International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

Explaining Drug Discovery: A Comparison of White-Box and Black-Box Models with XAI

DR. SHALINI LAMBA (HOD), PRATHAM GUPTA

Department of Computer Science, National P.G College, Lucknow, Uttar Pradesh, India

ABSTRACT

This article discusses how to transform a black box into a more interpretable model than a white box, especially in the field of bioinformatics, using artificial intelligence (XAI). Black box models such as deep learning and integration are quite accurate but lack transparency, which prevents them from being used in important applications such as healthcare. In our previous article, we proposed XAI techniques that can transform these opaque models into more understandable models, allowing researchers to understand decision-making models. In this work, we apply XAI techniques to black box and white box models in bioinformatics, focusing on accurate measurement and interpretation. We use SHAP values and visual models to highlight the impact of various aspects of the prediction model by experimenting with random forests (as a black box model) and logistic regression and decision trees (as a white box model). Our results show a trade-off: models with black bars (such as random forests) generally achieve higher accuracy in capturing patterns of complex objects, while models with white lines provide a clear ordering process that is crucial for clear understanding. This paper highlights the utility of XAI and highlights the importance of translation in bioinformatics to ensure that AI-driven models are not only accurate, but also interpretable and reliable for these important applications.

KEYWORDS: Explainable AI (XAI), Logistic Regression, Decision tree, White Box Model (WBM), Black Box Model (BBM)

INTRODUCTION

ACTIVITY NUMBER:

0 (Inactive): The compound does not exhibit the desired biological activity or therapeutic effect.

1 (Active): The compound shows the desired biological effect, indicating potential therapeutic value.

EVALUATION METRICS [5]

1. Precision

Precision measures the delicacy of the unborn prognostications made by the model. Specifically, it's the rate of rightly prognosticated positive compliances to the total prognosticated positive compliances. Precision answers the question How numerous of all the exemplifications the model rightly prognosticated?

The formula for precision:

Precision= True Positives/ (True Positives+ False Positives)

- High precision indicates a low false positive rate (the model rarely misclassifies negative cases as positive).
- Low precision means the model incorrectly labels many negative instances as positive.

2. Recall

Recall, also known as sensitivity or true goodness, measures the ability of the model to identify true goodness. It is the ratio of the correct prediction of a positive observation to each positive outcome. Repeat the question: How many examples of each good case were identified?

The formula for recall:

Recall=True Positives / (True Positives+ False Negatives)

- High recall indicates the model detects most of the actual positive cases.
- Low recall means the model misses many positive cases, labeling them incorrectly as negative.

3. F1 Score

The F1 score is a compromise between precision and recall, giving an equal measure of the two. It is especially useful when you want to balance accuracy and recall, or when you have an unbalanced class.

The formula for F1 score:

F1 Score=2 × (Precision * Recall)/(Precision + Recall)

- High F1 score suggests a good balance between precision and recall.
- Low F1 score indicates an imbalance, where the model may have high precision but low recall or vice versa.

4. Support

Support is the number of true values for each group in the dataset (i.e. the number of samples that belong to each group). Support helps you understand the distribution of classes in your data and identify the scores, returns, and F1 for each class.

Support =Count of true samples in each class

Difference between Logistic Regression and Decision Tree [6]

Feature	**Logistic Regression**	**Decision Tree**

Type	Linear model for binary classification	Non-linear model for classification and regression
Output	Predicts probabilities that can be mapped to classes	Produces a tree-like structure of decisions
Interpretability	Generally easy to interpret, coefficients indicate impact of features	Can be visualized; easy to understand the decision path
Handling Non-Linearity	Limited to linear relationships; may struggle with complex patterns	Naturally handles non-linear relationships
Overfitting	Less prone to overfitting; regularization can be applied	Prone to overfitting, especially with deep trees
Performance on Imbalanced Data	May not perform well without adjustments	Often requires techniques like pruning or balanced data to handle imbalance
Feature Importance	Coefficients indicate the importance of features	Easily provides feature importance via split criteria

BLACK BOX MODEL

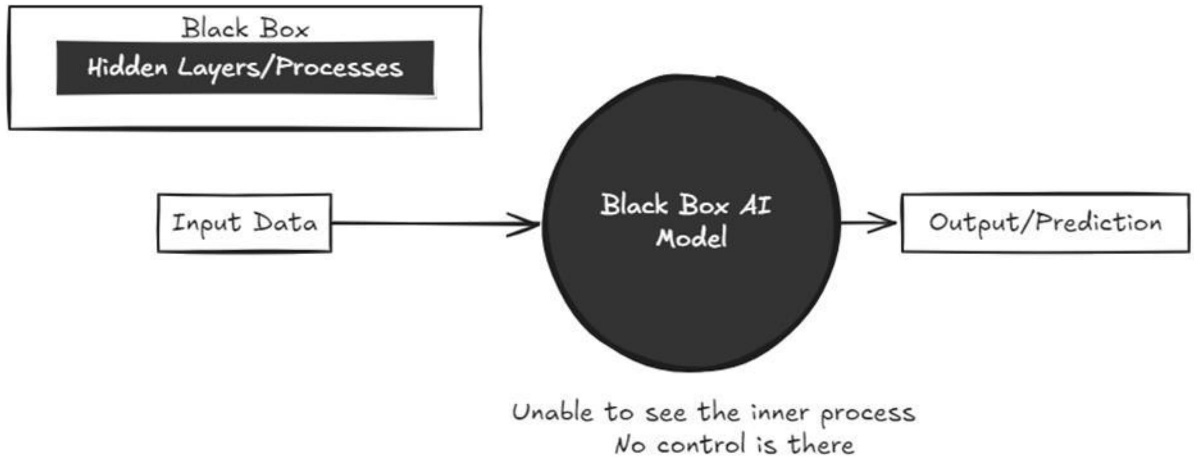


Figure 1 Showing the working of Black Box Model

A black box model in XAI refers to a machine learning model that operates as an invisible system where the inner workings of the model are not easily accessible or interpretable.[8] These models make predictions based on input data, but the decision-making process and the logic behind the predictions are not transparent to users.[10] This lack of transparency makes it difficult for users to understand the model's behavior and identify bias. [2]

OUTPUT OF BLACK BOX MODEL (Dataset reference[4]: Drug Data)

Classification Report:

Activity Number	Precision	Recall	F1 Score	Support
0	0.91	0.90	0.91	341
1	0.97	0.97	0.97	1134

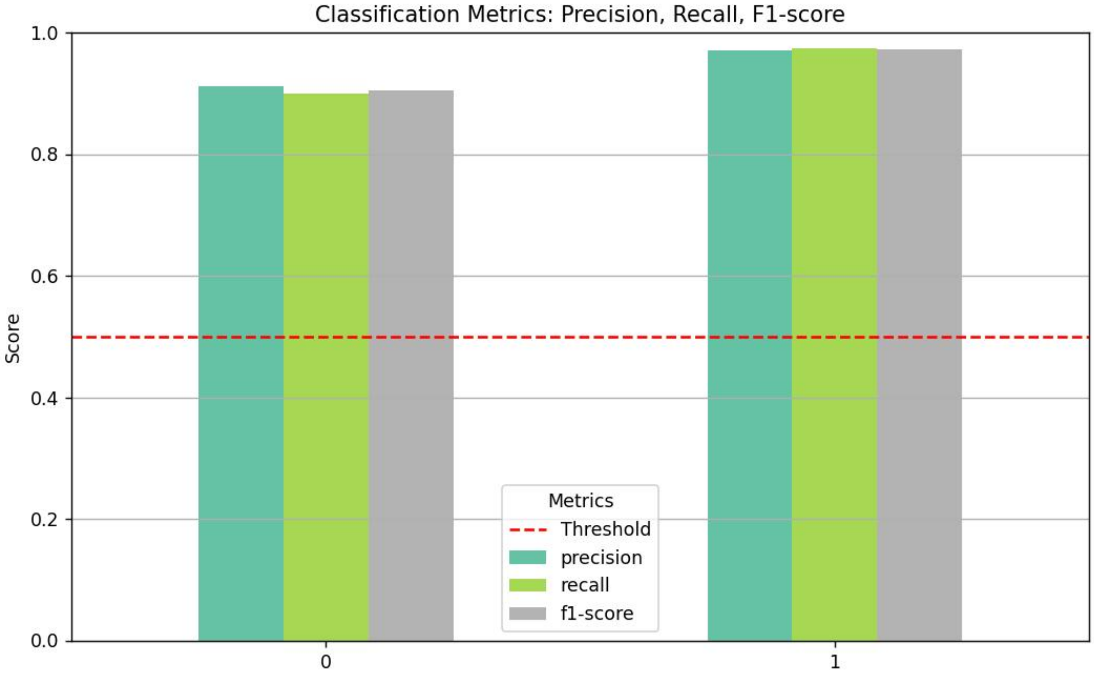


Figure 2 Bar Graph showing the output of Black Box Model

White Box Model

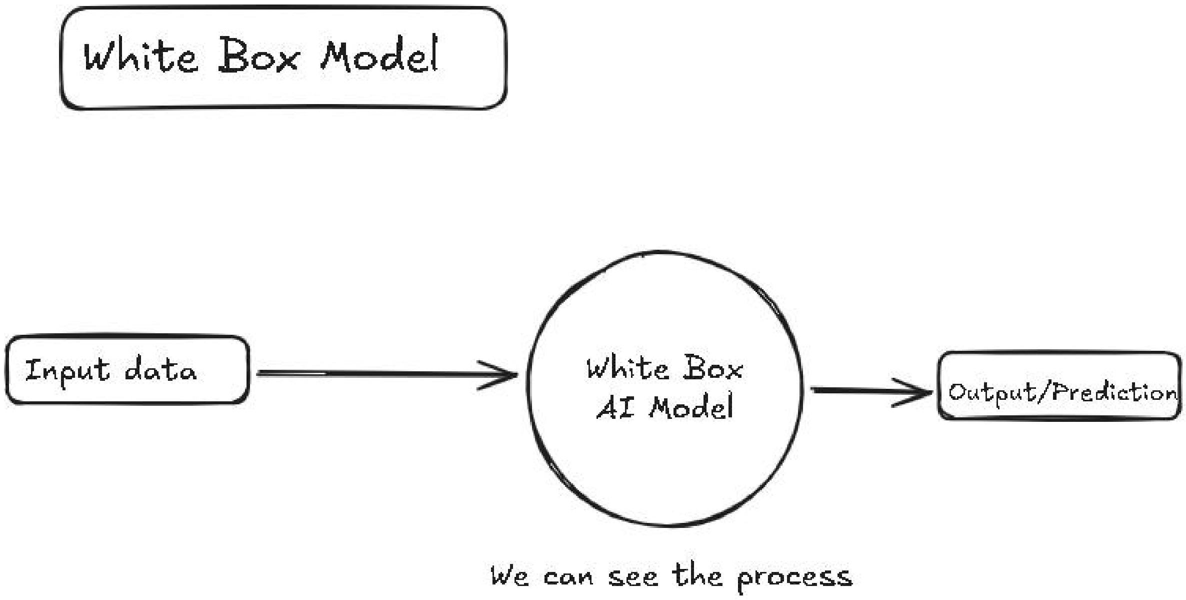


Figure 3 Showing the working of White Box model

The terms “white box,” “model understanding,” and “explanatory intelligence (XAI)” are used to describe machine learning models that provide results that can be easily interpreted by an expert. These standards typically strike a balance between accuracy and interpretation. The terms “understandable” and “explainable” are often used interchangeably to describe models that provide explanations to experts in a given domain. However, as noted, an understanding model needs additional models or features to generate explanations for experts. In contrast, a description model can provide answers on its own without the need for external assistance [3]

OUTPUT OF WHITE BOX MODEL (Dataset reference [4]: Drug Data)

Logical Regression Classification Report:

Activity Number	Precision	Recall	F1 Score	Support
0	0.91	0.90	0.91	341
1	0.97	0.97	0.97	1134

Activity Number	Accuracy
0	0.96
1	1475

	Weighted Average	Macro Average
Precision	0.96	0.94
Recall	0.96	0.94
F1 Score	0.96	0.94
Support	1475	1475

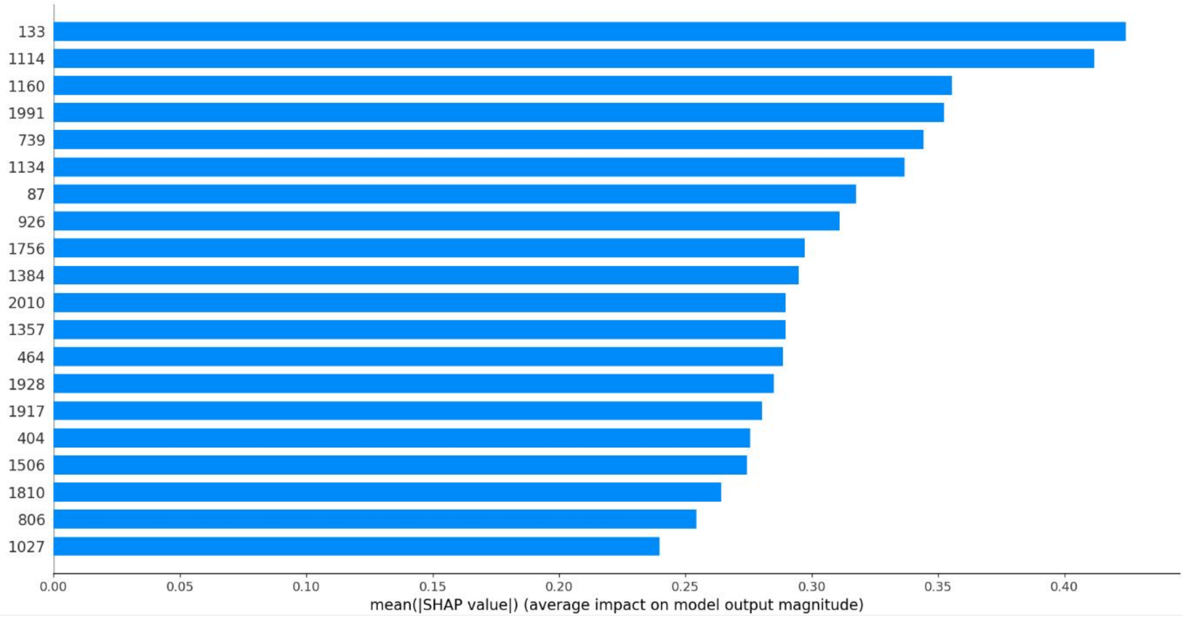
Decision Tree - Classification Report:

Activity Number	Precision	Recall	F1 Score	Support
0	0.82	0.59	0.69	341
1	0.89	0.96	0.92	1134

Activity Number	Accuracy
0	0.88
1	1475

	Weighted Average	Macro Average
Precision	0.87	0.85
Recall	0.88	0.78
F1 Score	0.87	0.80
Support	1475	1475

Figure 4 Feature Impact on Model Predictions Using SHAP Values



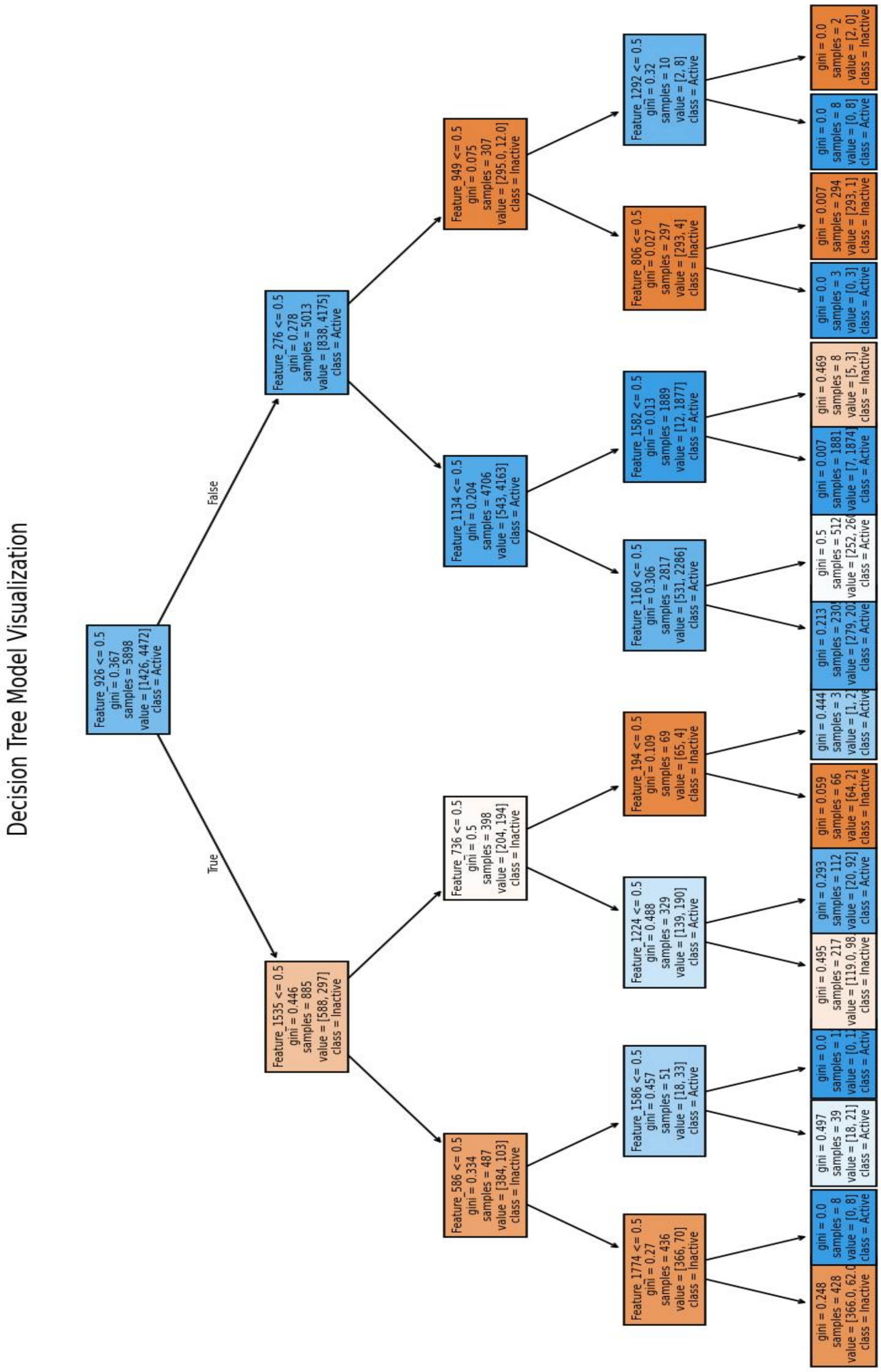


Figure 5 Decision Tree of White Box Model

USING SURROGATE MODELS:

Surrogate Models

A surrogate model is an interpretable model that approximates a more complex, less transparent “black-box” model (like a neural network or ensemble model) to make its decision process more understandable. Surrogate models are typically simpler algorithms, such as decision trees or linear models, which are trained to mimic the outputs of the black-box model.[11] By doing this, they provide a clearer view into which features or data aspects influence the predictions, helping stakeholders understand and trust the black-box model's behavior without compromising much on accuracy.[9]

For future work in improving transparency within black-box models, surrogate models offer a promising approach. They do not complete convert black box models to white box models. Studies indicate that using interpretable methods, such as decision trees or rule-based techniques, can help clarify black-box predictions.[12] For instance, research on approaches like SRules suggests that surrogate decision trees can distill complex ensemble models (e.g., random forests) into comprehensible rules.This strategy enables enhanced interpretability by focusing on feature importance and decision pathways, allowing users to understand model logic without reducing predictive effectiveness.[7]

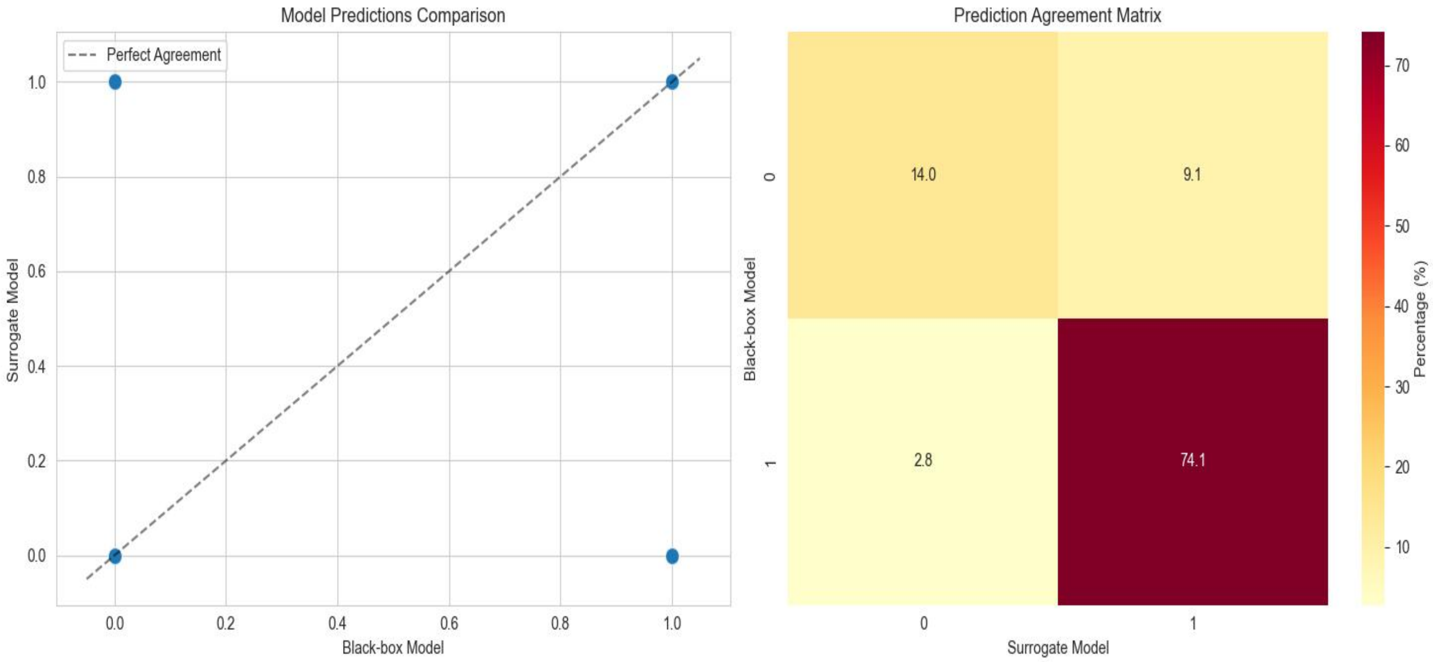


Figure 6 Comparison of Blackbox and Surrogate model

Surrogate Model (Decision Tree) Classification Report after using it on black box model:

Activity number	Precision	Recall	F1 Score	Support
0	0.81	0.59	0.69	341
1	0.89	0.96	0.92	1134

Activity Number	Accuracy
0	0.87

1	1475

	Macro Average	Weighted Average
Precision	0.85	0.87
Recall	0.78	0.87
F1 Score	0.80	0.87
Support	1475	1475

CONCLUSION:

We can conclude that both white-box and black-box have specific uses in machine learning, interpretation, and performance evaluation. In principle, we run two types of models on the same dataset: logistic regression and decision trees as white-box models, and random forests as black-box models. The white-box model provides transparency and clarity in the decision-making process. For example, decision tree modeling visualizes decision-making as a path, making it easier to follow the logic behind each prediction, while logistic regression specifically shows the effects of each. This transparency promotes trust and understanding, which is especially important in areas that often require translation, such as healthcare and finance. For example, black-box models like random forests are good at handling complex data and are often more predictive than simple, average models. However, this increase in accuracy comes at the cost of interpretation, because it is harder to understand the specific reasons behind the prediction in a random forest. Rules introduce trade-offs by running two types of models on the same data: White-box models provide clear insights into decision-making, while black-box models can be more efficient and better at interpreting applications.

In the future, one approach to enhance the interpretability of black-box models in drug discovery is the use of surrogate models. These models aim to approximate the predictions of complex, high-performance black-box models (such as Random Forests or Deep Neural Networks) using simpler, more interpretable models (such as Logistic Regression or Decision Trees). Surrogate models can be trained to replicate the behavior of a black-box model while offering greater transparency. This process involves creating a model that approximates the complex model's decision-making process but provides clearer, interpretable outputs. For example, a Logistic Regression or Decision Tree model can be used as a surrogate to simulate the predictions of a deep learning model, making it easier for researchers to understand the relationships between features and predictions.

REFERENCES:

[1] Enhancing Drug Discovery Efficiency Through Explainable Ai: Transitioning From Black Box To White Box Models Shalini Lamba, Pratham Gupta And Shrey Nagar, Department Of Computer Science, National P.G College, Lucknow, Uttar Pradesh, India

[2] Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Diyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, Amir Hussain, 24 August 2023

[3] Loyola-González, "Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View," in IEEE Access, vol. 7, pp. 154096-154113, 2019, doi: 10.1109/ACCESS.2019.2949286.

[4] https://github.com/yingzibu/JAK_ML/blob/main/new_data/JAK1_final.csv

[5] Chen, J., Zhang, J., Zhang, Y., & Chen, H. (2020). "A New Classification Model Based on Hybrid Deep Learning for Text Data." *IEEE Access*, 8, 201144-201154.

[6] Khan, M. I., & Ahmad, S. (2022). "Performance Analysis of Logistic Regression and Decision Tree Algorithms in Breast Cancer Detection." *Journal of Healthcare Engineering*, 2022, Article ID 8129203.

[7] An Interpretable Rule Creation Method for Black-Box Models based on Surrogate Trees – Srules Mario Parrón Verdasco Esteban García-Cuesta Departamento de Inteligencia Artificial, Universidad Politécnica de MadridMadridSpain.

[8] Enhancing black-box models: Advances in explainable artificial intelligence for ethical decision-making Jayesh Rane, Suraj Kumar Mallick, Ömer Kaya, Nitin Liladhar Rane

[9] The fidelity of global surrogates in interpretable Machine Learning Carel Schwartzenberg, Tom van Engers, and Yuan Li, University of Amsterdam, Amsterdam, The Netherlands

[10] Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology Jeremy Petch, PhD, MA, BA(H), Shuang Di, MSc, and Walter Nelson, BSc(H)

[11] From Black Boxes to Transparent Machines: The Quest for Explainable AI, Shalom Akhai Chandigarh College of Engineering, CGC Jhanjeri

[12] The Use of Surrogate Models to Analyse Agent-Based Models, Guus ten Broeke, George van Voorn, Arend Ligtenberg, Jaap Molenaar