Predictive Modeling of Drug-Induced Autoimmunity: A Machine Learning and Descriptor-Based Approach

Madhura Bokil[1]

[1] Harrisburg University of Science and Technology

Author Note

This project is submitted in fulfillment of the requirements for the ANLY 699 Applied Project in Analytics.

Correspondence concerning this article should be addressed to Madhura Bokil, 326 Market St, Harrisburg, PA 17101. E-mail: mbokil@my.harrisburgu.edu

Abstract

Drug induced autoimmunity (DIA), is a severe drug reaction, characterized by unintended immune responses that are elicited by drug molecules (Lopez, Wichmann, & Chen, 2020). Early prediction of DIA risk during drug development can significantly enhance patient safety and reduce the economic burden associated with drug recalls (Huang, Liu, & Huang, 2025). This study proposes an XGBoost-based predictive model trained on RDKit chemical descriptors for classifying compounds based on DIA risk. The model addresses inherent class imbalance using the Synthetic Minority Oversampling Technique (SMOTE) and employs Optuna for rigorous hyperparameter optimization. The final model demonstrated substantial predictive ability (AUC = 0.73), highlighting the potential of machine learning to assist in pharmaceutical toxicology screening (Huang et al., 2025).

Keywords: Drug-Induced Autoimmunity, Machine Learning, XGBoost, RDKit, SMOTE, Predictive Toxicology

Word count: 5238

Predictive Modeling of Drug-Induced Autoimmunity: A Machine Learning and Descriptor-Based Approach

Drug-induced autoimmunity (DIA) represents a critical area of concern in pharmaceutical safety and public health. Unlike other forms of drug toxicity, DIA arises when therapeutic compounds provoke aberrant immune responses targeting host tissues, causing conditions such as lupus-like syndromes, autoimmune hepatitis, and vasculitis (Lopez et al., 2020). The occurrence of DIA can severely impact patient outcomes and often results in the withdrawal of otherwise promising drugs from the market. Despite these significant impacts, accurately predicting which compounds might induce DIA remains an unresolved challenge within the drug development process.Structural characteristics of drug molecules play a key role in modulating DIA risk, as shown in prior modeling studies (Guo et al., 2022).

Conventional in vitro and in vivo screening methods for immunotoxicity are time-consuming, expensive, and typically lack the throughput needed for screening extensive compound libraries (Patel, Joshi, & Kuo, 2023). Additionally, there is a low availability of reliable biomarkers to help identify risks early during preclinical studies, making our understanding of how DIA works still incomplete. Consequently, there is a pressing need for computational approaches that leverage chemical structure and molecular properties to effectively predict DIA liabilities.

Recent advancements in machine learning (ML) and cheminformatics offer promising solutions in the diagnosis and prediction of autoimmune diseases (Stafford, Kellermann, & Mossotto, 2020). Molecular descriptors generated using computational chemistry toolkits, such as RDKit, provide comprehensive quantitative representations of chemical structures that can serve as inputs to predictive models (Hu, Song, & Li, 2024; Landrum, 2006). Specifically, gradient-boosted decision tree models, including XGBoost, have demonstrated superior performance on structured biomedical datasets due to their ability to handle complex nonlinear interactions and high-dimensional feature spaces effectively (Chen &

Guestrin, 2016).

In addition, hyperparameter optimization frameworks, like Optuna, facilitate the fine-tuning of model architectures and training parameters, thus enhancing generalization and predictive accuracy (Akiba, Sano, Yanase, Ohta, & Koyama, 2019). Class imbalance, a common issue in toxicity datasets where toxic compounds (positives) are typically underrepresented, can be addressed using techniques such as the Synthetic Minority Oversampling Technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Other machine learning approaches have also been proposed for predicting DIA risk using structural alerts and molecular descriptors (Yao et al., 2021).

This study aims to develop a reproducible ML pipeline integrating data preprocessing, class balancing via SMOTE, hyperparameter optimization with Optuna, and thorough performance evaluation. The data used for this research was sourced from a recent validated study published in Toxicology, titled “InterDIA: Interpretable Prediction of Drug-induced Autoimmunity through Ensemble Machine Learning Approaches” by Huang et al. (Huang et al., 2025). The authors of this previous work provided openly available training and external validation datasets, ensuring reproducibility and facilitating further methodological comparisons. Beyond the foundational work presented in InterDIA, there remains ample opportunity to further refine and extend predictive models for DIA. Other researchers have explored a variety of machine learning approaches for predictive toxicology more broadly, demonstrating that advanced ML models can outperform traditional rule-based systems in detecting complex toxicity outcomes such as hepatotoxicity and cardiotoxicity (Grisoni et al., 2024; **tian2022?**). For immunotoxicity domain, Guo et al. (Guo et al., 2022) and Yao et al. (Yao et al., 2021) have highlighted the promise of structure-based modeling approaches for predicting autoimmune-related adverse drug reactions, reinforcing the utility of RDKit-derived descriptors as informative features for such tasks.

Recent advancements in automated ML frameworks, such as DeepMol (Grisoni et al., 2024), have further streamlined the process of building performant predictive models in cheminformatics applications. The rapid development of end-to-end pipelines is supported by the frameworks that integrate feature engineering, model selection, and hyperparameter tuning. At the same time, Optuna-based optimization has been shown to significantly improve the performance of tree-based classifiers, including XGBoost, in various chemical and biological modeling tasks (Akiba et al., 2019; Shehab, Nayel, & Taha, 2025).

Predictive modeling for DIA remains a relatively underexplored area within the broader field of computational toxicology, even with the current advances. As Stafford et al. (Stafford et al., 2020) observed in their systematic review of AI applications in autoimmune diseases, one key challenge is ensuring that ML models account for the inherent biological variability and class imbalance often present in biomedical datasets. Techniques such as the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) are essential for mitigating the effects of class imbalance, while interpretability tools, such as SHAP analysis, can help elucidate which molecular features most strongly drive model predictions (Kruta et al., 2024; Lundberg & Lee, 2017).

It is important to leverage reproducible and openly available datasets, such as those provided by Huang et al. (Huang et al., 2025). The InterDIA study is critical for promoting transparency and enabling meaningful methodological comparisons across studies. The InterDIA dataset offers a valuable resource for training and validating structure-based DIA prediction models and provides an ideal starting point for this project.

Building upon these insights, this study seeks to advance DIA prediction by constructing a reproducible and interpretable ML pipeline that leverages XGBoost, Optuna-based hyperparameter tuning, and SMOTE for class balancing. By systematically evaluating model performance on the InterDIA dataset and openly sharing both code and results, this work aims to contribute to the growing body of research in predictive toxicology and demonstrate the practical utility of ML approaches for improving early-stage drug safety screening. In parallel with these methodological advances, there is growing recognition of the role that AI and ML can play in regulatory toxicology and pharmaceutical decision-making (Patel et al., 2023; Stafford et al., 2020). Regulatory agencies, such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA), are increasingly open to the use of ML-based predictive models to augment traditional safety assessments, provided that these models are transparent, reproducible, and interpretable (Kruta et al., 2024; Lundberg & Lee, 2017). Developing models that meet these standards is particularly important in the context of DIA, where adverse outcomes are relatively rare but can have severe consequences for patients and drug developers alike.

Modeling rare adverse events such as DIA presents unique challenges for ML workflows. The limited availability of positive cases can exacerbate class imbalance and hinder model generalization (Chawla et al., 2002). Moreover, the underlying biological mechanisms of DIA are often heterogeneous and poorly characterized (Lopez et al., 2020), further complicating predictive efforts. The selection of the correct model is a challenge, and so is the effective communication of the model limitations to the end user. Addressing both these challenges effectively requires a deep understanding of the domains.

One promising direction for future modeling efforts involves the integration of chemical and biological data sources. For example, combining molecular descriptors with transcriptomic or proteomic profiles could enhance model sensitivity and provide greater biological insight (Kruta et al., 2024). Multi-modal modeling approaches are increasingly being adopted in other areas of predictive toxicology (Grisoni et al., 2024; Wang et al., 2020), and extending such strategies to DIA represents a logical next step for the field.

In addition to methodological innovation, there is a strong emphasis within the ML toxicology community on transparency and reproducibility. By building upon the InterDIA dataset and adhering to principles of reproducibility and openness, this study aims to contribute meaningfully to the evolving landscape of predictive immunotoxicology.

Purpose

The purpose of this study is to explore whether machine learning models trained on chemical descriptors can effectively predict the risk of drug-induced autoimmunity (DIA) in chemical compounds. Given the limitations of traditional experimental methods for detecting DIA, a predictive computational approach could help flag high-risk compounds earlier in the drug development process.

Using chemical descriptor data from the InterDIA study dataset (Huang et al., 2025), I aim to address the following research questions:

H1: An XGBoost-based model trained on RDKit chemical descriptors, with class balancing via SMOTE and hyperparameter tuning via Optuna, will achieve an AUC score significantly above random chance in predicting DIA risk on a held-out test set.

H2: Key molecular descriptors identified through model feature importance analysis will provide interpretable insights into chemical properties associated with higher DIA risk.

Methods

Data

This study utilized the InterDIA dataset published by Huang et al. (Huang et al., 2025), comprising both labeled and unlabeled chemical compounds. The labeled subset includes compounds previously evaluated for drug-induced autoimmunity (DIA), each assigned a binary classification of DIA-positive or DIA-negative. The unlabeled subset served to assess the model's capacity for prospective compound screening.

Each compound was represented by a set of chemical descriptors generated using RDKit (Landrum, 2006). The descriptors encompass diverse molecular characteristics—such as topology, surface area, functional groups, and electronic properties—that inform cheminformatics modeling.

Predictors

The input features consisted of numerical chemical descriptors calculated via RDKit (Landrum, 2006). Prior to modeling, non-numeric columns (e.g., SMILES strings) were excluded, resulting in a feature set that quantitatively described each compound's molecular structure. Selected descriptors included topological polar surface area (TPSA), fragment counts (e.g., fr_ketone), and charge-related properties (e.g., PEOE_VSA2), among others.

Outcome

The target variable was a binary indicator of DIA association and the compounds were labeled as DIA-positive (1) or DIA-negative (0) based on their known association with the condition.

Data Analytic Plan

The machine learning pipeline was implemented in Python, utilizing the scikit-learn and XGBoost libraries (Chen & Guestrin, 2016). The label data set was divided into-

Training- 60%, Test- 20% and Validation- 20%. Stratification was applied to ensure balanced class distributions across all subsets.

Given the class imbalance—marked by a limited number of DIA-positive compounds—the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) was applied to the training set to generate synthetic minority class instances and mitigate potential bias.

We selected XGBoost for its capacity to model structured, high-dimensional data effectively. Its widespread use in bioinformatics research provided additional justification for this selection. Hyperparameter optimization was performed using the Optuna framework (Akiba et al., 2019) (Shehab et al., 2025), which efficiently explored combinations of learning rate, tree depth, number of estimators, and class weighting.

The model was retrained using the combined training and validation data once the optimization was complete, and was then tested on an independent set. Its performance was evaluated using several metrics, including AUC, accuracy, precision, recall, and confusion matrix results Additionally, feature importance scores derived from the XGBoost model were analyzed to identify key chemical descriptors influencing DIA risk predictions.

In the final step, the trained model was used to estimate DIA-related probabilities for each compound in the unlabeled dataset These predictions can support early-stage screening and help prioritize compounds for further drug development efforts.

Results

The final XGBoost model, optimized using Optuna and trained on RDKit chemical descriptors, achieved an AUC score of 0.73 on the held-out test set, with an overall accuracy of 58.3%. This indicates that the model was able to distinguish between DIA-positive and DIA-negative compounds with reasonable accuracy, providing a meaningful improvement over random classification.

Table 1

Table 1. Classification report for final XGBoost model (AUC = 0.73, Accuracy = 0.58).

Class	Precision	Recall	F1_Score	Support
0	0.83	0.56	0.67	18
1	0.33	0.67	0.44	6

A confusion matrix was used to evaluate the model further. The model correctly classified most DIA-negative compounds but showed a typical tradeoff between sensitivity and specificity, with some false positives and missed DIA-positive predictions. The confusion matrix and classification report (Table 1) further illustrate the model's behavior. For DIA-negative compounds (Class 0), the model achieved a precision of 0.83, a recall of 0.56, and an F1 score of 0.67. For DIA-positive compounds (Class 1), the model attained a precision of 0.33, a recall of 0.67, and an F1 score of 0.44. These results indicate that while the model is able to identify a reasonable portion of DIA-positive compounds, improving precision for the minority class remains an opportunity for future refinement.This outcome is consistent with the challenges of modeling rare toxicological events in imbalanced biomedical datasets.

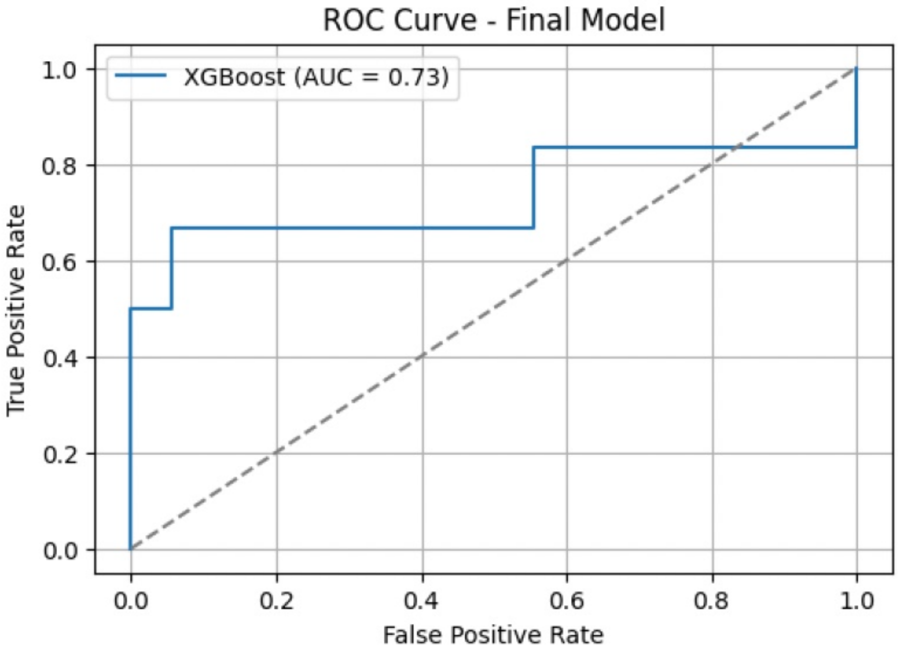

Figure 1. ROC Curve for final XGBoost model.

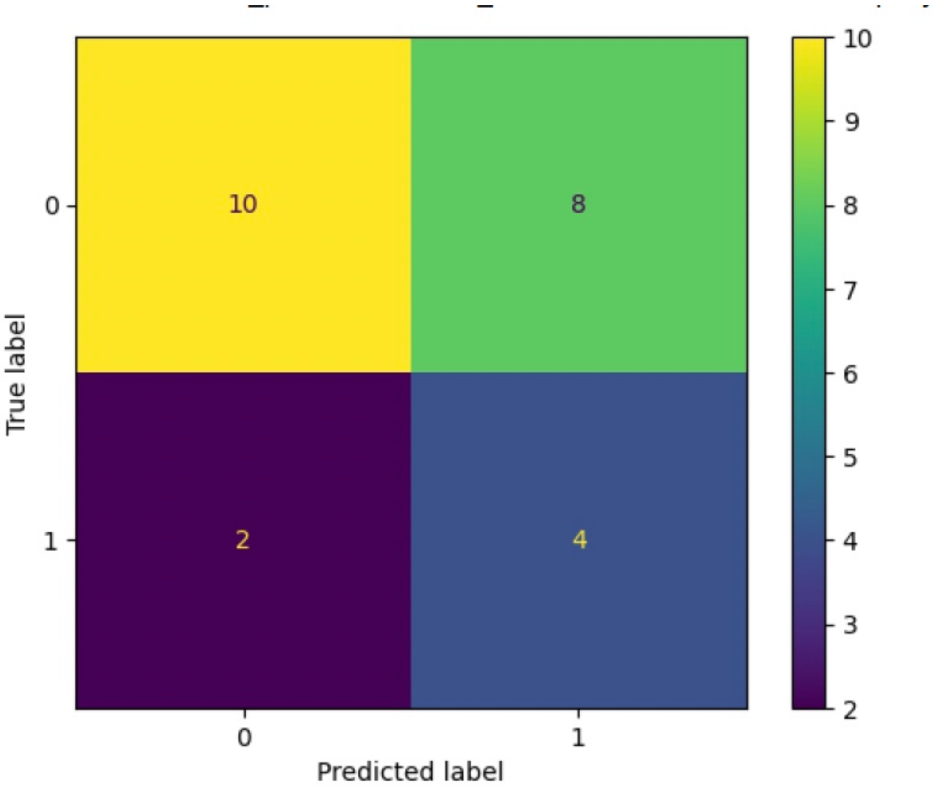

Figure 2. Confusion Matrix of final XGBoost model.

Feature importance analysis provided additional insights into which chemical properties contributed most strongly to the model's predictions. Descriptors such as topological polar surface area (TPSA), fr_ketone, and PEOE_VSA2 emerged as key predictors of DIA risk. These features align with established structure-activity relationships in immunotoxicity, indicating that molecular size, polarity, and electronic properties may affect a compound's potential to induce autoimmunity.

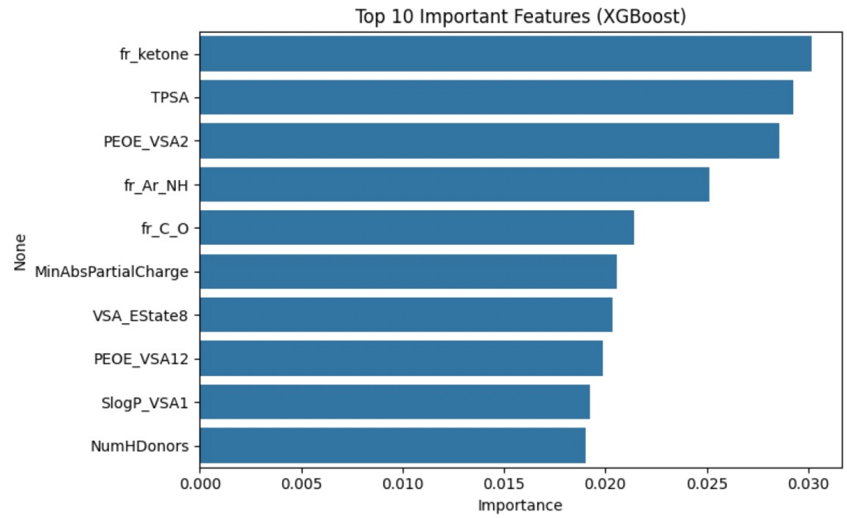

Figure 3. Feature Importance of final XGBoost model.

To assess the model's potential utility for prospective screening, I deployed it on the unlabeled set of compounds. The resulting distribution of predicted probabilities showed that most compounds fell within a moderate risk range (approximately 0.4 to 0.55), while a smaller subset exceeded 0.65, indicating potential high-risk candidates. The results suggest that the model helps prioritize compounds for further investigation during the early stages of drug development.

Overall, the pipeline demonstrated that machine learning models trained on chemical descriptors can provide valuable predictive insights into DIA risk and could serve as a useful complement to traditional screening methods.

The best XGBoost model achieved the following results on the held-out test set:

Accuracy: 0.58

AUC: 0.73

Discussion

XGBoost was used to develop the machine learning model that predicts the risk of drug-induced autoimmunity from molecular descriptors generated using RDKit. The goal was to explore whether chemical structure alone could offer predictive insights into DIA risk, supporting early screening efforts during drug development.

With an of AUC 0.73 and accuracy of 58% , the model shows promise. The model shows good ability to discriminate between DIA-positive and DIA-negative compounds, even though the accuracy was moderate. It also identified molecular patterns linked to DIA risk.

The classification report revealed that the model achieved higher precision on DIA-negative compounds, but also succeeded in identifying 67% of DIA-positive compounds, which are typically underrepresented in such datasets. This is an encouraging result given the known challenges of modeling rare toxicological outcomes in imbalanced biomedical data.

Feature importance analysis provided useful insights into the chemical properties most associated with DIA predictions. Descriptors such as fr_ketone, TPSA, and PEOE_VSA2 emerged as top contributors to model performance. These findings suggest that molecular polarity, surface area, and specific functional group patterns may play a role in modulating immunogenic potential. This aligns with prior work suggesting that certain structural motifs and physicochemical properties can influence immune system interactions. Given the small number of DIA-positive compounds in the dataset, the model's ability to generalize—particularly in terms of precision for positive samples—was likely constrained, representing one of its key limitations.

Additionally, the chemical descriptors used do not capture all possible biological mechanisms that could contribute to DIA — factors such as metabolism, immune system variability, and patient-specific responses are not represented in the feature set. Future work can be carried out by incorporating additional data sources like in-vitro assay that can build comprehensive models and address the limitations.

Hyperparameter optimization using Optuna contributed meaningfully to model performance, with parameters such as learning rate and scale_pos_weight emerging as influential. This highlights the value of tuning in improving model robustness, especially in imbalanced settings.

Overall, this project demonstrates that machine learning models based on chemical descriptors can provide valuable predictive insights into DIA risk and may serve as a useful complement to experimental screening approaches. From the current study we can say that further refinement of the machine learning model is required, and a more extensive validation. However, these models show promise and have the potential to improve early stage detection of DIA and reduce the risks in late stage detection.

These results provide an important proof of concept that machine learning models can augment early-stage drug safety screening by flagging potential DIA-inducing compounds prior to costly experimental validation. However, the relatively modest recall for the positive class suggests that further work is needed to enhance model sensitivity, potentially through incorporating biological assay data, genetic risk factors, or immune-modulating properties.

Future research should focus on integrating multi-modal data sources (e.g., transcriptomics, immunogenicity assays) with chemical descriptors to improve predictive power. It also should leverage automated machine learning frameworks like DeepMol (Grisoni et al., 2024) to further streamline optimization and development of the model. Additionally, explainability techniques such as SHAP analysis could help elucidate which molecular features are most strongly associated with DIA risk, aiding both model trustworthiness and scientific understanding.

An additional are for future work for predictive modeling involves exploring the potential applicability of structure-based immunotoxicity predictors in food safety systems.

Predictive models can be adapted to assess the immunogenic potential of novel food additives, contaminants, and packaging migrants, thereby supporting proactive risk management in food production and regulatory oversight.

Due to the small sample size and class imbalance, the model exhibited some variability across runs, with AUC values ranging from 0.63 to 0.73 in different splits. This is expected in biomedical datasets of this type, and suggests that future work should aim to stabilize performance through larger datasets and more robust validation.

In conclusion, this work contributes to the growing body of research leveraging machine learning for predictive toxicology and underscores the potential of data-driven approaches in improving the safety profile of drug candidates.

Conclusion

This study demonstrates that machine learning models — specifically optimized XGBoost classifiers trained on RDKit molecular descriptors — can effectively predict the risk of drug-induced autoimmunity (DIA) based on chemical structure. By combining class balancing with SMOTE, hyperparameter tuning with Optuna, and interpretability through feature importance analysis, the model achieved promising predictive performance on a challenging DIA classification task. These results show that such models can provide a valuable complement to experimental toxicology workflows, supporting earlier and more efficient prioritization of compounds in the drug development pipeline.

Enhancing model sensitivity through the incorporation of bio-assay data, access to larger and more diverse chem libraries and the use of techniques like SHAP for an indepth insight into structure-activity relationships should be part of the scope of future work. As data-driven approaches continue to evolve, machine learning holds considerable promise for advancing predictive toxicology and improving patient safety in pharmaceutical development.

Code Availability Statement

The code used to implement the machine learning pipeline for this study is available at: https://github.com/Mbokil19/DIA-ML-prediction.

Data Availability Statement

The dataset used in this study was obtained from *InterDIA: Interpretable prediction of drug-induced autoimmunity through ensemble machine learning approaches* (Huang et al., 2025), published in *Toxicology*. The authors of the original study provided the dataset publicly as part of their supplementary materials.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2623–2631. ACM.

Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://doi.org/10.32614/CRAN.package.papaja

Barth, M. (2023). *tinylabels: Lightweight variable labels*. Retrieved from https://cran.r-project.org/package=tinylabels

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. ACM.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., … Yuan, J. (2025). *Xgboost: Extreme gradient boosting*. Retrieved from https://CRAN.R-project.org/package=xgboost

Grisoni, F. et al. (2024). DeepMol: An automated machine and deep learning framework for chemical property prediction. *Journal of Cheminformatics*, *16*, 37. https://doi.org/10.1186/s13321-024-00937-7

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from https://www.jstatsoft.org/v40/i03/

Guo, H., Zhang, P., Zhang, R., Hua, Y., Zhang, P., Cui, X., … Li, X. (2022). Modeling and insights into the structural characteristics of drug-induced autoimmune diseases. *Frontiers in Immunology*, *13*, 1015409. https://doi.org/10.3389/fimmu.2022.1015409

Hu, J., Song, Y., & Li, C. (2024). Machine learning-driven toxicity prediction using RDKit molecular descriptors: A review and practical guidelines. *Journal of Cheminformatics*, *16*, 45.

Huang, L., Liu, P., & Huang, X. (2025). InterDIA: Interpretable prediction of drug-induced autoimmunity through ensemble machine learning approaches. *Toxicology*, *511*, 154064.

Kruta, J., Carapito, R., Trendelenburg, M., Rizzi, M., Mollet, A., Capri, M., … Miho, E. (2024). Machine learning for precision diagnostics of autoimmunity using multi-omics and clinical data. *Scientific Reports*, *14*(1), 27848. https://doi.org/10.1038/s41598-024-76093-7

Kuhn, & Max. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Landrum, G. (2006). *RDKit: Open-source cheminformatics.*

Lopez, S., Wichmann, K., & Chen, M. (2020). Drug-induced autoimmunity: Mechanisms and clinical manifestations. *Frontiers in Immunology*, *11*, 1054.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, *30*, 4765–4774.

Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames.* Retrieved from https://CRAN.R-project.org/package=tibble

Müller, K., Wickham, H., James, D. A., & Falcon, S. (2024). *RSQLite: SQLite interface for r.* Retrieved from https://CRAN.R-project.org/package=RSQLite

Patel, V., Joshi, A., & Kuo, T. (2023). Regulatory readiness of machine learning models for pharmaceutical risk assessment: Current trends and future outlook. *Regulatory Toxicology and Pharmacology*, *137*, 105373. https://doi.org/10.1016/j.yrtph.2023.105373

R Core Team. (2024). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for r and s+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r.* New York: Springer. Retrieved from http://lmdvr.r-forge.r-project.org

Shehab, E. A., Nayel, H., & Taha, M. (2025). OPTUNA optimization for predicting chemical respiratory toxicity using machine learning models. *Journal of Computer-Aided Molecular Design*, *39*(1), 21–35. https://doi.org/10.1007/s10822-025-00597-1

Siriseriwan, W. (2024). *Smotefamily: A collection of oversampling techniques for class imbalance problem based on SMOTE.* Retrieved from https://CRAN.R-project.org/package=smotefamily

Stafford, I. S., Kellermann, M., & Mossotto, E. (2020). A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digital Medicine*, *3*(30). https://doi.org/10.1038/s41746-020-0229-3

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., & Bryant, S. H. (2020). PubChem's BioAssay database. *Nucleic Acids Research*, *48*(D1), D400–D412. https://doi.org/10.1093/nar/gkz981

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors).* Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2023b). *Stringr: Simple, consistent wrappers for common string operations.* Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation.* Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2023). *Purrr: Functional programming tools.* Retrieved from https://CRAN.R-project.org/package=purrr

Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data.* Retrieved from https://CRAN.R-project.org/package=readr

Wickham, H., Pedersen, T. L., & Seidel, D. (2023). *Scales: Scale functions for visualization.* Retrieved from https://CRAN.R-project.org/package=scales

Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data.* Retrieved from https://CRAN.R-project.org/package=tidyr

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.org/knitr/

Yao, K. et al. (2021). Machine learning for predicting risk of drug-induced autoimmune diseases using structural alerts and molecular descriptors. *International Journal of Environmental Research and Public Health*, *18*(13), 7139. https://doi.org/10.3390/ijerph18137139

Zhu, H. (2024). *kableExtra: Construct complex table with 'kable' and pipe syntax.* Retrieved from https://CRAN.R-project.org/package=kableExtra