Slovak Technical University in Bratislava

Faculty of Informatics and Information Technologies

FIIT-00000-000000

Bc. Matej Skyčák

Prediction Model for Dermatoxicity Determination

Master thesis

Thesis supervisor: Ing. Marta Šoltésová Prnová PhD.

May 2025

Slovak Technical University in Bratislava

Faculty of Informatics and Information Technologies

FIIT-00000-000000

Bc. Matej Skyčák

Prediction Model for Dermatoxicity Determination

Master thesis

Study programme: Informatics

Study field: Computer Science

Workplace: Institute of Informatics and Software Engineering, FIIT STU, Bratislava

Thesis supervisor: Ing. Marta Šoltésová Prnová PhD.

May 2025

Declaration of honor

I hereby declare that I have independently wrote this master's thesis using the referenced literature.

Bratislava, May 2025 .....

Acknowledgement

I would like to thank my thesis supervisor Ing. Marta Šoltésová Prnová PhD. for the time spent, the patience, and willingness to lead this work.

Also many thanks to my family for supporting my studies and always being there for me. Last but not least, I want to express gratitude to my friends for all their help and for keeping me motivated.

Annotation

Slovak University of Technology Bratislava
Faculty of Informatics and Information Technologies
Degree Course: Informatics

Author: Bc. Matej Skyčák
Master Thesis: Prediction Model for Dermatoxicity Determination
Supervisor: Ing. Marta Šoltésová Prnová PhD.
May 2025

In today's world, where a growing number of chemicals are synthesized, it is essential to develop efficient and reliable methods for assessing their potential toxicity. This master's thesis focuses on predicting the dermatotoxicity of chemical substances using machine learning models based on Quantitative Structure-Activity Relationships (QSAR). The primary objective was to create models capable of accurately determining whether a particular substance causes skin irritation, thereby contributing to health and environmental protection.

The thesis includes theoretical insights into toxicology and machine learning, as well as the implementation and evaluation of models such as Random Forest, XGBoost, and Support Vector Classifier. These models were trained on datasets of real chemical substances and optimized through various descriptor selection methods. Results demonstrate high accuracy and reliability, with the XGBoost model achieving the best performance with an AUC score of 0.9508.

This approach provides both scientific contributions and practical applications, such as reducing animal testing and accelerating chemical assessments. The work represents a step toward more sustainable and ethical solutions in the field of chemical safety.

Anotácia

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
Študijný program: Informatika

Autor: Bc. Matej Skyčák
Diplomová práca: Prediction Model for Dermatoxicity Determination
Vedúci diplomovej práce: Ing. Marta Šoltésová Prnová PhD.
May 2025

V dnešnom svete, kde sa syntetizuje stále väčšie množstvo chemických látok, je nevyhnutné zabezpečiť efektívne a spoľahlivé metódy pre posúdenie ich potenciálnej toxicity. Táto diplomová práca sa zaoberá predikciou dermatotoxicity chemických látok prostredníctvom modelov strojového učenia, ktoré využívajú kvantitatívne vzťahy medzi štruktúrou a aktivitou (QSAR). Hlavným cieľom bolo vytvoriť modely schopné presne určiť, či konkrétna látka spôsobuje podráždenie kože, a tým prispieť k ochrane zdravia a životného prostredia.

Práca zahŕňa teoretické pozadie v oblasti toxikológie a strojového učenia, ako aj implementáciu a vyhodnotenie modelov, vrátane Random Forest, XGBoost a Support Vector Classifier. Modely boli trénované na dátových sadách reálnych chemických látok a optimalizované pomocou rôznych výberových metód deskriptorov. Výsledky ukazujú vysokú presnosť a spoľahlivosť modelov, pričom najlepšie výsledky dosiahol model XGBoost s AUC hodnotou 0,9508.

Predložený prístup prináša nielen vedecký prínos, ale aj praktické využitie, napríklad pri redukcii testovania na zvieratách a urýchlení hodnotenia chemických látok. Táto práca tak predstavuje krok smerom k udržateľnejšiemu a etickejšiemu riešeniu v oblasti chemickej bezpečnosti.

Contents

Chapter 1

Introduction

Machine learning has been beneficial in many aspects of research, with toxicology prediction being one of them. The main objective of this master thesis is the prediction of dermatotoxicity using machine learning models based on the structure of the examined substances called QSAR (Quantitative Structure-Activity Relationship). With the help of these models, we have tried to determine whether the substance is toxic or not, specifically, we looked into the skin irritation effects.

The fact that the results of this thesis can have a direct impact in the area of health protection and the environment highlights the importance of further research and implementation of the proposed methodologies, ensuring that advancements contribute to sustainable and safe practices for society. The use of machine learning in this field can significantly accelerate and increase the effectiveness of the process of identifying potentially harmful substances, which is especially important nowadays given the vast number of newly synthesized chemicals.

The testing of substances in this manner is recognized and accepted by the Global Organization for Economic Cooperation and Development (OECD), which means

that the results provided by the QSAR models are accepted internationally. Additionally, this approach has the potential to reduce the number of tests conducted on animals, which is not only ethically preferred, but also legislatively supported in many countries [1].

Building such models based on the structure of the substance to predict dermatotoxicity requires extensive training sets. These sets contain information about both toxic and non-toxic chemical substances, which enables us to find connections between structure and toxicity. This master thesis is a combination of the necessary theoretical background in machine learning and toxicology and practical implementations with the primary objective of creating a tool to predict skin irritation of chemical substances. We hope that the results of this thesis will bring a better understanding of the problem and possibly take it a few steps in the right direction to help protect health and the environment.

Chapter 2

Related Work

2.1 Skin Sensitization Quantitative QSAR Models Based on Mechanistic Structural Alerts

2.2 Skin sensitization QSAR models based on structural alerts

The paper "Skin Sensitization Quantitative QSAR Models Based on Mechanistic Structural Alerts" focuses on the development and validation of QSAR (Quantitative Structure-Activity Relationship) models for predicting the skin sensitization potential of chemicals. The authors present eight multilinear regression models, which integrate structural alerts related to chemical reactivity. These models were developed using data from the Local Lymph Node Assay (LLNA), specifically EC3 values, for 366 chemicals. Each model uses 36 descriptors to predict sensitization potential.

Key insights from the study include:

- **Model development:** The models are based on the largest available dataset of LLNA EC3 values, sourced from public databases. Each model uses around 10 chemicals per descriptor to ensure statistical reliability.
- **Mechanistic insights:** The study emphasizes the importance of understanding the chemical mechanisms behind skin sensitization. The models are compared with previous research, highlighting their alignment with known chemical behaviors.
- **Risk assessment application:** These models are valuable for assessing the skin sensitization risk of chemicals, especially in regulatory settings such as cosmetics safety assessments. This can help replace or reduce the need for animal testing.
- **Practical use:** The results can enhance tools like SpheraCosmolife, which support risk assessments for cosmetics, aligning with the Next Generation Risk Assessment (NGRA) approach aimed at improving human health safety.
- **Statistical performance:** The models show strong statistical validation, with measures like R^2, Q^2 (cross-validated R^2), and external validation metrics (e.g., Q^2F1, Q^2F2) demonstrating their robustness and predictive accuracy.

This study provides robust QSAR models that offer reliable predictions of skin sensitization potential. By incorporating mechanistic insights, the models contribute to advancing safety evaluations, particularly in the cosmetic industry, and can support regulatory decision-making. Additionally, the strong predictive performance and statistical validation of the models make them a promising alternative to traditional animal testing, supporting the shift towards more ethical and efficient risk

assessment practices in chemical safety.

Chapter 3

Problem analysis

This chapter outlines the theoretical framework of the thesis, focusing on dermatotoxicity and its assessment using in vivo, in vitro, and in silico methods. It emphasizes Quantitative Structure-Activity Relationship (QSAR) modeling as a key tool for predicting toxicity using molecular descriptors and structural properties, along with machine learning algorithms used for the prediction of toxicity.

3.1 Dermatotoxicity

Toxicology is a field of science that helps us understand the harmful effects that substances can have on people, animals, or the environment. Its origins are ancient, starting well before the recorded history. The early references typically involve poisons or potions. In the book Dermatotoxicity, there is a notable connection between poisons and remedies, encapsulated by Paracelsus's famous quote: "All substances are poisons; there is none which is not a poison. The right dose differentiates a poison and a remedy." Traditionally, toxicology was descriptive, focusing on individual cases of poisoning with clear immediate effects. This knowledge was

often used to prevent future incidents or refine poisoning methods [2].

In the twentieth century, toxicology evolved significantly, with the focus shifting from descriptive approaches and acute poisoning cases to predicting and estimating toxicity in diverse exposure scenarios. Toxicologists now often need to predict adverse health effects at exposure levels much lower than those with human data available, and for substances without human data at all [2, 3].

The emphasis has moved from acute effects to long-term diseases such as cancer, which can manifest years after exposure to a toxicant. These diseases occur at low incidences, adding to the baseline rates of cancer expected in large populations. The current concern is more about the effects at the population level rather than individual cases of poisoning. Consequently, toxicologists must extrapolate data across species, from high to low exposure levels, and from small-scale studies to large populations. Human data is often insufficient, requiring the reliance on animal studies, tissues, and molecular systems at exposure levels higher than typical human exposures. Thus, multiple layers of extrapolation are necessary to accurately predict human risk [2].

The term **dermatotoxicity** refers to the adverse reactions of the skin when exposed to harmful chemical substances. This type of toxicity can cause significant damage depending on how deeply the chemicals penetrate the different layers of the skin. This exposure can cause a variety of skin conditions, including irritation, allergic contact dermatitis, contact urticaria, and various forms of photosensitization. In addition, other skin reactions can occur as a result of this exposure. To protect public health, it is crucial to thoroughly test new products and industrial chemicals for their potential dermatotoxic effects. This process helps identify and mitigate the risks associated with exposure of the skin to harmful substances, ensuring that the products are safe for consumer use and preventing potential

skin-related health problems [4].

When predicting and analyzing dermatotoxic effects, it is crucial to consider endpoints such as skin and eye irritation, corrosion, and sensitization, including photoallergy. In addition, assess skin carcinogenicity and dermal genotoxicity. Considering cumulative exposure, occupational dermatotoxicity, and contact dermatitis provides a comprehensive view. Utilizing in vivo and in vitro tests in combination with computational models ensure a thorough assessment of dermatotoxicity.

There are many types of tests for which it is possible to predict toxicity in vivo, in vitro, and in silico. Each method plays their role in understanding toxicity with their pros and cons.

- **In vivo** toxicity testing involves studying the effects of a substance within a whole living organism, such as animals or humans, providing comprehensive data on biological responses. Typical testing objects are rats or rabbits. The purpose of this test was to monitor acute toxicity responses in the system. These tests can be partially replaced by other alternatives; however, this type of testing might be required by the authorities [5].

- **In vitro** toxicity testing uses cell cultures or isolated tissues to examine toxic effects in a controlled environment outside of a living organism. It offers insights into the direct impact of toxins on cells with the cost being significantly lower compared to in vivo tests. This type comes with its disadvantages, two of them being lack of toxin elimination and cell-to-cell interaction [5].

- **In silico** toxicity testing utilizes computer-based models and simulations to predict toxicological effects based on existing data and theoretical approaches gained from the substances. They complement the classical in vivo and in vitro tests, thereby reducing the need for animal testing even more, while

saving time and possibly improving the predictions by trying a different approach. The quantitative structure-activity relationship is a group of models used to predict toxicity based on molecular decriptors [6].

Contact dermatitis (CD) is an inflammatory skin condition caused by irritant chemicals, metal ions, or contact allergens, leading to two main types: **irritant CD** and **allergic CD**. Irritant CD arises from direct chemical damage to the skin, causing inflammation without involving the immune system or requiring prior sensitization, and tends to affect most individuals exposed to aggressive substances; this is the type we are trying to predict. In contrast, allergic CD is a delayed hypersensitivity reaction (type IV) driven by T-cell-mediated immune responses triggered by allergens that modify skin proteins. CD can be the result of various factors, including irritants, allergens, UV radiation, microbes, or autoimmune responses, and can occur acutely or chronically. Occupational CD is also common, linked to workplace exposures such as disinfectants, detergents, or latex gloves. Although irritant CD is more prevalent, allergic CD affects an estimated one fifth of the population. [7]

3.2 QSAR

The QSAR models, as the name implies, are a family of models that find mathematical correlation between the qualitative molecular structures of substances and their activity, or, in our case, their toxicity. In this context, what we are trying to predict is also called the endpoint. Examples of such structural properties used for the development of a model might include the melting point, the boiling point, or the surface tension. QSAR modeling serves as a powerful tool for predicting the behavior of new or hypothetical chemicals, allowing researchers to estimate their toxicological and biological activities without the need for extensive experimental

or animal testing [8, 6].

There are other types of model depending on the specific correlations we aim to analyze. For example, we know the QSPR (Quantitative Structure-Property Relationship) and the QSTR (Quantitative Structure-Toxicity Relationship). The QSTR family of models is relevant to our work, as we are trying to find a correlation between the structures of a substance and the dermatoxicity of [8].

A general QSAR model for predicting toxicity (T) is typically expressed as a function of a feature vector of chemical properties:

$$T = f(vector)$$

This function f translates the chemical properties into a predicted toxicological response. The development of a QSAR model involves generating molecular descriptors, which can be obtained from both theoretical calculations and experimental measurements. Selecting those descriptors that significantly correlate with the toxicity in question, might also be considered [8].

QSAR models can be classified into local and global models. Local QSAR models are developed from congeneric chemicals, compounds with similar structures, and are typically more accurate because they are tailored to a specific group of chemical substances. However, developing local QSAR models for each type of chemical can be resource-intensive. On the other hand, global QSAR models are derived from diverse sets of chemicals and, while generally more practical and broadly applicable, they may lack the precision and performance of local models [6, 9].

Toxicity prediction using QSAR models offers some advantages:

- It is possible to model categorical and continuous endpoints [6].
- The interpretation is simple, based on the descriptors used and the complexity of the model [6].

However, there are some disadvantages and disadvantages with QSAR models:

- Dose and duration of exposure to the toxin in not taken into consideration [6]
- Training of such model requires large models, which are hard to obtain [6].

3.2.1 Molecular descriptors

A molecular descriptor is a quantitative parameter that characterizes specific information about a molecule, serving as numerical values associated with its chemical structure. Several types of descriptors based on the dimension on the property are known: **0D, 1D, 2D, 3D**. Each group tells about some type of properties of a given molecule. For example, for 0D it is the number of atoms or the number of bonds, for 2D it is the topological or structural parameters, and for 3D it is the electronic or spatial parameters [6, 10].

3.3 Machine learning

Machine learning provides sophisticated algorithms that can handle the complex, non-linear relationships often found in chemical data. These algorithms automate the learning process from large datasets, improving the ability to predict the biological activity of new compounds, for example, skin irritation in our case [11, 12].

In our implementation, we have used several machine learning models to find which

offers the best results for predicting skin irritation. To understand these models in more depth, we will cover them theoretically in the following sections.

3.3.1 Support Vector Classifier (SVC)

Support Vector Classification (SVC) is a supervised machine learning algorithm that aims to find the optimal hyperplane that separates data into distinct classes by maximizing the margin between them. The concept of maximization of the margin is essential for improving generalization and avoiding overfitting. SVC works by determining the support vectors, the data points closest to the decision boundary, which directly influence the position of this hyperplane. The optimization problem in SVC involves maximizing the margin while minimizing classification errors. This trade-off is controlled by the regularization parameter, C, which influences the model's ability to generalize. A higher C value prioritizes minimizing errors, which can lead to overfitting, while a lower value emphasizes maximizing the margin but allows for more misclassifications [13, 14].

One of the key strengths of SVC is its ability to handle non-linear data using kernel functions. Kernels, such as linear, polynomial, and radial basis function (RBF) kernels, map data to a higher-dimensional space where linear separation is possible. Choosing the right kernel is crucial for model performance. SVC's training involves solving a quadratic optimization problem, and although it can be computationally expensive, especially with non-linear kernels, it is effective in high-dimensional spaces. Its applications are wide-ranging, from text classification to bioinformatics, particularly when dealing with complex datasets [13, 14].

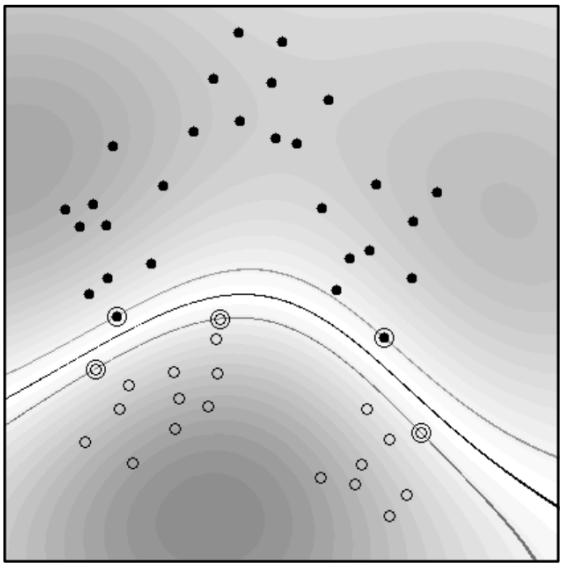

Figure 3.1: A two-dimensional example of SVC with radial basis function kernel splitting points of 2 classes using optimal hyperplane from a book *Learning with Kernels* [13]

3.3.2 Random Forest

Random Forest is an ensemble learning algorithm designed for classification and regression tasks, built upon the foundation of decision trees. A decision tree is a model that splits data recursively based on feature values to create a tree-like structure, where each internal node represents a decision rule, and each leaf node corresponds to an output class or prediction. While decision trees are intuitive and interpretable, they are prone to overfitting, particularly when grown to their full depth. This overfitting occurs because a single tree can capture noise or irrelevant patterns in the training data, reducing its generalization ability [15, 16, 11].

To mitigate these issues, Random Forest combines the outputs of multiple decision trees, improving accuracy and robustness. The algorithm utilizes bagging (Bootstrap Aggregating), where each tree is trained on a randomly sampled subset of the data with replacement. Additionally, Random Forest introduces random-

ness during tree construction by selecting a random subset of features for each split. These techniques ensure diversity among the trees, reducing overfitting and enhancing generalization. By aggregating predictions—using a majority vote for classification or averaging for regression—Random Forest creates a more stable and reliable model [15, 16].

Random Forest also offers insights into feature importance, making it a valuable tool for feature selection tasks. However, its computational demands can increase with the number of trees or the dataset size. Despite these challenges, the algorithm is widely used in fields such as bioinformatics, risk prediction, and image classification, where robust performance is essential [15, 16].

3.3.3 Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced machine learning algorithm that builds on the gradient boost framework. In gradient boosting, an ensemble of weak learners, typically decision trees, is constructed iteratively. Each new tree is trained to correct the errors made by the previous ones, focusing on the residuals or gradients of the prediction errors. The model improves progressively by minimizing a specified loss function, such as the mean squared error for regression or the log loss for classification. XGBoost enhances this framework by introducing several key innovations, such as regularization (L1 and L2), which helps to control model complexity and reduce the risk of overfitting, especially in complex datasets. Additionally, XGBoost uses a sophisticated tree pruning strategy, stopping tree growth early when further splitting does not significantly reduce loss, thus building more compact and interpretable models [17, 18].

The core of XGBoost's efficiency lies in its optimization techniques. It implements a gradient-based optimization algorithm with second-order approximation (using

both the first and second derivatives of the loss function), which leads to faster convergence and better model performance compared to traditional gradient boosting methods. The algorithm also uses a shrinkage technique, adjusting the learning rate after each iteration to ensure stability and prevent overshooting. XGBoost supports parallelization, allowing it to train on large datasets quickly, and its use of a weighted objective function allows it to handle various loss functions and fine-tune model performance [17, 18].

Chapter 4

Solution proposal

In this chapter we will go through the proposed solution for our implementation of QSAR model for predicting skin irritation.

4.1 Our approach

The visualisation of our approach is in the diagram 4.1. This approach can be used basically on any dataset, where we have chemical substances and some target value, either continous (Simulation Index) or descrete (Toxic or non-toxic). At first, EDA and dataset pre-processing has to be done, removing rows with missing values and duplicates. Next, we use the PubChem API to fetch the SMILES Code from their database. Some substances might not be found or have corrupt name, those are deleted from the dataset. Based on the SMILES Code, we use another library RDKit to calculate molecular descriptors. Some descriptors hold 0 for each substance, we can say that those do not provide any additional information, so we remove those. After that based on the correlation matrix, where we calculate correlation of each descriptors to the target variable, we select those with the

highest correlation. This feature selection saves computational time and might prevent over-fitting of the model on the training set. The dataset is split into 3 parts training set, which is used for training the model, validation set to find the best hyper-parameters and test set to evaluate the performance of the best model.

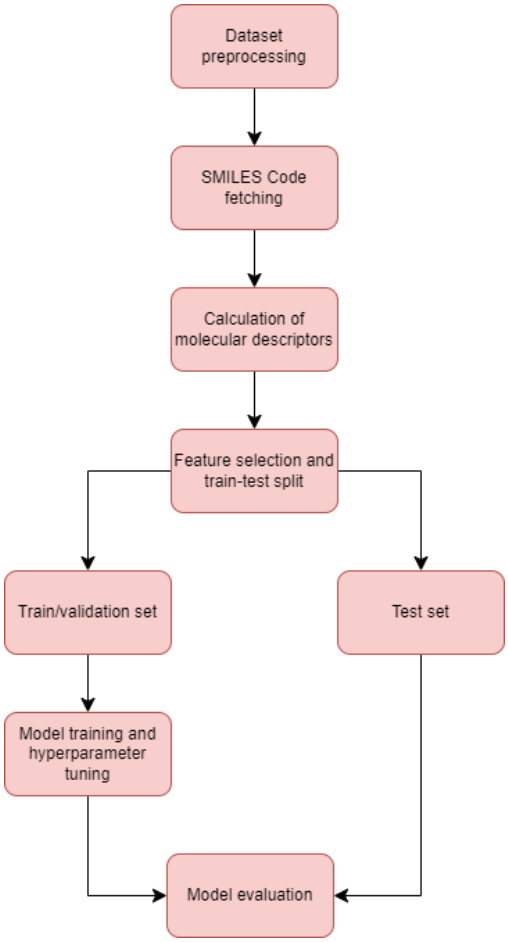

Figure 4.1: Proposed solution for building QSAR Model

Chapter 5

Implementation

We have build a model on a dataset of almost 200 substances to predict skin irritation using scikit-learn. We are limited in data resources, therefore we needed to pick models, which should perform well also on the smaller amounts of data. We chose 3 different models which we trained and optimised, namely Random Forest, Support Vector Classifier and XGBoost.

5.1 Dataset

The data set on which we trained our models was obtained from the thesis supervisor. It is a collection of 190 entries, each representing an in vitro test with the result indicating whether a substance caused irritation. The format of the data is as follows: Name, CAS number, SMILES code and result (0 or 1).

SMILES code has been used to calculate the molecular descriptors, specifically 210 descriptors for each compound. We have checked if the features have zero variance or hold only 0 values across the whole dataset. We found 35 features with these characteristics, so we dropped them. We also checked for duplicates, which

were not present in our dataset. Also, duplicate rows with the same substance and substances without SMILES code have also been removed. The ratio between irritant and non-irritant should ideally be balanced, but in our case there is a slight imbalance with approximately two thirds being irritant substances and one third non-irritant. For easier use we saved this dataset. All of the molecular descriptors have been scaled using Min-Max scaler for better model performance.

5.2 Model

As already mentioned, we have built 3 different models, we started on the default parameters provided by the model implementations in Python. Then we predicted whether the substance is a skin irritant or not. For each model, we performed feature selection trying 3 different subsets (10, 20, 30) of the most predictive attributes with the target value. This was performed using 3 different feature selection methods, to further optimize our models. There are many different ways to obtain the most predictive attributes; we used correlation, SHAP values, and mutual information. For training, we used 80% of the dataset and the remaining 20% for the testing of the model. After initial training we performed hyper-parameter tuning to find the parameter values with the best results. This was performed using a grid search where we defined a set of values to test. After that, the model was also cross-validated using five folds to analyze the robustness and consistency of the models. In the end we evaluated the models using classification report which includes accuracy, precision, recall, f1-score, AUC.

The best models were serialized and saved for further prediction, which will be done later by prompting a substance using a user interface, and using the best model we will try to predict the outcome.

Chapter 6

Evaluation

To evaluate the performance of a previously trained model, several commonly used metrics are used. These include accuracy, precision, recall, and F1 score, each of which provides valuable insights into the model's effectiveness and its ability to make accurate predictions across different classes.

6.1 Models evaluation

To evaluate the models, several steps were followed to ensure a thorough assessment of their performance. First, the dataset was split into training and test sets using a stratified approach to maintain class distribution. Various feature selection techniques, including correlation-based, mutual information-based, and SHAP-based methods, were applied to identify the most relevant features for model training.

Each model was then trained with hyperparameter tuning using GridSearchCV to optimize parameters such as max_depth, learning_rate, and n_estimators, selecting the best configuration based on the AUC (Area Under the Curve) metric.

Model performance was evaluated using key metrics: accuracy, precision, recall, and F1 score. These metrics were obtained from the classification report, providing insights into the model's effectiveness for both classes. Additionally, the AUC score was calculated to assess the model's ability to distinguish between the classes, with higher values indicating better performance.

The results were compared across different models and feature selection methods to identify the most effective model for predicting skin irritation. This evaluation process provided a comprehensive view of each model's strengths and allowed for informed decision-making regarding model selection.

6.1.1 Random forest

In this evaluation, the performance of the Random Forest model was assessed across different feature selection techniques with varying numbers of features (10, 20, and 30). The best results can be seen in Table 6.1.

Key findings

- **Best performing method:** The highest AUC score of 0.9697 was achieved by the SHAP-Based method using the top 30 features. This indicates that the model is highly effective at distinguishing between the classes and has a strong predictive capability.
- **AUC score:** The AUC score, which measures the model's ability to rank positive instances higher than negative ones, was also highest with the SHAP-based method (top 30 features) at 0.9697. This is a strong indicator of the model's robustness and ability to discriminate between the two classes (irritant vs. non-irritant).
- **Precision, recall, and F1-score:** The SHAP-based (top 30 features)

method also performed well across these metrics. The precision and recall for both classes (0 and 1) were balanced, indicating that the model correctly classified both classes without significant bias. The F1-score, which combines precision and recall, was also high for both classes.

- **Feature selection:** The correlation-based and mutual information-based methods showed good performance with AUC scores but the SHAP-based method outperformed them, particularly for the top 30 features, where it achieved the highest AUC and other evaluation scores.

Features	Method	Hyperparameters	AUC	Precision	Recall	F1
10	Corr-Based	max_depth: 10, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 150	0.93	0.79	0.79	0.79
20	SHAP-Based	max_depth: 10, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 15	0.90	0.92	0.91	0.91
30	SHAP-Based	max_depth: None, min_samples_leaf: 2, min_samples_split: 2, n_estimators: 50	0.96	0.95	0.94	0.94

Table 6.1: Best results for random forest with different feature selection methods and feature counts

6.1.2 XGBoost

The performance of the XGBoost model was evaluated using different feature selection techniques (correlation-based, mutual information-based, and SHAP-based) with varying feature counts (10, 20, and 30). The key findings are summarized below. The results are displayed in the Table 6.2.

Key findings

- **Best performing method:** The SHAP-based method with 20 features

achieved the highest AUC score of 0.9508. This demonstrates that SHAP-based feature selection effectively identifies the most predictive features for this dataset.

- **AUC score:** While the highest AUC was observed with SHAP-based (20 features), correlation-based and mutual information-based methods also performed well, achieving AUC scores exceeding 0.91 for larger feature sets (20 and 30 features).

- **Precision, recall, and F1-score:** The SHAP-based (20 features) method provided strong classification metrics across all measures. Balanced precision and recall highlight the model's robustness and reliability in predicting both classes.

- **Feature selection comparison:** Although mutual information-based and correlation-based methods performed competitively, SHAP-based consistently outperformed them in terms of AUC and overall classification performance, particularly with 20 features.

Features	Method	Hyperparameters	AUC	Precision	Recall	F1
10	SHAP-Based	learning_rate: 0.2, max_depth: 3, n_estimators: 150	0.88	0.78	0.76	0.77
20	SHAP-Based	learning_rate: 0.1, max_depth: 3, n_estimators: 150	0.95	0.88	0.88	0.88
30	Corr-Based	learning_rate: 0.1, max_depth: 5, n_estimators: 100	0.95	0.91	0.94	0.94

Table 6.2: Best results for XGBoost with different feature selection methods and feature counts

6.1.3 Support Vector Classifier (SVC)

The performance of the Support Vector Classifier (SVC) was evaluated using correlation-based and mutual information-based feature selection methods with varying feature counts (10, 20, and 30). The key findings are summarized below. We have not used SHAP method, because we couldn't train the model, for unkown reason, the training was not stopping. The results are displayed in Table 6.3.

Key findings

- **Best performing method:** The best model was the **rbf kernel** with **30 mutual information-Based features**, achieving the highest **AUC score of 0.9091**. This shows that the rbf kernel performs optimally with the selected features, excelling in distinguishing between irritant and non-irritant cases.
- **AUC score:** The **rbf kernel** with **30 mutual information-based features** outperformed all other models in terms of **AUC**, highlighting its ability to capture complex non-linear relationships.
- **Accuracy and class performance:** The **linear kernel** with **30 correlation-based features** achieved the highest **accuracy of 88.24%**, showing solid performance with linear decision boundaries. However, the **rbf kernel** consistently had better **AUC scores**, making it a more robust choice despite slightly lower accuracy.
- **Precision, recall, and F1-score:** The **linear kernel** with **30 correlation-based features** demonstrated the best balance of **precision** and **recall**, particularly for class '1' (Irritant), achieving high F1-scores. Conversely, the **poly kernel** models showed significant weaknesses in class '0' predictions, with low precision and recall.

- **Feature selection comparison:** The **correlation-based method** generally outperformed **mutual information-based selection** for the **linear kernel**, while the **mutual information-based method** was more suitable for **rbf kernels**, producing the highest **AUC score** and optimal performance with non-linear relationships.

Features	Method	Kernel	AUC	Precision	Recall	F1
10	Correlation-Based	Linear	0.87	0.76	0.76	0.75
20	Correlation-Based	Polynomial	0.89	0.81	0.74	0.68
30	Mutual Information-Based	Radial basis function	0.90	0.72	0.71	0.65

Table 6.3: Best results for SVC with different feature selection methods and feature counts

6.2 Summary

In this study, several machine learning models, including Random Forest, SVC, and XGBoost, were evaluated using different feature selection techniques (Correlation-Based and Mutual Information-Based) with varying feature counts. Among the models, XGBoost achieved the highest AUC score of 0.9508 with the SHAP-Based feature selection method and 20 features, demonstrating its strong capability in identifying the most predictive features. Random Forest performed competitively, with its best result coming from the SHAP-Based method with 20 features, achieving an AUC of 0.9446. SVC, while generally performing slightly lower than XGBoost and Random Forest, showed its best result using the Radial Basis Function (RBF) kernel with 30 features from the Mutual Information-Based selection, yielding an AUC of 0.9091. Overall, XGBoost emerged as the top-performing model, providing robust and reliable predictions for skin irritation, with SHAP-Based feature selection contributing significantly to its success.

Chapter 7

Conclusion and Future Work

This thesis focused on predicting the dermatotoxicity of chemical substances using QSAR-based machine learning models. We successfully trained and evaluated three models: Random Forest, XGBoost, and Support Vector Classifier, with XGBoost showing the best performance.

In the future, we plan to expand the work by integrating two additional models to explore alternative approaches and improve accuracy. Additionally, we aim to create a user-friendly interface to input chemical substances and determine their irritant potential using the trained models.

These developments will enhance the practical applicability of our work, contributing to safer chemical usage and reducing reliance on animal testing, aligning with ethical and scientific standards in toxicology.

References

[1] OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. OECD, Sept. 2014. ISBN: 9789264085442. DOI: 10.1787/9789264085442-en. URL: http://dx.doi.org/10.1787/9789264085442-en.

[2] F.N. Marzulli and H.I. Maibach. *DERMATOTOXICOLOGY 4ED*. Taylor & Francis, 1991. ISBN: 9781560320555. URL: https://books.google.co.in/books?id=w4cDg59Bz30C.

[3] Lewis Casarett and John Doull. *Toxicology: Basic science of poisons*. McGraw-Hill, 1992.

[4] N Ali and FW Oehme. "A literature review of dermatotoxicity". In: *Veterinary and human toxicology* 34.5 (1992), 428—437. ISSN: 0145-6296. URL: http://europepmc.org/abstract/MED/1455613.

[5] Magda Sachana and Alan J. Hargreaves. "Chapter 9 - Toxicological Testing: In Vivo and In Vitro Models". In: *Veterinary Toxicology (Third Edition)*. Ed. by Ramesh C. Gupta. Third Edition. Academic Press, 2018, pp. 145–161. ISBN: 978-0-12-811410-0. DOI: https://doi.org/10.1016/B978-0-12-811410-0.00009-X. URL: https://www.sciencedirect.com/science/article/pii/B978012811410000009X.

[6] Arwa B. Raies and Vladimir B. Bajic. "In silico toxicology: computational methods for the prediction of chemical toxicity". In: *WIREs Computational Molecular Science* 6.2 (Jan. 2016), 147–172. ISSN: 1759-0884. DOI: 10.1002/wcms.1240. URL: http://dx.doi.org/10.1002/wcms.1240.

[7] Gaby Novak-Bilić. "Irritant and Allergic Contact Dermatitis – Skin Lesion Characteristics". In: *Acta Clinica Croatica* (2018). ISSN: 0353-9466. DOI: 10.20471/acc.2018.57.04.13. URL: http://dx.doi.org/10.20471/acc.2018.57.04.13.

[8] Kunal Roy, Supratik Kar, and Rudra Narayan Das. *A Primer on QSAR/QSPR modeling fundamental concepts*. Springer International Publishing, 2015.

[9] Christoph Helma. *Predictive Toxicology*. Mar. 2005. DOI: 10.1201/9780849350351. URL: http://dx.doi.org/10.1201/9780849350351.

[10] Roberto Todeschini and Viviana Consonni. *Molecular descriptors for Chemoinformatics*. Wiley-VCH, 2009.

[11] Batta Mahesh. "Machine learning algorithms-a review". In: *International Journal of Science and Research (IJSR).[Internet]* 9.1 (2020), pp. 381–386.

[12] Artem Cherkasov et al. "QSAR Modeling: Where Have You Been? Where Are You Going To?" In: *Journal of Medicinal Chemistry* 57.12 (Jan. 2014), 4977–5010. ISSN: 1520-4804. DOI: 10.1021/jm4004285. URL: http://dx.doi.org/10.1021/jm4004285.

[13] Bernhard Scholkopf and Alexander J. Smola. *Learning with kernels support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[14] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine Learning* 20.3 (Sept. 1995), 273–297. ISSN: 1573-0565. DOI: 10.1007/bf00994018. URL: http://dx.doi.org/10.1007/BF00994018.

[15] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), 5–32. ISSN: 0885-6125. DOI: 10.1023/a:1010933404324. URL: http://dx.doi.org/10.1023/A:1010933404324.

[16] J. R. Quinlan. "Induction of decision trees". In: *Machine Learning* 1.1 (Mar. 1986), 81–106. ISSN: 1573-0565. DOI: 10.1007/bf00116251. URL: http://dx.doi.org/10.1007/BF00116251.

[17] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. ACM, Aug. 2016, 785–794. DOI: 10.1145/2939672.2939785. URL: http://dx.doi.org/10.1145/2939672.2939785.

[18] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5 (2001), pp. 1189 –1232. DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.