datasets were recorded as 0.3739 and 0.4070, respectively. These scores indicate chemical diversity and AD compatibility within the datasets. Molecular weight (MW) and Ghose-Crippen LogKow (ALogP) values were utilized to analyze the distribution of chemical space (Figure 1). The MW values of the molecules ranged from 131.0946 to 2637.0983 g/mol, while their ALogP values were from -26.7021 to 3.8871. This visualization confirmed that the test set components were adequately included in the chemical domain of the training set.

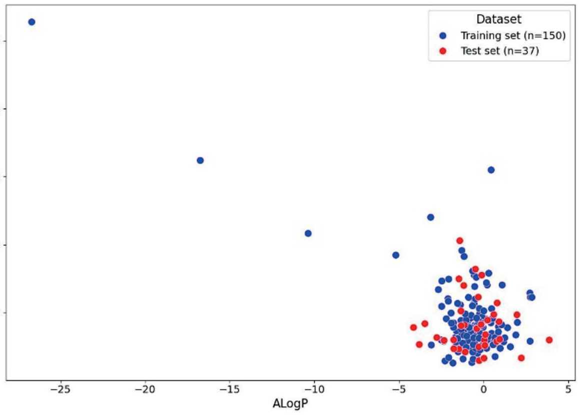

Figure 1. Distribution analysis in chemical space

(n: number of molecules; MW: Molecular Weight; AlogP: Ghose-Crippen LogKow)

These techniques enhanced the current model's validity and clarified AD's boundaries. The IBk model shows promise in producing robust and reliable predictions across various chemical domains. These findings support the selection of the IBk algorithm to ensure inter-class consistency.

Explanation of the top-performing model

The top-performing IBk model was explained using PFI analysis (Figure 2). The graph's vertical axis (Y-axis) lists the features utilized by the model, arranged from bottom to top according to increasing relative importance—features positioned higher have a greater effect on the model's predictions. The horizontal axis (X-axis) indicates the importance score attributed to each feature within the model. This score is computed by assessing the decrease in model accuracy when the values of a particular feature are randomly permuted. Variables with higher values along the X-axis strongly influence the model's decisions. Variables near zero minimally affect model predictions. This analysis highlights the features exerting the greatest influence on predictions [27].

The most significant descriptor for the IBk model is JGI6, followed by maxsOH, MATS4s, and nAtomLAC. Next, MATS2c, MIC3, and MIC2 show equal influence. After these, VC-3 and SpMin7_Bhm yield similar effects. Following this are VE3_D and maxHCsats, which also exhibit comparable impacts. Then, WTPT-5 and SCH-5 rank similarly in their effects. Next is hmin, followed by minHBint10. Additionally, minHBint5, DELS, and VE2_Dzp share equal importance, trailing behind nF9Ring. Following these are naaS, VE3_Dt, and VE3_DzZ, which indicate equal impact. Lastly, GGI7 and ATSC8p are noted as having a negligible effect.

DISCUSSION

The majority of the descriptors in the model belong to the Electrotopological State Atom Type (E-State) Feature class (maxsOH, maxHCsats, hmin, minHBint10, minHBint5, DELS, and naaS). The E-state index encodes both electronic and topological information at the atomic and sub-molecular levels [29]. This class plays a crucial role in identifying functional regions of molecules with potential pharmacophore or toxicophore properties. The capability to evaluate electronic structures and topological properties via a comprehensive approach has established E-state indices as a crucial instrument for QSAR analyses [30]. Our model's descriptors mainly consist of E-state indices consistent with the chemoinformatics literature. This model contains three descriptors from the Autocorrelation Descriptor class [31]: MATS4s, MATS2c, ATSC8p. Some drugs known to cause rhabdomyolysis have been reported to trigger this condition through direct toxicity to skeletal muscle, by increasing intracellular free ionized calcium levels, and by decreasing serum coenzyme Q levels

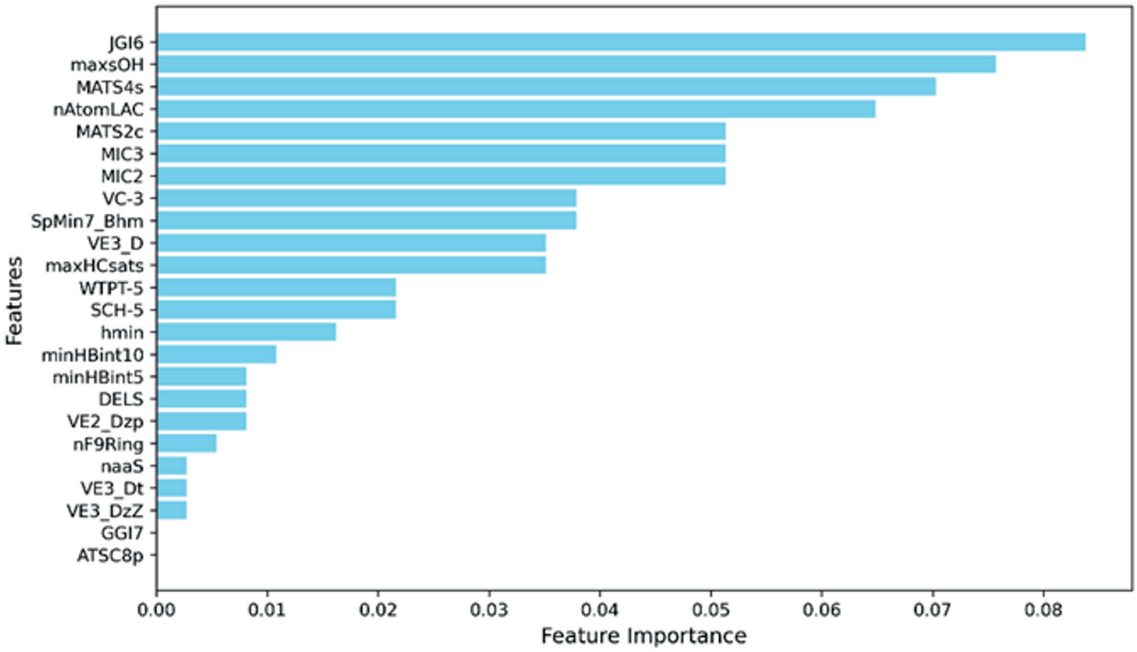

Figure 2. The contribution of molecular descriptors to the top-performing model

[9]. In a QSAR model for predicting the activity of 3-Hydroxy-3-methylglutaryl-CoA (HMG-CoA) reductase inhibitors, various descriptors from the E-State, Autocorrelation, Topological Charge, Detour Matrix, and Barysz Matrix Descriptor classes were utilized, similar to our study. These characteristics significantly impacted the model's estimating capacity [32]. Additionally, a study aimed at identifying newly synthesized HMG-CoA reductase inhibitors through [ITALIC]in silico[/ITALIC] methods employed chemical similarity analysis and QSAR model integration, with Autocorrelation descriptors serving as a significant component [33]. Our QSAR model also included descriptors from Topological Charge (JGI6 and GGI7), Information Content (MIC3 and MIC2), Detour Matrix (VE3_D and VE3_Dt), and Barysz Matrix (VE2_Dzp and VE3_DzZ). Given that rhabdomyolysis is a serious side effect of statins, using these descriptor classes in our study is significant. In the model of statin activity created by Ancuceanu et al. [32], the JGI5 descriptor from the Topological Charge index was used. Higher HMG-CoA reductase inhibitory activity was linked to higher JGI5 levels. The sixth-order derivative, JGI6, was the strongest descriptor in the current model. JGI6 tends to rise as the molecular structure becomes more complicated, with more ring systems, branching, and more heteroatoms [32]. Based on the chemoinformatic literature and current findings, it can be suggested that a higher JGI6 value may increase the risk of drug-induced rhabdomyolysis.

The current model includes one descriptor for each class in the Chi Cluster Descriptor (VC-3), Chi Chain Descriptor (SCH-5), Longest Aliphatic Chain Descriptor (nAtomLAC), Burden Modified Eigenvalues Descriptor (SpMin7_Bhm), Weighted Path Descriptor (WTPT-5), and Ring Count Descriptor (nF9Ring). Rajathei et al. developed a 2D-QSAR model by correlating the structural features of a series of atorvastatin analogs identified as HMG-CoA reductase inhibitors with their biological activities [33]. One of the significant descriptors in the HMG-CoA reductase inhibitors model was SCH-7, a Chi Chain Descriptor family member. The SCH-x descriptors of the Kier and Hall molecular connectivity indices [34] are the x-th degree chain (or ring) type versions, describing the x-th degree of connectivity of non-hydrogen atoms in molecules. According to a study by Rajathei et al., the inhibitory action of SCH-7 may be attributed to its higher degree of connectivity, resulting from its increasing positive value [33]. Similarly, the 5th-level derivative of this descriptor, SCH-5, emerged as a significant predictor in the current DIR model. The current model suggests that the susceptibility of medications to rhabdomyolysis may increase as the SCH-5 value increases. Another significant descriptor highlighted in the aforementioned study is the VE3_Dt descriptor derived from the Detour Matrix index [33], which is also included in our model. Furthermore, the present model employed VE3_D, while the HMG-CoA reductase inhibitors model used VE1_D and VE2_D [33].

The PIF analysis revealed that GGI7 and ATSC8p have the lowest impact on the current model prediction. After removing these features and rebuilding the model, a slight decrease in performance was observed. This implies that these features may interact with other variables as they are not entirely independent. In the presence of interrelated variables, assessing feature combinations yields more reliable insights than analyzing individual predictors [35]. The descriptors presented in this study are suggested as potential toxicophore structures responsible for a molecule's rhabdomyolysis risk. Before modifying drugs at the molecular level to reduce the risk of rhabdomyolysis, an assessment should consider the descriptors supported by other studies and the importance ranking determined by the PIF analysis (Figure 2) to ensure the decision-making process is well-founded.

Prior molecular modifications aimed at minimizing DIR, a comprehensive assessment should be conducted based on the importance hierarchy determined through PIF analysis (Figure 2) and descriptors validated by previous studies to ensure a robust and evidence-based decision-making framework.

Strengths and limitations of the model

DIR is a potentially fatal idiosyncratic adverse drug reaction. Diagnosing DIR in clinics can often be challenging due to the limited availability of methods. Current methods, particularly monitoring CK levels and testing for myoglobinuria in urine, are inadequate for the early and definitive detection of DIR [36]. Various commonly prescribed medications, such as antidepressants (e.g., promazine, trifluoperazine), lipid-lowering medications (e.g., clofibrate, lovastatin), and antihistamines (e.g., doxylamine, diphenhydramine), have the potential to trigger rhabdomyolysis. As more newly synthesized drugs are introduced into clinical use, the risk of rhabdomyolysis associated with these drugs is expected to increase [9]. Thus, accurately identifying the DIR risk before marketing is crucial. Utilizing [ITALIC]in silico[/ITALIC] analysis is recommended as the quickest and most economical approach in the initial phase [23].

The limited number of DIR cases and the inclusion of drugs without conclusive safety evidence regarding rhabdomyolysis risk in the 'safe' group limit the accuracy and reliability of several DIR models available in the literature. To address this gap in the literature, a 2D-QSAR model was developed in this study to predict the pharmaceutical-related risk of rhabdomyolysis using DIRA data. In this context, models were developed utilizing the IBk, SL, and BN algorithms with data categorized as DIR-positive and DIR-negative. The IBk model, noted for its strong predictive performance, was analyzed in detail. The binary IBk model achieved an accuracy of 81.08% on the test set. In the multiclass DIR QSAR model developed by Zhou et al., a success rate of 73.00% was recorded using the Random Forest (RF) algorithm. Although assessing various levels of rhabdomyolysis risk in the RF model is considered an advantage, it requires improvements to increase its success rate [13].

In the current study, collecting data from a single source increased homogeneity and positively contributed to data consistency and model accuracy. However, depending on a single source dataset can present challenges in creating comprehensive datasets. Conversely, the accuracy of both the positive and negative groups in the present dataset is supported by the literature. In this respect, compared to other models that accept medications without rhabdomyolysis data as negative, it offers a more reliable dataset. For example, Cui et al. developed a binary DIR prediction model using 163 drug molecules associated with rhabdomyolysis risk and 1341 drug molecules with no reported rhabdomyolysis risk. In the study, the RF algorithm achieved the highest success rate of 79.28% [12]. Besides, the binary rhabdomyolysis QSAR model developed using the Support Vector Machine (SVM) algorithm, based on a dataset of similar size to the current study, achieves an accuracy rate of 84.50%. Although the study demonstrated a higher performance than the 81.08% ACC rate obtained in our study, a direct comparison is not appropriate due to dataset differences. The SVM model's dataset includes various ingredients, pharmaceuticals, and chemicals [9]. The diversity in the dataset is a crucial factor directly affecting the model's generalization capacity. Concentrating exclusively on pharmaceutical compounds has both advantages and limitations. Utilizing only pharmaceutical-grade ingredients ensured consistency in the dataset, enhancing the model's sensitivity to a specific chemical group and contributing to safer drug development processes. Additionally, the presented model explanations provide new insights into minimizing or preventing potential drug side effects. In this context, the present model excludes non-pharmaceutical substances. Furthermore, as in traditional QSAR modeling, salts and inorganic compounds have been excluded from the scope of analysis. Restructuring the algorithms used in the model requires high expertise; however, the availability of the datasets simplifies the process of creating the model.

In conclusion, the QSAR models developed in the present study can support DIR assessment during drug development and the early stages of preclinical research. Integrating QSAR-based approaches into drug safety management offers an ethically sustainable alternative, enhancing both economic efficiency and time effectiveness. Alongside the strong predictive performance demonstrated by the DIR model, this study aims to contribute to safer drug development by providing structural insights. The study anticipates that the presented physicochemical properties and the developed DIR models will serve as significant guides and effective analytical tools for assessing DIR-related risks.

Author contribution

Study conception and design: FKÇ; data collection: FKÇ; analysis and interpretation of results: FKÇ; draft manuscript preparation: FKÇ. The author reviewed the results and approved the final version of the manuscript.

Ethical approval

For this study, which was conducted entirely using *in silico* methods, ethics committee approval was not required as no human participants, animals, or identifiable personal data were involved.

Funding

The author declare that the study received no funding.

Conflict of interest

The author declare that there is no conflict of interest.

REFERENCES

[1] Waldman W, Kabata PM, Dines AM, et al. Rhabdomyolysis related to acute recreational drug toxicity-A Euro-DEN study. PLoS One 2021;16(3):e0246297. https://doi.org/10.1371/journal.pone.0246297

[2] Baeza-Trinidad R. Rhabdomyolysis: A syndrome to be considered. Med Clin (Barc) 2022;158(6):277-283. https://doi.org/10.1016/j.medcli.2021.09.025

[3] Hohenegger M. Drug induced rhabdomyolysis. Curr Opin Pharmacol 2012;12(3):335-339. https://doi.org/10.1016/j.coph.2012.04.002

[4] Morin AG, Somme D, Corvol A. Rhabdomyolysis in older adults: Outcomes and prognostic factors. BMC Geriatr 2024;24(1):46. https://doi.org/10.1186/s12877-023-04620-8

[5] Nance JR, Mammen AL. Diagnostic evaluation of rhabdomyolysis. Muscle Nerve 2015;51(6):793-810. https://doi.org/10.1002/mus.24606

[6] Amanollahi A, Babeveynezhad T, Sedighi M, et al. Incidence of rhabdomyolysis occurrence in psychoactive substances intoxication: A systematic review and meta-analysis. Sci Rep 2023;13(1):17693. https://doi.org/10.1038/s41598-023-45031-4

[7] Wang Y, Lin Y, Lin Q, Liang H, Cai W, Jiang D. Exploring the association between selective serotonin reuptake inhibitors and rhabdomyolysis risk based on the FDA pharmacovigilance database. Sci Rep 2023;13(1):12257. https://doi.org/10.1038/s41598-023-39482-y

[8] Wen Z, Liang Y, Hao Y, et al. Drug-Induced Rhabdomyolysis Atlas (DIRA) for idiosyncratic adverse drug reaction management. Drug Discov Today 2019;24(1):9-15. https://doi.org/10.1016/j.drudis.2018.06.006

[9] Hu X, Yan A. In silico prediction of rhabdomyolysis of compounds by self-organizing map and support vector machine. Toxicol in Vitro 2011;25(8):2017-2024. https://doi.org/10.1016/j.tiv.2011.08.002

[10] Chang Y, Hawkins BA, Du JJ, Groundwater PW, Hibbs DE, Lai F. A guide to in silico drug design. Pharmaceutics 2022;15(1):49. https://doi.org/10.3390/pharmaceutics15010049

[11] Kelleci Çelik F, Yılmaz Sarıaltın S. An explainable prediction model for drug-induced interstitial pneumonitis. J Res Pharm 2025;29(1):322-334. https://doi.org/10.12991/jrespharm.1644357

[12] Cui X, Liu J, Zhang J, Wu Q, Li X. In silico prediction of drug-induced rhabdomyolysis with machine-learning models and structural alerts. J Appl Toxicol 2019;39(8):1224-1232. https://doi.org/10.1002/jat.3808

[13] Zhou Y, Li S, Zhao Y, et al. Quantitative Structure-Activity Relationship (QSAR) model for the severity prediction of drug-induced rhabdomyolysis by using random forest. Chem Res Toxicol 2021;34(2):514-521. https://doi.org/10.1021/acs.chemrestox.0c00347

[14] Poorsarvi Tehrani P, Malek H. Early detection of rhabdomyolysis-induced acute kidney injury through machine learning approaches. Arch Acad Emerg Med 2021;9(1):e29. https://doi.org/10.22037/aaem.v9i1.1059

[15] Kim S, Chen J, Cheng T, et al. PubChem 2023 update. Nucleic Acids Res 2023;51(D1):D1373-D1380. https://doi.org/10.1093/nar/gkac956

[16] 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

[17] 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

[18] Python 3.9.5. Software. Available at: https://www.python.org/ (Accessed on August 20, 2024).

[19] Frank E, Hall MA, Witten IH. The WEKA workbench. Online appendix for "Data mining: Practical machine learning tools and techniques". 4th ed. Burlington, MA, USA: Morgan Kaufmann; 2016.

[20] 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;20:197–211. https://doi.org/10.1007/s41060-024-00509-w

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

[22] Sumner M, Frank E, Hall M. Speeding up logistic model tree induction. In: Jorge AM, Torgo L, Brazdil P, Camacho R, Gama J, editors. Knowledge Discovery in Databases: PKDD 2005. Berlin, Heidelberg: Springer; 2005: 675-683. https://doi.org/10.1007/11564126_72

[23] Hogg RV, Tanis EA. Probability and statistical inference. Macmillan: New York; 1977: 993.

[24] Organisation for Economic Co-operation and Development (OECD). Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)Sar] Models. In: OECD Series on Testing and Assessment. Paris: OECD Publishing; 2017: 1-154. https://doi.org/10.1787/20777876

[25] Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: Where have you been? Where are you going to? J Med Chem 2014;57(12):4977-5010. https://doi.org/10.1021/jm4004285

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

[27] Kaneko H. Cross-validated permutation feature importance considering correlation between features. Anal Sci Adv 2022;3(9-10):278-287. https://doi.org/10.1002/ansa.202200018

[28] 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

[29] Hall LH, Kier LB. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 1995;35:1039-1045. https://doi.org/10.1021/ci00028a014

[30] Roy K, Mitra I. Electrotopological state atom (E-state) index in drug design, QSAR, property prediction and toxicity assessment. Curr Comput Aided Drug Des 2012;8(2):135-158. https://doi.org/10.2174/157340912800492366

[31] Todeschini R, Consonni V. Molecular descriptors for chemoinformatics. Weinheim: Wiley VCH; 2009. https://doi.org/10.1002/9783527628766

[32] Ancuceanu R, Popovici PC, Drăgănescu D, Busnatu Ș, Lascu BE, Dinu M. QSAR regression models for predicting HMG-CoA reductase inhibition. Pharmaceuticals (Basel) 2024;17(11):1448. https://doi.org/10.3390/ph17111448

[33] Rajathei DM, Parthasarathy S, Selvaraj S. Combined QSAR model and chemical similarity search for novel HMGCoA reductase inhibitors for coronary heart disease. Curr Comput Aided Drug Des 2020;16(4):473-485. https://doi.org/10.2174/1573409915666190904114247

[34] Kier LB, Hall LH. Molecular connectivity in chemistry and drug research. New York: Academic Press; 1976.

[35] Naik AK, Kuppili V. Dynamic relevance and interdependent feature selection for continuous data. Expert Systems with Applications 2022;191:116302. https://doi.org/10.1016/j.eswa.2021.116302

[36] Keltz E, Khan FY, Mann G. Rhabdomyolysis. The role of diagnostic and prognostic factors. Muscles Ligaments Tendons J 2014;3(4):303-312.