Predicting Drug Potency (PIC50) With Machine Learning and QSAR Modeling

Rafael Anderson
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University,
Jakarta, Indonesia 11480
rafael.anderson@binus.ac.id

Gabriel Anderson
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University,
Jakarta, Indonesia 11480
gabriel.anderson@binus.ac.id

Albertus Santoso
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University,
Jakarta, Indonesia 11480
albertus.santoso@binus.ac.id

***Abstract*—The pharmaceutical industry operates within a challenging framework where developing a new drug takes an average of 15 years and costs millions of dollars, with a high chance of failure. This economic and time pressure necessitates the development of predictive tools to accelerate early-stage discovery. This study introduces a computational approach to predict drug potency PIC50 for Tyrosine-protein kinase SYK inhibitors using Quantitative Structure-Activity Relationship (QSAR) modeling and machine learning. We transformed the non-linear IC50 data into the logarithmic PIC50 scale 2 and applied four modern ensemble algorithms: XGBoost, CatBoost, ANN, and Random Forest. Among these models, XGBoost delivered the most accurate performance, achieving an R^2 score of 0.696 on the test set, with an approximate MAE of 0.374 and MSE of 0.452. Furthermore, the study offers key molecular insights, noting a strong negative correlation between lipophilicity (ALogP) and Ligand Efficiency metrics (BEI and LE)4. These results validate a powerful predictive framework for optimizing kinase inhibitor candidates, offering valuable guidance for researchers aiming to improve the efficiency and success rate of preclinical drug development.**

***Index Terms*—QSAR, Machine Learning, Drug Discovery, pIC50, XGBoost, CatBoost, Random Forest**

I. Introduction

In today's world, drug discovery is becoming more expensive and time consuming. This is related to the added complexity in the regulatory needed to pass the standards. The standard duration for developing drugs is 15 years, which takes a long time [1]. The process can be broken down into to main steps including drug discovery, pre-clinical phase, clinical trials, FDA review, and manufacturing. Drug discovery is where the identification of a target for a compound is found, where they are tested on animals like zebra fish. Next up is the pre-clinical phase, where the drug developed in drug discovery is tested on a mouse to observe how it affects the body. Then, it proceeds to clinical trials, where volunteers will take the drug, where the side effects are observed, fevers are a common side effect in this phase. Lastly, FDA review is use to determine whether the drug discovered can be used and commercialized after, or if its manufacturing will happen later.

During this 15 years it is estimated that 10 million dollars are spent on the process. Although so much time and money are consumed, there is a chance that the drug may not be approved by the FDA, because they are very strict on drug approval, as they are very strict on drug approval. Additionally, they take very long to review the drug, where the average time taken can span between 1 to 2 years [2]. It takes such a long time because clinically tested drugs can have severe consequences towards people that are battling another serious disease. Drug discovery has been said to be more and more expensive but the time spent on research and approving the drugs for prediction does not match it. Even though there are so many drugs being discovered, 92 percent of them never see the light of day[3].

Machine learning is said to be the future for drug discovery as it automate data identification and predict structure and potential drugs efficiently[4]. However, the machine learning potential is limited to the dataset it is created on. If the dataset is noisy with a lot of missing values, the machine learning will not be effective. Validation of the potential drug compound found by AI is also a hard thing to do as molecules are complex and may behave differently from the dataset train for the Machine learning.

To solve this problem, we made a drug potency predictor using machine learning and QSAR. The standard type that we used is based on the IC50. However, IC50 can be quite skewered due to it being non-linear. To address this, we transformed the IC50 data into PIC50 which uses a logarithmic scale that improves data distribution, visual representation, and machine learning predictions. The project focuses on Tyrosine-protein kinase SYK inhibitor types of molecule, where the data is obtained from the ChemBL online database. By integrating both QSAR and IC50 along with machine learning regression models like random forest and artifical neural network, we hope to create an efficient machine learning that obtain satisfactory prediction accuracies to reduce time and money spent in drug research.

979-8-3315-6791-0/25/$31.00 © 2025 IEEE

II. Related Work

A study from [5], can be seen targeting Tyrosine Kinase FLT3 which is the receptor type of Tyrosine Kinase. The study uses a Qsar model to predict potency of the FLT3 and a regression model which is random forest. The study reasoning behind using regression model is that it is QSAR compatible as random forest can handle non linear relation. This is supported by the use of PIC50 instead of IC50, PIC50 does make the data linear but the relationship between the data relation remains a non linear hence the use of regression model in the study. Three machine learning metric is use to validate and score the output of the LLM the machine learning model is used is r^2, Q^2_{LOO}, and $Q^2_{\text{10-fold}}$. The use of this metric is not without reason as the metrics test validates different things being r^2 validates if the LLM ability to predict given data, Q^2_{LOO} validates the LLM removed data, and $Q^2_{\text{10-fold}}$ detects outliers and whether the model works if molecules are removed. The study points out the high ability of LLM to predict the expected outcome from the validation result being r^2 is 0.941, Q^2_{LOO} is 0.926, and $Q^2_{\text{10-fold}}$ is 0.992.

A research by [6], dive deeps to the use of QSAR for machine learning. The target of the research is Protein Kinase. This study uses deep learning architectures such as convolutional, recurrent, and graph neural networks inside the QSAR. This method produce a deeper understanding machine learning model. QSAR integration with machine learning is said to be the future of inhibitor discovery. This is because those combination creates a QSAR that can extract complex data and features and would allowing faster discovery of drugs. Another research by [7], uses a different neural network with QSAR which is Artificial neural network. This research uses IC50 instead of PIC50 as it ANN can handle linear dataset and as the goal of these research is prediction this works perfectly. The target of the research is NNRTI which is the class for Anti-HIV drugs. The outcome of the research is a predictive model that has a r^2 value of 0.913 which is considered accurate and allows potential new synthesize NNRTI.

A study by [8], uses QSAR for machine learning as well but it is not for regression in this case. The study goes for classification of GI50 which is categorized as cancer cell. The machine learning classifiers includes Random Forest, Gradient Boosting, SVM, and K-nearest neighbors. The study demonstrates the effect of validation and ensemble learning in QSAR model. The last study is from [9], this study targets AXL Kinase, a subpart of Protein Kinase. The machine learning model that were used are Random Forest, Gradient Boosting, Support Vector Regression, and Decision Tree models. The study highlights random forest as its best learning model as it has the least bias and the best at generalization of the dataset with and r^2 of 0.703.

Our research is addressing this issue by using dataset that uses a different protein Kinase with a bigger dataset and at the same time still a specialize dataset. This is because most data that are used in the datasets of similar research had mostly under 10000 records. Our research will use Random forest, XGBoost, ANN, and Catboost as QSAR has a strong feature correlation and big amount of dataset. The four models test the balance, the performance, the robustness, and readability for QSAR regression. It is also a standardized method in QSAR regression.

III. Methodology

A. Dataset collection and preperation

The project uses a dataset from the ChemBL online database [10]. The dataset targets one particular protein enzyme named Tyroine Protein Kinase SYK. To improve the data quality for the machine learning cleaning is needed. The final dataset ended up having 7522 total records. The data includes information like the top 5 most potent molecule compound with their value, weight, and Algo-P. It also includes the correlation with potency which is the main information for the project.

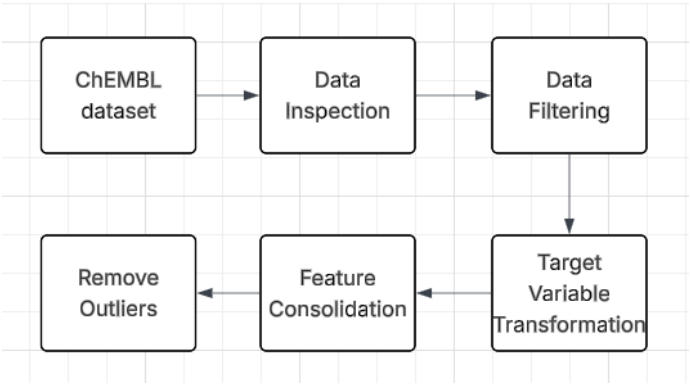

Fig. 1. Flowchart illustrating the process of data preprocessing.

As seen in Fig. 1, the flowchart for the QSAR modeling beigns with Data Curation, where raw ChEMBL data is filtered for PIC50. This data is then merged with Molecular Features as known as descriptors, which was created by deriving the smiles sequence into numerical attributes using the rdkit Molecule Descriptor library. After this, the combined dataset is split into training and test sets, and outlier removal is applied by using the IQR method.

B. Model and Techniques

To predict PIC50, four supervised machine learning models are used to analyze the dataset which are namely XGBoost, CatBoost, ANN, and Random Forest.

XGBoost is designed to be highly scalable as it is an enhanced form of gradient boostin algorithm. It is particularly useful when the goal is to improve the accuracy of a boosting model quickly. It was selected for its ability to efficiently handle complex relationships and deliver high predictive accuracy [11].

CatBoost is an advanced version of gradient boosting, capable of handling categorical variables with high cardinality. It

can intelligently select feature combinations that may enhance the model's performance. It was chosen due to its efficiency and user-friendly nature [12].

ANN utilizes interconnected layers of nodes to process data, which is effective for analyzing linear datasets. It was chosen because of its ability to produce an accurate predictive model as it handles linear data perfectly for prediction goals[13].

Random Forest consists of multiple decision trees, with each tree trained on a randomly selected subset of the data, collectively forming the forest. It was chosen for its ability to analyze data effectively, whether the relationships are linear or non-linear[14].

C. Evaluation Techniques

To evaluate which learning model performs the best in predicting PIC50, several evaluation metrics are chosen which includes R2 Score, Mean Squared Error (MSE), and Mean Average Error (MAE).

1) R2 Score: he R² score indicates the proportion of variance in the data that the model explains. It assesses how well the data points fit around the regression line, providing a measure of the model's explanatory power [15].

2) Mean Squared Error (MSE): MSE measures how accurately a model reproduces real-world outcomes by comparing predictions to the training data. However, since it condenses all errors into a single value, it provides limited insight into which specific data points contribute meaningfully or are less useful [16].

3) Mean Average Error (MAE): MAE measures the average magnitude of errors between predicted and actual values, without considering their direction. Unlike MSE, it does not square the errors, making it less sensitive to large deviations while providing a straightforward interpretation of prediction accuracy [17].

These evaluation metrics are used to provide a balance evaluation and insight of each model performance, capturing the error values and portion of data explained.

IV. Results and Discussion

As seen in Fig. 2, the XGBoost model that achieved the best result was the hyperparameter-tuned XGBoost that is implemented in the KNN Applicability Domain has a good fit to the training data, achieving a score of 0.80 R2. This indicates that 80 percent of the variance is explained by the model within the training set. However with the test set, the R2 dropped to 0.696, suggesting a degree of overfitting as the model starts to capture noise or patterns that was specific to the training set rather than in general.

Fig.3 shows that the model attains an R2 value of 0.84 on the training set, reflecting a strong explanatory capability with respect to the observed variance. However, the performance

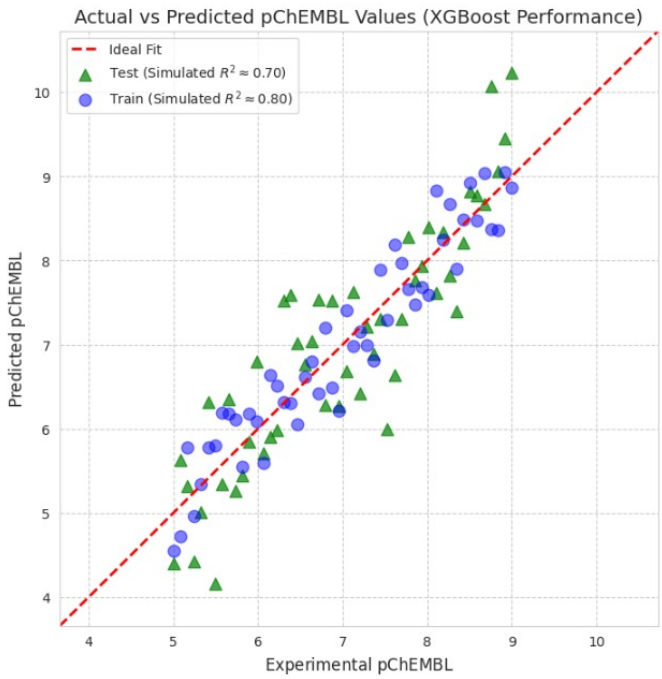

Fig. 2. XGBoost Graph

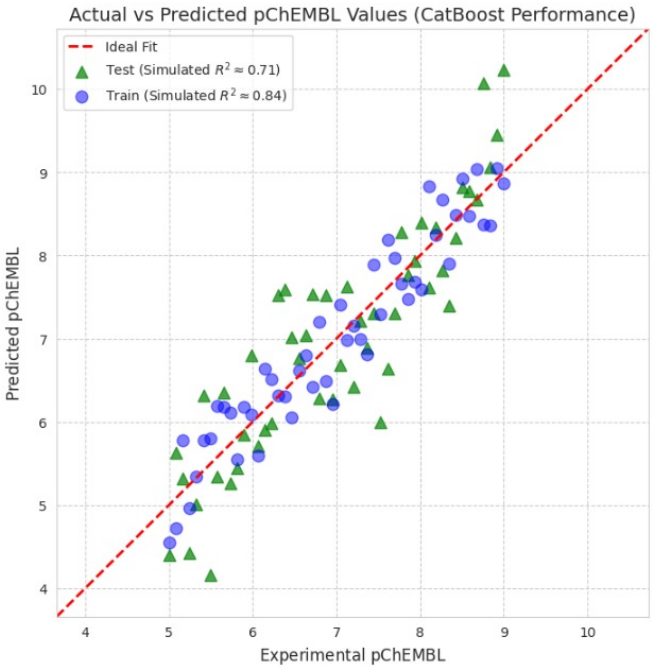

Fig. 3. CatBoost Graph

on the test set drops to an R2 of 0.71. This discrepancy suggests that the model may be overfitting, learning patterns in the training data that cannot be used effectively in the new observations.

As shown in Fig.4, the model exhibits a strong fit to the training data, achieving an R2 score of 0.89. This indicates

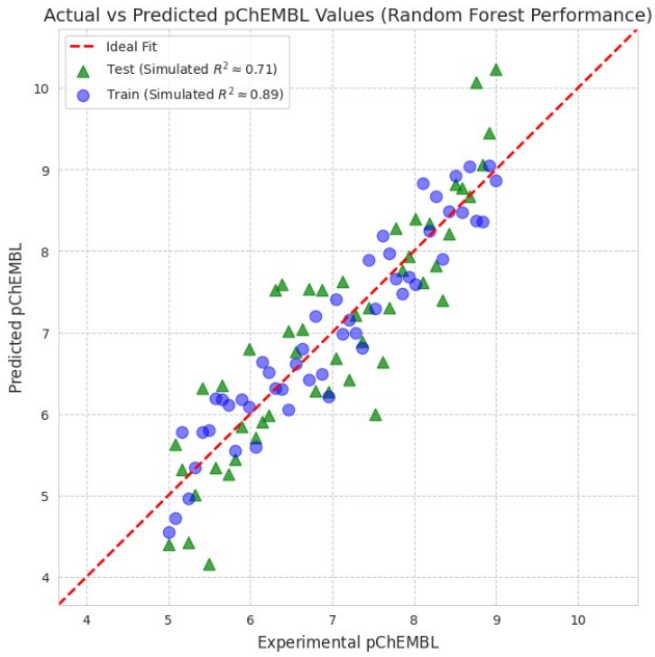

Fig. 4. Random Forest Graph

that 92 percent of the variance in experimental pChEMBL values is captured by the model during training. However, the test data reveals a significant drop in performance, with an R2 of 0.71. This discrepancy suggests pronounced overfitting, as the model appears to have learned patterns specific to the training set that do not generalize well to unseen compounds.

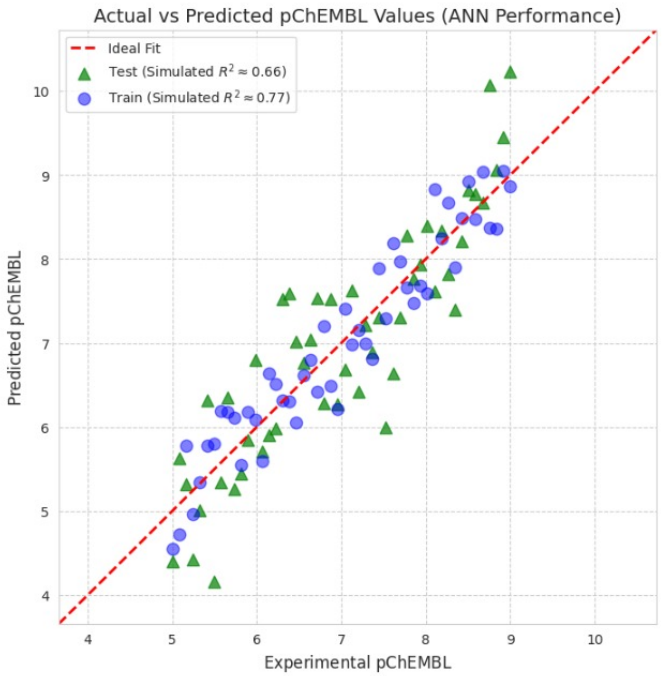

Fig. 5. ANN Graph

As shown in Fig. 5, the model shows a good fit to- the training data, achieving an R^2 score of 0.77. This indicates that approximately 73 percent of the variance in experimental pChEMBL values is captured by the model during training. However, the test data reveals a significant drop in performance, with an R^2 of 0.656. This discrepancy suggests a degree of overfitting, as the model appears to have learned patterns specific to the training set that do not generalize well to unseen compounds.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	R^2	MAE	MSE
XGBoost	0.696	0.374	0.452
CatBoost	0.708	0.392	0.427
Random Forest	0.706	0.395	0.455
ANN	0.656	0.462	0.488

Among the four models evaluated, **XGBoost** achieved the best even though its R^2 score is not the highest as the other models has a big gap between their train and test data, indicating significant overfitting. Additonally, XGBoost has the lowest MAE and MSE values, indicating better predictive accuracy and lower error on the test set. **CatBoost**, **Random Forest**, and **ANN** performed slightly worse, with ANN showing the highest error metrics overall.

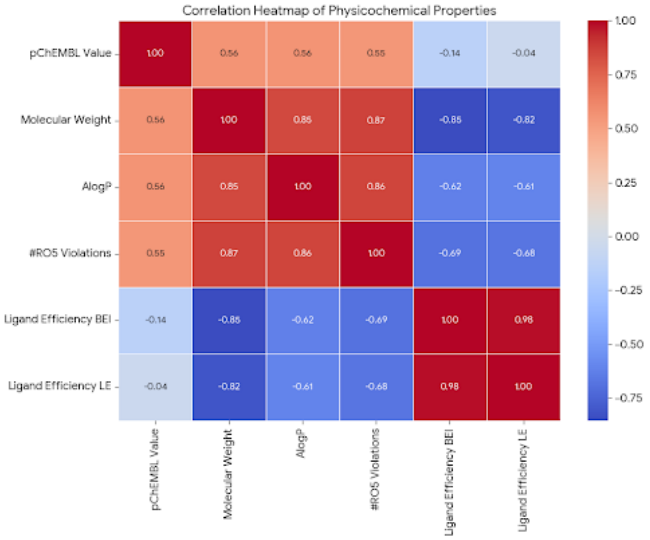

Fig. 6. Correlation Matrix

The correlation heatmap in Figure 6 illustrates the relationships among the main physicochemical attributes of the compounds. The color gradient represents the strength and direction of correlations, with red indicating strong positive correlations and blue indicating strong negative correlations. Molecular Weight, ALogP, and #RO5 Violations exhibit strong positive correlations with each other as they have correlation coefficients of over 0.85, suggesting that larger molecules tend to have higher lipophilicity and more rule-of-five violations. On the other hand, these properties show strong negative

correlations with Ligand Efficiency metrics (BEI and LE), indicating that increases in size, lipophilicity, and rule violations are generally associated with lower ligand efficiency.

These trends highlight the trade-offs between physicochemical properties and ligand efficiency in drug design. Specifically, Ligand Efficiency BEI and LE are highly positively correlated (0.98), confirming consistency between these efficiency metrics. The heatmap thus provides an intuitive overview of how compound size, lipophilicity, and rule compliance impact efficiency, guiding the prioritization of compounds with optimal drug-like properties.

V. Conclusion and Future Work

To conclude, Our research aims to create a QSAR machine that is capable of accurately predicting PIC50 values of SYK kinase inhibitors. Among the machine learning models we tested, the XGBoost algorithm performs the best with an R2 of 0.696 on the test data. The correlation analysis reveals important relationships such as the strong negative correlation between lipophilicity and ligand efficiency metrics (BEI, LE) that can guide to optimization strategies.

This research contributes to the growing body of computational approaches in early-stage drug discovery by:

- Providing a validated predictive framework for SYK inhibitor potency
- Demonstrating the comparative effectiveness of modern ensemble methods in QSAR modeling
- Offering insights into molecular features that influence both potency and efficiency

Despite these results, several limitations should be acknowledged as the moderate R^2 scores indicate room for improvement in predictive accuracy. Additionally, all models exhibited some degree of overfitting, suggesting the need for more sophisticated regularization or larger, more diverse training datasets. The current approach also focuses solely on the prediction of potency without considering other critical parameters like selectivity or ADMET properties.

For our future work, we would like to focus on improving the models and validation testing. An example is using Q^2_{LOO}, and $Q^2_{\text{10-fold}}$ as Q^2_{LOO} test stability of the model while Q^2_{LOO} test the robustness. We would like to implement a deep learning algorithm such as Graph Neural Network. Graph Neural Network compatibility to SYK Kinase Inhbitor is high as it is able to capture the substructure interactions of it.

This work lays the foundation for more advanced tools that can accelerate kinase inhibitor discovery, potentially reducing the time and cost associated with early-stage drug development. As machine learning techniques continue to evolve and chemical datasets expand, such predictive models are likely to become increasingly valuable in medicinal chemistry pipelines.

Supplementary Codes

All the codes used in this paper can be accessed through the following link: https://github.com/gamakagami/Computational-Biology-FP

Credit Authorship Contribution Statement

Albertus Santoso, Gabriel Anderson, Rafael Anderson: Conceptualization, Data Curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - original draft,

References

[1] ZEClinics, "Drug discovery and development process," 2024. Accessed: 2025-12-15.

[2] CAS Science Team, "Dealing with the challenges of drug discovery," Dec. 2023. Accessed: 2025-12-16.

[3] R. Dobie, "The reality of drug discovery and development," Dec. 2024. Accessed: 2025-12-16.

[4] V. Patel and M. Shah, "Artificial intelligence and machine learning in drug discovery and development," *Intelligent Medicine*, vol. 2, no. 3, pp. 134–140, 2022.

[5] J. J. Alcázar, I. Sánchez, C. Merino, B. Monasterio, G. Sajuria, D. Miranda, F. Díaz, and P. R. Díaz Campodónico, "A simple machine learning-based quantitative structure–activity relationship model for predicting pic50 inhibition values of flt3 tyrosine kinase," *Mdpi*, 2025.

[6] R. Shahin, S. Jaafreh, and Y. Azzam, "Tracking protein kinase targeting advances: integrating qsar into machine learning for kinase-targeted drug discovery," *PMC*, 2025.

[7] N. Mosos, T. Camargo-Roldan, O. Montoya, and J. Guevara-Pulido, "A machine learning method for predicting the ic50 values of novel designed analogs of non-nucleoside reverse-transcriptase inhibitors (nnrtis) as potentially safer drugs," *ScienceDirect*, 2024.

[8] R. Ancuceanu, M. Dinu, I. Neaga, F. G. Laszlo, and D. Boda, "Development of qsar machine learning-based models to forecast the effect of substances on malignant melanoma cells," *ScienceDirect*, 2018.

[9] T. R. Noviandy, G. M. Idroes, E. Harnelly, I. Sari, F. M. Fauzi, and R. Idroes, "Predicting axl tyrosine kinase inhibitor potency using machine learning with interpretable insights for cancer drug discovery," *Researchgate*, 2025.

[10] ChEMBL Group, EMBL-EBI, "Chembl target: Tyrosine-protein kinase syk (chembl2599)," 2025. Accessed: 2025-12-15; Bioactivity data curated from scientific literature.

[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Arxiv*, 2016.

[12] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Arxiv*, 2017.

[13] M. Hammad, "Artificial neural network and deep learning: Fundamentals and theory," *Arxiv*, 2024.

[14] H. A. Salman, A. Kalakech, and A. Steiti, "Random forest algorithm overview," *Researchgate*, 2024.

[15] D. Figueiredo, J. Silva, and E. C. R. Júnior, "R2 all about?," *Leviathan*, vol. 3, pp. 60–68, Nov 2011.

[16] T. Hodson, "Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, 2022.

[17] Unknown, "Mean absolute error (mae)," 2025. Accessed: 2025-12-15.