Jurnal Mandiri IT
Vol. 14 No. 1, July (2025), pp. 180-190

ISSN 2301-8984 (Print), 2809-1884 (Online)
Published by Institute of Computer Science

Comparative study of machine learning algorithms for predicting drug induced autoimmunity using molecular descriptors

Yusuf Rio Delfiero[1], Ajeng Hidayati [2], Bagus Hendra Saputra[3]
[1,2,3] Informatika, Universitas Pertahanan Republik Indonesia, Bogor, Indonesia

ARTICLE INFO

Article history:

Received Jun 30, 2025
Revised Jul 16, 2025
Accepted Jul 20, 2025

Keywords:

Drug Induced Autoimmunity;
Machine Learning;
Model Comparison;
Molecular Descriptors;
SHAP Analysis.

ABSTRACT

Drug induced autoimmunity (DIA) poses significant challenges in pharmaceutical development due to its complex immunological mechanisms and delayed clinical manifestations. This study proposes a comparative evaluation of three ensemble machine learning models CatBoost, XGBoost, and Gradient Boosting for predicting DIA using molecular descriptors. A curated dataset of drug compounds with known autoimmune outcomes was analyzed through a systematic workflow incorporating preprocessing, stratified sampling, and model evaluation using accuracy, F1 score, and ROC AUC. Results indicate that CatBoost achieved the highest ROC AUC, while XGBoost demonstrated superior balance between precision and recall, as reflected by its F1 score. Feature importance analysis using SHAP highlighted key molecular properties such as SlogP_VSA10 and fr_NH2 as major contributors to prediction outcomes. The study provides a reproducible and interpretable framework for early toxicity screening, offering valuable insights for data driven decision making in drug safety assessment.

Corresponding Author:

Yusuf Rio Delfiero,
Informatika,
Universitas Pertahanan Republik Indonesia,
IPSC Sentul Area, Sukahati, Citeureup District, Bogor Regency, West Java 16810, Indonesia
Email: riodelpiero15@gmail.com

1. INTRODUCTION

Ensuring drug safety remains a fundamental priority in pharmaceutical research and development, particularly during the early stages of compound screening (Husnain et al., 2023; Kabir et al., 2024). Among the various adverse drug reactions, drug induced autoimmunity (DIA) is a serious and complex immunological event that can lead to significant clinical consequences and even post marketing drug withdrawals (Chand et al., 2023). Conventional approaches for assessing autoimmune risk, such as *in vitro* assays and animal models, are often time consuming, resource intensive, and may not accurately reflect human immunopathological responses (Javadzadeh et al., 2025). These challenges highlight the growing need for computational methods capable of providing early and reliable predictions of autoimmune potential. In this context, molecular descriptors, which quantitatively encode chemical and physicochemical properties of drug compounds, have gained increasing attention in chemoinformatics driven toxicology (PANDEY, 2024). When integrated with machine learning algorithms, these descriptors enable the development of predictive models that can uncover complex, non linear relationships between molecular structure and biological effects. Despite their promise, predictive studies focusing specifically on DIA remain scarce, and there is limited consensus on which ML algorithms perform best for this task, given the high dimensional and heterogeneous nature of chemical data. This situation calls for a systematic investigation into the comparative performance of different ML methods for predicting drug induced autoimmunity using molecular descriptors (Xu et al., 2020).

Although drug induced autoimmunity poses significant clinical and regulatory challenges, its prediction during the early stages of drug development remains a largely unresolved problem (Yang et al., 2024). The complexity of autoimmune mechanisms, which often involve delayed onset and patient specific immunological responses, makes experimental detection difficult and often retrospective (Pisetsky, 2023). While computational approaches using molecular descriptors offer a promising alternative, current predictive efforts are hindered by several methodological limitations (Niazi & Mariam, 2023). Furthermore, the high dimensionality and redundancy inherent in molecular descriptor data exacerbate the risk of overfitting and model instability if not addressed properly (Jiang et al., 2025). These challenges underscore the pressing need for a systematic and comparative study of multiple machine learning algorithms to identify robust, high performing models for predicting drug induced autoimmunity.

Previous studies from (Smith et al., 2023) showed that the study employed XGBoost and Random Forest machine learning models to predict the potential of drugs to induce autoimmunity using transcriptional data. Both models achieved high classification accuracy, with XGBoost reaching an AUC of up to 0.875 and Random Forest 0.794, highlighting that disruptions in cell cycle regulation and proliferation are key features associated with autoimmune risk. This study by (Zheng et al., 2021) used 14 machine learning and regression models to predict blood concentrations of tacrolimus in autoimmune disease patients, based on real world therapeutic drug monitoring data. Among the models, XGBoost showed the best performance (R^2 = 0.54, MAE = 0.25, RMSE = 0.33), and using SHAP analysis, the study identified nine key predictors—such as height, daily dose, hematocrit, and LDL cholesterol—that influenced tacrolimus concentration, offering a practical tool for optimizing clinical dosing decisions. Meanwhile, a study by (Wu et al., 2021) developed a machine learning model to predict the risk of drug induced autoimmune diseases based on reactive metabolite related structural alerts and daily drug dose. Using CatBoost as the main predictive model, they found that a specific structural alert a benzene ring with a nitrogen containing substituent combined with a high daily dose (≥100 mg), significantly increased the prediction performance (AUC = 0.70), reducing false positives and improving positive predictive value, thus providing a practical prescreening tool for drug safety in early development stages.

The main objective of this study is to develop and evaluate predictive models to identify the risk of drug induced autoimmunity based on molecular descriptors. By applying a comparative approach, this study systematically examines the performance of several machine learning algorithms including Gradient Boosting, XGBoost, and CatBoost on a curated dataset of drug compounds annotated for autoimmune outcomes. This research aims to determine which algorithms offer the most reliable prediction accuracy, robustness, and generalizability for these specific toxicity endpoints. In addition, this research also aims to create a standardized modeling framework that integrates feature preprocessing, model training, and performance evaluation using metrics such as accuracy, F1 score, and area under the ROC curve. The results from this study are expected to contribute to the development of more effective in silico screening tools for early stage drug safety assessment, especially in detecting autoimmune liabilities before clinical testing.

Despite the increasing use of machine learning in drug toxicity prediction, the modeling of drug induced autoimmunity remains underexplored (Bhasuran et al., 2025). Most studies focus on general toxicity or rely on single algorithms without comparing alternative methods. Challenges such as high dimensional molecular data and class imbalance are often overlooked (Jiang et al., 2025). Comprehensive frameworks that include rigorous validation, multiple algorithms, and reproducible pipelines are rare, leaving limited guidance for researchers. This study aims to fill that gap by systematically evaluating machine learning models specifically for predicting autoimmune toxicity.

This study presents a rigorous and comprehensive approach to predicting drug-induced autoimmunity by comparing multiple machine learning algorithms using molecular descriptors. Unlike previous research focused on general toxicity or single models, it benchmarks diverse classifiers within a standardized framework that includes preprocessing, and multi metric evaluation. This enhances reproducibility and provides a foundation for future improvements. By addressing a gap in computational toxicology, the study contributes to the development of in silico tools for early detection of autoimmune risks, supporting data driven decisions in both research and drug safety practices..

2. METHODOLOGY AND EVALUATION

Research Design

This study applied a structured experimental workflow to develop predictive models for drug induced autoimmunity (DIA) using molecular descriptors. The process began with minimal preprocessing, as the dataset was already clean and well formatted, requiring only label encoding for categorical variables. The data was then split into training and testing sets to ensure unbiased model evaluation. Three supervised machine learning algorithms Gradient Boosting, XGBoost, and CatBoost were trained using the molecular descriptors as input features and autoimmune outcomes as the target variable. Each model was trained and assessed based on key classification metrics including Accuracy, F1 Score, and ROC AUC. A 2×2 confusion matrix was used to summarize prediction outcomes, while SHAP analysis was conducted to interpret feature contributions. This systematic approach ensures model robustness, comparability, and interpretability, supporting the development of reliable in silico tools for early toxicity prediction.

Population and Sample

The study population consisted of drug compounds with known immunological outcomes, curated from publicly available toxicological databases and prior published datasets on autoimmune reactions. The sample comprised molecular data from these compounds, each represented by a set of numerical molecular descriptors generated using RDKit, a cheminformatics toolkit. A total of *n* compounds were included in the dataset, with a balanced representation of compounds classified as either "autoimmune inducing" or "non autoimmune." To ensure model generalizability and minimize bias, the dataset was partitioned into a training set and a testing set using stratified sampling. The training set was used for model development and validation, while the testing set was held out for final performance evaluation.

Data Collection

The dataset used in this study was sourced from UCI Irvice (Drug Induced Autoimmunity Prediction). This dataset contains RDKit generated molecular descriptors designed for predicting drug induced autoimmunity using ensemble machine learning. It includes numerical features representing drug properties and structures, split into training and testing sets. The class labels are defined as 1 for DIA-positive drugs and 0 for DIA-negative drugs. The data enables the development of predictive, interpretable models for autoimmune risk assessment, supporting advances in computational toxicology and drug safety.

Data Pre Processing

The dataset required minimal preprocessing. The SMILES column was excluded, and features (X) were obtained by dropping the Label and SMILES columns. The Label column was used as the target (y), and label distribution was checked to assess class balance

Split Data

Before model training, the dataset was split into 80% training data (477 records) and 20% test data (120 records). The training data was then used to train with 3 models.

Gradient Boosting

Gradient Boosting is a machine learning technique that builds predictive models by combining multiple weak learners, typically decision trees, in a sequential manner. Each new model is trained to correct the errors made by the previous ones by minimizing a loss function using gradient descent. This approach is effective for both classification and regression tasks, offering high accuracy and flexibility.
(Bahad & Saxena, 2020).

Gradient Boosting models the prediction function as a summation of weak models (usually shallow decision trees):

$$\boldsymbol{F(x)} = \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}} \boldsymbol{y_m h_m(x)} \quad (1)$$

$\boldsymbol{F(x)}$: final prediction function, $\boldsymbol{h_m(x)}$: m th weak model (usually tree), $\boldsymbol{\gamma_m}$: step size or learning rate $\boldsymbol{M}$: number of iterations/tree

XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting algorithms designed for speed and performance (Demir & Sahin, 2023). It builds an ensemble of decision trees sequentially, where each new tree attempts to correct the residual errors made by the previous ones. XGBoost incorporates regularization (both L1 and L2) to prevent overfitting, handles missing data internally, and uses a second order Taylor approximation of the loss function for more accurate optimization. One of its key strengths lies in its ability to scale efficiently to large datasets while maintaining high predictive accuracy. XGBoost is widely used in machine learning competitions and practical applications due to its robustness and versatility (Johnson, 2025).
For a given iteration $\boldsymbol{t}$, the objective function to minimize is:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right)\right] + \Omega(f_t) \quad (2)$$

$\boldsymbol{l}$: the loss fuction, $\hat{\boldsymbol{y}}_{\boldsymbol{i}}^{(\boldsymbol{t-1})}$: the prediction from previous iteration, $\boldsymbol{f_t}$: the new tree added at iteration $\boldsymbol{t}$, $\boldsymbol{\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2}$ is the regularization term, (with $\boldsymbol{T}$ = number of leaves, $\boldsymbol{w_j} = \boldsymbol{leaf\ weight}, \boldsymbol{\gamma}, \lambda$ = regularization parameters)

CatBoost

The CatBoost model is a powerful gradient boosting algorithm developed by Yandex, designed to handle both numerical and categorical data efficiently (Sayyad et al., 2024). It builds an ensemble of decision trees in a sequential manner, where each tree is trained to minimize the errors of the previous trees using gradient descent optimization. CatBoost stands out from other boosting algorithms by employing ordered boosting and advanced techniques to reduce overfitting and handle categorical variables without extensive preprocessing (Hancock & Khoshgoftaar, 2020). This makes it particularly suitable for high dimensional datasets like those involving molecular descriptors. The prediction of the CatBoost model can be represented by the following formula:

$$\hat{y} = \sum_{m=1}^{M} \eta_m . T_m(x) \quad (3)$$

$\hat{\boldsymbol{y}}$: final prediction value, $\boldsymbol{M}$: the number of boosting iterations (trees), $\boldsymbol{\eta_m}$: the learning rate or weight for the $\boldsymbol{m}$-th tree, $\boldsymbol{T_m(x)}$: the output of the $\boldsymbol{m}$-th decision tree for input $\boldsymbol{x}$.

Evaluation

Metrics

After training the data with three classification algorithms, the models were evaluated using Accuracy, F1 Score, and ROC AUC to assess their performance. Accuracy measures the proportion of correct predictions, F1 Score balances precision and recall which is useful for imbalanced datasets, and ROC AUC reflects the model's ability to distinguish between classes. These metrics collectively provide a comprehensive view of how well each model predicts drug induced autoimmunity. The accuracy, precision, recall, and f1 score, and ROC AUCE values can be calculated with the following equations (Maulana et al., 2024) (8) - (11).

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (5)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (6)$$

$$\text{F1 Score} = \frac{2*Precision*Recall}{Precision+Recall} \quad (7)$$

$$\text{ROC AUC Score} = \int_0^1 TPR(FPR)dFPR \quad (8)$$

$$TPR = \frac{TP}{TP+FN} \quad (9)$$

$$FPR = \frac{TP}{FP+TN} \quad (10)$$

Confusion Matrix

The next stage is to evaluate the model's predictions using the confusion matrix, which is a table that summarizes the relationship between actual and predicted class labels. It has a square

shape of size L×L, where L denotes the total number of classes in the classification task (Luque et al., 2021). Since this study focuses on a binary classification problem, a 2×2 confusion matrix is used to represent the outcomes (Larner, 2024). The structure and components of this matrix are shown in Figure 2.

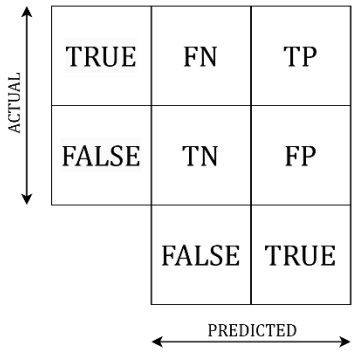

Figure 1 Confusion Matrix 2x2

SHAP

SHAP (SHapley Additive exPlanations) is a method for interpreting machine learning models by quantifying how much each feature contributes to a model's prediction output (Ekanayake et al., 2022). SHAP is founded on Shapley value theory from cooperative game theory, which ensures a fair allocation of each feature's contribution by evaluating all possible feature combinations (Liu et al., 2024; Veeramsetty, 2021).

The SHAP value for a feature *i* is defined as:

$$\emptyset_i = \mathbf{1} \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!.(|N| - |S| - \mathbf{1})!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (11)$$

$\emptyset_i$ is the SHAP value for feature iii, representing its contribution to the model prediction.
N is the set of all features.
S is a subset of features not containing iii.
$f(S)$ is the model prediction using only the features in subset SSS.
The expression $f(S \cup \{i\}) - f(S)$ measures the marginal contribution of feature iii when added to subset *S*.
The term $\frac{|S|!.(|N|-|S|-\mathbf{1})!}{|N|!}$ is a weighting factor that ensures fairness by averaging over all possible orderings of features.

3. RESULTS AND DISCUSSIONS

The training process, performed on 477 records using three different algorithms and evaluated on a test set of 120 records, produced accuracy results as shown in Figure 1 below.

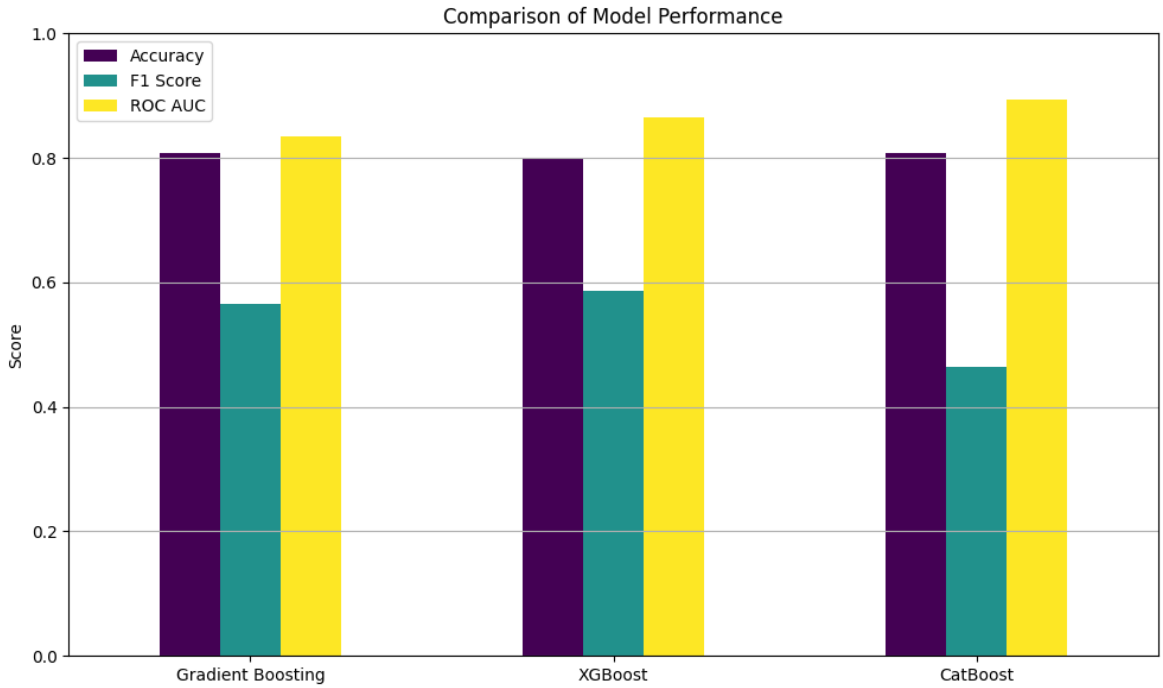

Figure 2 Metric Value Comparison of Each Model

Figure 2 shows that CatBoost gave the highest ROC AUC score, which means it was the best at distinguishing between drugs that could and couldn't trigger an autoimmune reaction. Meanwhile, XGBoost had the highest F1 score, showing that it was better at balancing between detecting positive cases and avoiding false alarms something important for rare reactions like DIA. Although all models had similar accuracy, the differences in F1 and AUC scores show why it's important to use more than just accuracy to judge a model's quality.

These findings highlight the value of testing several models instead of relying on just one. Each model has its strengths XGBoost works well when the data is imbalanced, and CatBoost is great at separating classes clearly. By comparing them in the same setup, we can see which one is most reliable for early stage drug screening. This is especially useful in drug development, where early prediction of side effects like autoimmunity can help avoid serious problems later in clinical trials or after the drug is released.

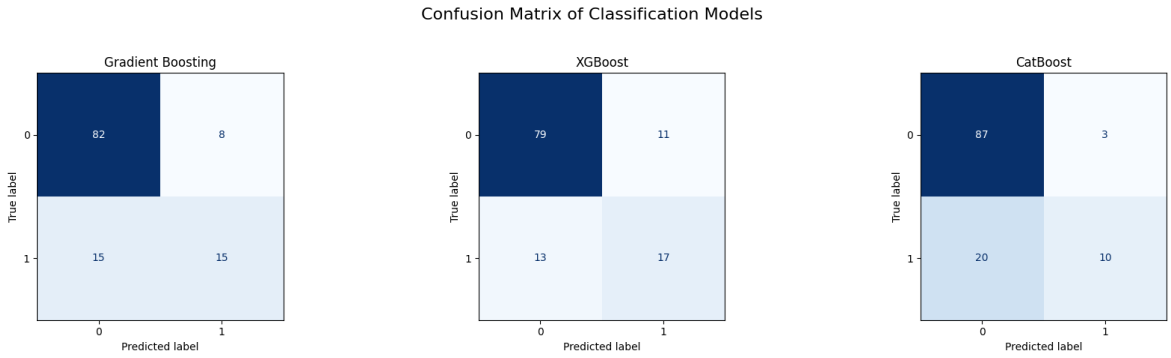

Figure 3 Metric Value Comparison of Each Model

The confusion matrices in Figure 3 illustrate the classification performance of the Gradient Boosting, XGBoost, and CatBoost models in predicting drug induced autoimmunity. Gradient Boosting shows a balanced performance with 15 true positives and 15 false negatives, indicating moderate sensitivity. XGBoost improves slightly, correctly identifying 17 positive cases with fewer false negatives (13), reflecting better recall. CatBoost, while achieving the highest number of true negatives (87), misclassifies more actual positive cases (20 false negatives), suggesting a tendency

toward conservative predictions. These differences highlight the trade offs each model makes between sensitivity and specificity, with XGBoost offering a better balance for detecting rare positive cases.

Additionally, to identify which features contribute the most to predicting grade values, we can use visual tools such as SHAP plots.

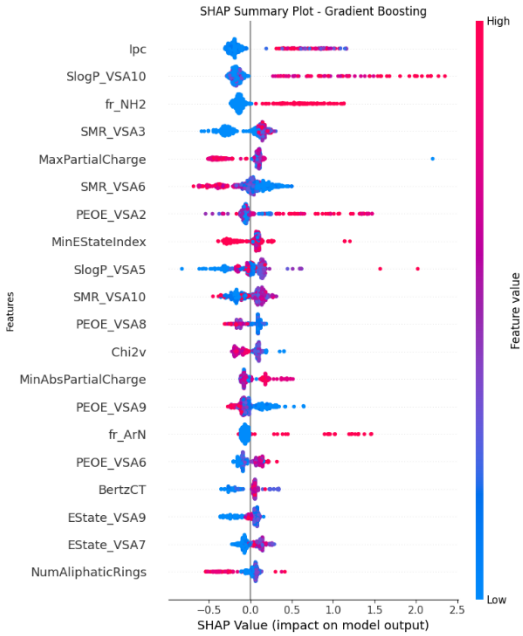

Figure 4 SHAP Summary Plot for Gradient Boosting Model

The SHAP summary plot for the Gradient Boosting model (Figure 4) reveals which molecular descriptors most influence the prediction of drug induced autoimmunity. The most impactful features include *lpc*, *SlogP_VSA10*, *fr_NH2*, and *SMR_VSA3*, all showing wide SHAP value ranges that indicate strong contributions to the model's output. Positive SHAP values suggest a feature increases the likelihood of autoimmunity, while negative values reduce it. The color gradient represents the actual value of each feature red for high and blue for low highlighting, for instance, that high values of *SlogP_VSA10* are associated with a higher autoimmune risk. Other descriptors such as *MaxPartialCharge*, *PEOE_VSA2*, and *MinEStateIndex* also contribute to predictions but to a lesser extent. Features lower in the plot like *NumAliphaticRings* and *EState_VSA7* show minimal impact. Overall, the plot offers a clear visual explanation of feature importance and directionality,

providing valuable insights into how specific molecular characteristics influence the model's decision making in predicting autoimmune outcomes.

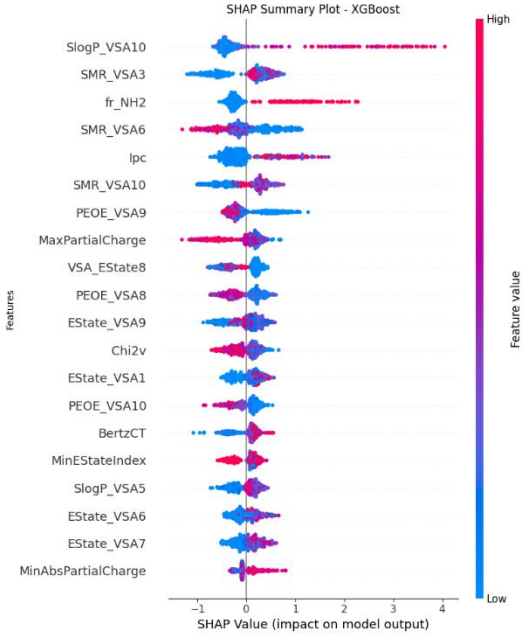

Figure 5 SHAP Summary Plot for XGBoost Model

Figure 5 showed that several molecular descriptors particularly SlogP_VSA10, SMR_VSA3, fr_NH2, SMR_VSA6, and lpc have the highest impact on the model's output in predicting drug induced autoimmunity. High values of SlogP_VSA10 and fr_NH2 increase the predicted risk, indicating that surface area associated with lipophilicity and the presence of primary amine groups are strongly associated with autoimmune potential. Descriptors such as SMR_VSA3 and SMR_VSA6 also contribute significantly, although their effects vary depending on feature values. In contrast, features like MinAbsPartialCharge and EState_VSA7 contribute minimally, as reflected by SHAP values clustered around zero. These results highlight the key molecular features that drive the model's predictions and provide evidence based insights into structural characteristics associated with increased autoimmune risk.

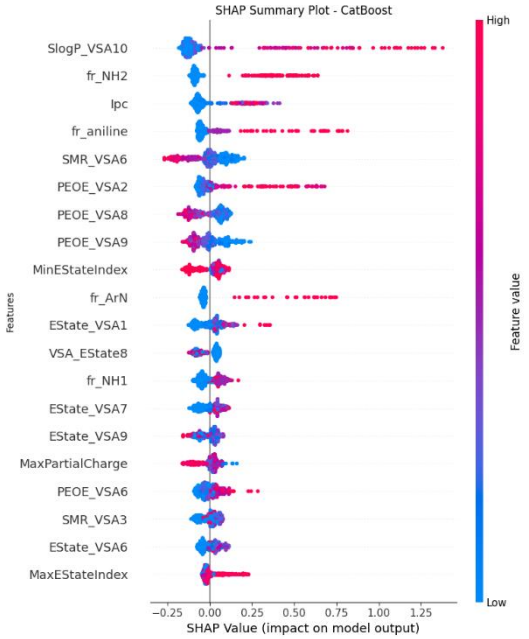

Figure 6 SHAP Summary Plot for CatBoost Model

Figure 6 highlights SlogP_VSA10, fr_NH2, and lpc as the most influential features in predicting drug induced autoimmunity, with high values of these descriptors increasing the predicted risk. Other descriptors such as fr_aniline, SMR_VSA6, and PEOE related features also contribute notably, though with more variable effects. In contrast, features like MaxEStateIndex and EState_VSA6 show minimal impact, with SHAP values clustered near zero. Overall, the plot underscores the importance of lipophilicity, electronic properties, and specific functional groups in driving the model's predictions.

Discusion

In the context of defense and strategic infrastructure such as military bases, Internet of Things (IoT)-based security systems play a vital role in detecting and responding to potential threats in real time. However, existing systems face several significant obstacles that limit their effectiveness in large-scale operational scenarios. First, most IoT security systems rely on short-range communication protocols such as WiFi and Bluetooth, which are only effective within limited ranges and are not suitable for large and remote military environments. Second, delays or high latency in motion detection and notification delivery hinder rapid decision-making, which is critical in emergency situations or attacks. Third, high energy consumption due to the continuous operation of sensors and cameras causes devices to run out of power quickly, especially in locations that are difficult to reach by conventional power supplies. These three challenges indicate a critical gap in the design of currently used IoT security systems, especially when faced with operational demands in strategic environments that require wide coverage, fast response, and high energy efficiency. Therefore, an innovative approach is needed that can holistically address these three issues to improve the reliability and sustainability of IoT-based security systems.

This research aims to address the main challenges in IoT security systems through the development of a solution that integrates PIR motion sensor technology, ESP-32 WROVER camera modules, and LoRa (Long Range) wireless communication. LoRa was chosen for its ability to transmit data with low power consumption and long communication range, up to 20–30 km, making it highly suitable for implementation in large-scale military environments. The system is designed to minimize latency by processing data locally on the device, thereby avoiding reliance on cloud-based processing, which often causes delays. Additionally, the use of the ESP-32 WROVER camera module enables direct image capture upon detecting movement, which is then automatically sent via Telegram notifications. This combination allows users to receive instant visual confirmation of any suspicious events. Thus, the system functions not only as a detection tool but also as a visual threat verification system. This research also focuses on energy efficiency by implementing a low-power standby mode approach that only activates sensors and cameras when needed, thereby reducing energy consumption and extending the operational lifespan of devices in the field.

By designing an IoT security system that integrates motion detection, image capture, long-range communication, and real-time notifications, this research makes a significant contribution to the advancement of IoT-based security technology. The proposed system demonstrates a significant performance improvement compared to conventional solutions, particularly in terms of response time, which ranges from 1.1 to 1.6 seconds—faster than cloud-based systems with higher latency. The energy efficiency achieved through the use of LoRa makes this system more suitable for deployment in environments without stable power access, such as remote military bases or other critical surveillance areas. This innovation also offers scalability flexibility, enabling the system to be expanded to multiple surveillance points without requiring complex network infrastructure. By combining reliable communication, fast response, and energy savings, this system provides a cost-effective yet robust solution to address security challenges in the digital age. Overall, the approach proposed in this research not only addresses existing gaps in IoT security systems but also opens new directions for developing adaptive, responsive, and energy-efficient systems for national security needs and the protection of strategic infrastructure.

4. CONCLUSION

This study demonstrates the effectiveness of ensemble learning models CatBoost, XGBoost, and Gradient Boosting in predicting drug induced autoimmunity using molecular descriptors. While all models showed comparable accuracy, CatBoost achieved the highest ROC AUC, and XGBoost had the best F1 score, emphasizing the importance of evaluating multiple metrics beyond accuracy alone. SHAP analysis further revealed that descriptors related to lipophilicity, electronic distribution, and

functional groups, such as SlogP_VSA10, fr_NH2, and lpc, consistently played dominant roles in model predictions. These findings suggest that combining model performance evaluation with feature interpretability provides a robust framework for early stage drug screening, enabling more informed decisions to minimize autoimmune risks in drug development. For future research, expanding the dataset with more diverse compounds, integrating deep learning based molecular embeddings, and validating the models on external datasets or in vitro studies could further enhance the reliability and generalizability of predictive frameworks in pharmacovigilance.

REFERENCES

Bahad, P., & Saxena, P. (2020). Study of adaboost and gradient boosting algorithms for predictive analytics. *International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019*, 235–244.

Bhasuran, B., Jin, Q., Xie, Y., Yang, C., Hanna, K., Costa, J., Shavor, C., Han, W., Lu, Z., & He, Z. (2025). Preliminary analysis of the impact of lab results on large language model generated differential diagnoses. *Npj Digital Medicine*, *8*(1), 166.

Chand, S., Dikkatwar, M. S., Pant, R. D., Misra, V., Pradhan, N., Ansari, U., & Debnath, G. (2023). Drug Induced Hematological Disorders: An Undiscussed Stigma. In *Drug Metabolism and Pharmacokinetics*. IntechOpen.

Demir, S., & Sahin, E. K. (2023). An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost. *Neural Computing and Applications*, *35*(4), 3173–3190.

Ekanayake, I. U., Meddage, D. P. P., & Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials*, *16*, e01059.

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *Journal of Big Data*, *7*(1), 94.

Husnain, A., Rasool, S., Saeed, A., & Hussain, H. K. (2023). Revolutionizing pharmaceutical research: Harnessing machine learning for a paradigm shift in drug discovery. *International Journal of Multidisciplinary Sciences and Arts*, *2*(4), 149–157.

Javadzadeh, S. M., Barati, M., Ghahremani, A., & Ahmadabad, H. N. (2025). Challenges and Strategies for Developing Effective Neoantigen-Based Dendritic Cell Vaccines for Cancer Immunotherapy: A Literature Review. *Reviews in Clinical Medicine*, *12*(2).

Jiang, J., Zhang, C., Ke, L., Hayes, N., Zhu, Y., Qiu, H., Zhang, B., Zhou, T., & Wei, G.-W. (2025). A review of machine learning methods for imbalanced data challenges in chemistry. *Chemical Science*.

Johnson, R. (2025). *XGBoost in Practice: Definitive Reference for Developers and Engineers*. HiTeX Press.

Kabir, M., Rana, M. R. H., & Debnath, A. (2024). The Role of Quality Assurance in Accelerating Pharmaceutical Research and Development: Strategies for Ensuring Regulatory Compliance and Product Integrity. *Journal of Angiotherapy*, *8*(12), 1–11.

Larner, A. J. (2024). *The 2x2 matrix: contingency, confusion and the metrics of binary classification*. Springer Nature.

Liu, Y., Fu, Y., Peng, Y., & Ming, J. (2024). Clinical decision support tool for breast cancer recurrence prediction using SHAP value in cooperative game theory. *Heliyon*, *10*(2).

Luque, A., Mazzoleni, M., Carrasco, A., & Ferramosca, A. (2021). Visualizing classification results: Confusion star and confusion gear. *IEEE Access*, *10*, 1659–1677.

Maulana, R., Narasati, R., Herdiana, R., Hamonangan, R., & Anwar, S. (2024). Komparasi Algoritma Decision Tree Dan Naive Bayes Dalam Klasifikasi Penyakit Diabetes. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(6), 3865–3870. https://doi.org/10.36040/jati.v7i6.8265

Niazi, S. K., & Mariam, Z. (2023). Recent advances in machine-learning-based chemoinformatics: a comprehensive review. *International Journal of Molecular Sciences*, *24*(14), 11488.

PANDEY, R. (2024). CHEMOINFORMATICS: A COMPUTATIONAL GATEWAY. *Omics Applications and Avenues*, 241.

Pisetsky, D. S. (2023). Pathogenesis of autoimmune disease. *Nature Reviews Nephrology*, *19*(8), 509–524.

Sayyad, J., Attarde, K., & Saadouli, N. (2024). Optimizing e-commerce Supply Chains with Categorical Boosting: A Predictive Modeling Framework. *IEEE Access*.

Smith, G. L., Walker, I. G., Aubareda, A., & Chapman, M. A. (2023). A machine learning approach to predict drug-induced autoimmunity using transcriptional data. *BioRxiv*, 2023.04.04.533417. https://www.biorxiv.org/content/10.1101/2023.04.04.533417v1%0Ahttps://www.biorxiv.org/content/10.1101/2023.04.04.533417v1.abstract

Veeramsetty, V. (2021). Shapley value cooperative game theory-based locational marginal price computation for loss and emission reduction. *Protection and Control of Modern Power Systems*, *6*(4), 1–11.

Wu, Y., Zhu, J., Fu, P., Tong, W., Hong, H., & Chen, M. (2021). Machine learning for predicting risk of drug-induced autoimmune diseases by structural alerts and daily dose. *International Journal of*

Environmental Research and Public Health, *18*(13). https://doi.org/10.3390/ijerph18137139

Xu, L. L., Young, A., Zhou, A., & Röst, H. L. (2020). Machine learning in mass spectrometric analysis of DIA data. *Proteomics*, *20*(21–22), 1900352.

Yang, Y., Liu, Y., Chen, Y., Luo, D., Xu, K., & Zhang, L. (2024). Artificial intelligence for predicting treatment responses in autoimmune rheumatic diseases: advancements, challenges, and future perspectives. *Frontiers in Immunology*, *15*, 1477130.

Zheng, P., Yu, Z., Li, L., Liu, S., Lou, Y., Hao, X., Yu, P., Lei, M., Qi, Q., Wang, Z., Gao, F., Zhang, Y., & Li, Y. (2021). Predicting Blood Concentration of Tacrolimus in Patients With Autoimmune Diseases Using Machine Learning Techniques Based on Real-World Evidence. *Frontiers in Pharmacology*, *12*(September), 1–10. https://doi.org/10.3389/fphar.2021.727245