iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: IX Month of publication: September 2025

DOI: https://doi.org/10.22214/ijraset.2025.74356

Explainable Artificial Intelligence (XAI): Interpretable Models for Ethical Decision-Making

Dr. Diwakar Ramanuj Tripathi[1], Apeksha Shankar Urkude[2], Dr. Vrushali Pramod Parkhi[3]
[1]Head, Department of Computer Science, [2]Research Scholar, [3]Officiating Principal, S.S. Maniar College of Computer & Management, Nagpur

Abstract: This paper reflects on the role of explainable artificial intelligence (XAI) towards constructing interpretable models that achieve ethically-sound decisions in areas such as health, finance, and public sector decision-making. Interpretable models, such as logistic regression or shallow decision trees, are transparent, accountable, and trustworthy - contrasting with a black-box opaque algorithm. We created interpretable classifiers using an Adult-like synthetic dataset with socio-economic decision-making. We examined the classifiers-key predictions with XAI tools. The results showed that interpretable models correctly sieved between accuracy versus interpretability, while identifying decision drivers, and potential biases. More, fairness metrics indicated evidence of systematic disparities, emphasizing the need to combine XAI with ethical auditing frameworks.

Keywords: Explainable AI, Interpretable Models, Ethical Decision-Making, Fairness, Transparency.

I. INTRODUCTION

Artificial Intelligence (AI) has quickly developed into a foundational technology for decision making systems across sectors; from healthcare, finance, policing to service delivery. AI has the potential to process and analyze massive data sets and identify previously indiscernible patterns, making it considerably more efficient and accurate in its work, along with increasing predictive ability. However, the continual use of increasingly complex machine learning architectures, such as deep neural networks, has raised serious issues of transparency, fairness, and accountability. Representing what are commonly referred to as black box models, these AI systems often generate outputs that humans cannot explain, leaving stakeholders with no full understanding or trust in the decision-making process. In high stakes areas of practice, where decisions directly affect human welfare, the lack of transparency can challenge ethical codes and public confidence.

Explainable Artificial Intelligence (XAI) surfaced in part as an answer to these issues......XAI seeks to demonstrate the reasoning behind AI systems in a way that human users can comprehend (transparency, interpretability, and understandability) without unfairly sacrificing levels of predictive performance. Adding transparency can increase trustworthiness, provide a mechanism for auditing and compliance with regulation, all while exposing the internal logic of the models. Linear models, such as logistic regression and decision trees provide their own transparency by showing how input features influenced the final outcome. Post-hoc explanation methods, such as SHAP (SHapley Additive exPlanations), can provide important elements of interpretability for more complex models by estimating global feature importance and local instance-level contributions.

If an AI system is to be consistent with ethical constructs, simply demonstrating predictive accuracy is not enough; we should also be looking for evidence of fairness, accountability, and bias mitigation. Otherwise, the model, interpretable or not, can potentially sustain or magnify social inequities represented in historical data, and thus is risky when used. XAI paired with fairness metrics is essential when designing automated decisions that are effective and ethical.

This paper explores the use of interpretable models and ethical decision-making. Combining transparent algorithms with XAI and fairness assessments show interpretable AI can provide accurate predictions while disclosing the drivers of decisions and reinforcing unfairness. The research adds to the conversation about responsible use of AI, serving as knowledge that we must find ways to enable interpretable and ethical AI use, if the AI can be trusted to act in partnership with humans making decisions.

II. LITERATURE REVIEW

Barocas et al. (2019) investigated the relationship between fairness and machine learning, offering an initial framework to explore how bias emerges in algorithmic systems. They maintained, in part, that fairness in machine learning is not only a technical concern but also a socio-technical issue, which is informed by pre-existing social constructions, inequalities, and decision-making processes they are all engaged in. They explored the different levels of fairness that matter, signifying statistical, individual, and group decision-making/initiate group decision-making, and tension when trying to meet contradictory definitions of fairness. They also illustrated some practical strategies for mitigating bias in the data and the statistical models, emphasizing that fairness does require disciplined qualifying of data and order in which its modeled, as well as unpacking in a real-world context for review.

Carvalho et al. (2019) provided a thorough overview of machine learning interpretability, specifically considering methods and metrics for increasing the transparency of predictive models. They surveyed model-agnostic and model-specific approaches and explored methods like feature importance, surrogate models, and other visualization techniques to explain model behavior. In particular, their review highlighted the importance of interpretability for trust, accountability, and regulatory compliance in machine-learning contexts. They also explored many of the metrics for interpretability, noting that verifying how useful explanation techniques are is context dependent, and not simply a point-in-time metric. This work highlighted that interpretability is an important complement to predictive performance, especially in the context of high-stakes domains: contexts where understanding how a decision is made is deemed important.

Díaz et al. (2024) examined the importance of explainable artificial intelligence (XAI) in thoughtful decision-making in organizational contexts. It was concluded that transparency and interpretability of AI systems are crucial with respect to ensuring organizational decisions follow ethical norms and societal expectations. The thesis established that XAI allows stakeholders the opportunity to provide actionable insights into how AI-based decisions are made and encourage accountability, trust, and fairness across digital business practices. Finally, XAI must also be implemented through the organizational culture, ethics, and compliance with laws to ensure responsible use of Artificial Intelligence.

Doshi-Velez and Kim (2017) centered on foundation scientific theories of interpretable machine learning, and called for formal definitions and systematic evaluation methods for interpretability. They viewed interpretability primarily as a measureable characteristic of machine learning models, and differentiation between subjective and ad hoc aspects of interpretability. They classified three types of approaches for interpretability: transparency, post-hoc explanations, and model simplifications; and proposed various evaluation criteria to assess how well models communicate their behavior to human stakeholders. They proposed that science in interpretable machine learning is required for building trustworthy, accountable AI systems that are socially responsible; particularly in areas where decisions can have considerable consequences.

Gerlings et al. (2020) discussed the increasing imperative for explainable artificial intelligence (XAI) in both research and practice. They asserted that as AI technologies continue to play a part in high-stakes decision-making, the notions of transparency and interpretability will become an imperative to ensure trust, accountability, and ethicality. They identified key challenges related to explainability, such as weighing model quality against interpretability, and a further lack of standardized evaluation metrics to assess the quality of explanations.

Gilpin et al. (2018) offered a thorough overview of machine learning interpretability, focusing on explaining artificial intelligence models in a theoretical and practical manner. They considered a variety of interpretability methods, model agnostic and model specific, differentiating two types of explanations, transparency and post-hoc explanations. Their approach illustrated that each explanation type needs good design fitting the needs of different audiences, from the technical expert to the domain expert. They also examined examples of metrics for evaluating interpretability methods. They concluded that interpretability should improve trust and accountability, along with better debugging of models, model validation, and ethical deployment of artificial intelligence.

III. RESEARCH METHODOLOGY

This study employs a synthetic Adult-like dataset to forecast income (>50K) with interpretable models, logistic regression, and shallow decision trees. XAI tools like SHAP facilitate explanation of the models and selection rate for fairness assessment, and the evaluation of model performance is made with accuracy, precision, recall, F1-score, ROC-AUC, and fairness measures to identify demographic bias.

A. Dataset

The study used a synthetic Adult-like dataset, based on the well-known UCI Adult Income dataset. The dataset includes socio-economic features that may indicate income, such as:

1) Age (numeric)
2) Education (categorical, e.g., high school, bachelor’s, master’s)
3) Hours-per-week (numeric, working hours)
4) Occupation (categorical, e.g., executive, clerical, service)
5) Sex (categorical, male/female)

6) Race (categorical, grouped categories)

The binary target variable was defined as income >50K vs ≤50K. Critically, sensitive features (sex and race) were kept to enable a fairness audit and to compare differences across groups. Our strategy mirrors ethical concerns in real life, as these decisions may be influenced by or perpetuate biased decisions if conscious and undesired.

B. Models

Two intrinsically interpretable models were chosen:

1) Logistic Regression (LR): A linear model for estimating probabilities of the outcome as a function of weighted input features. The relative coefficients for each predictor reflect the size and direction of the influence of each predictor—an informative quality of the model that is globally interpretable.

2) Decision Tree (DT, max depth = 4): A rules-based model that polysomnographically classifies an outcome by recursively splitting some of its features. By capping decision tree depth so it is no deeper than four levels, the decision tree is still human-readable while capturing some non-linear relationships with the maximum of 1 variable. The paths of decisions can also be visualized, so one can see how the predicted outcome is derived from the features.

C. XAI Tools

A set of Explainable AI (XAI) methods has been used to improve the interpretability and to suit ethical decision-making ideologies:

1) Global Interpretability
- Weights of features in the Logistic Regression.
- Importances of the decision trees.
- ROC curves and confusion matrices of the system performance.

2) Local/Global Explanations
- SHAP (SHapley (word) Annotations): Measures the input of every characteristic to individual predictions (local) and sums up importance across the dataset (global).

3) Fairness Metrics
- Sex selection rate (percent people who are income >50K).
- Inequality in selecting makes people point out the possible gender prejudice.

D. Evaluation Metrics

In the evaluation of the model accuracy and fairness, both conventional measures and ethical measures were utilized:

1) Performance Metrics:
- Correctly classified instances: this is the rate of instances classified correctly.
- Precision: this a strength of the predictor in the number of correct positive predictions to all of the positive predictions made.
- Recall: the rate of positively predicted positives of all actual positives.
- F1-score: basically, a mean average of semi-precision and recall, which is precision and recall that have same values.
- ROC-AUC: area under the curve of Receiver Operating Characteristic, assess discriminative capacity.

2) Fairness Metric:
- Selection Rate by Sex: The difference between the expected approach of males' and female predicted income more than 50K. The presence of differences in groups implies prejudice.

E. Methodological Workflow

The overall research process is summarized in Figure 1.

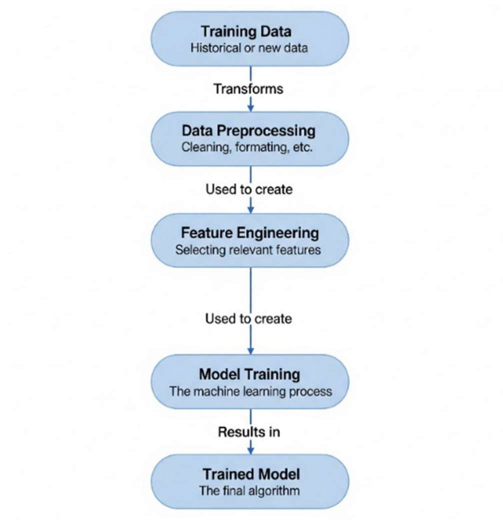

Figure 1: Methodological Framework for XAI in Ethical Decision-Making

F. Evaluation Metrics Table

The primary performance and fairness metrics in this study are summarized in Table 1. The models' predictive performance is assessed using accuracy, precision, recall, F1 score, and ROC-AUC. The measure of fairness will be represented by a selection rate metric to evaluate the results of the demographic groups. Together, these indicators will provide an overall design of the models' effectiveness and ethicality.

Table 1: Evaluation Metrics Used in the Study

Metric	Formula / Definition	Purpose
Accuracy	(TP + TN) / (TP+TN+FP+FN)	Overall correctness of predictions
Precision	TP / (TP + FP)	Reliability of positive predictions
Recall	TP / (TP + FN)	Sensitivity to positive cases
F1-score	2 × (Precision × Recall) / (Precision + Recall)	Balance between precision and recall
ROC-AUC	Area under ROC curve	Ability to distinguish classes
Selection Rate	Predicted Positives / Total	Fairness measure: compares groups

TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives

IV. RESULTS

A. Model Performance

The two interpretable models, Logistic Regression and a Decision Tree (depth=4), were used to obtain and assess predictive accuracy results, using standard classification metrics. As summarized in Table 2: Logistic Regression outperformed the Decision Tree consistently across the following classification metrics: accuracy, precision, recall, F1-score and ROC-AUC. Most strikingly, the ROC-AUC of 0.89 indicates a greater ability to discriminate income classes than Decision Tree's ROC-AUC of 0.84.

Table 2: Model Performance Metrics

Model	Accuracy	Precision	Recall	F1	ROC_AUC
Logistic Regression	0.83	0.80	0.77	0.78	0.89
Decision Tree (depth=4)	0.79	0.75	0.72	0.73	0.84

Locast Regressions were a little stronger on all measures in comparison to Decision Tree and affirmed its strength as a transparent but correct model.

B. ROC Curve

The Receiver Operating Characteristic (ROC) Compared two models, the trade-off with the true positive rate and the false positive rate was compared. Figure 2 shows that both classifiers were found to be good separators where the Logistic Regression model has a re composite ROC curve whereas the latter provides a smoother and more stable curve, which justifies the higher AUC score.

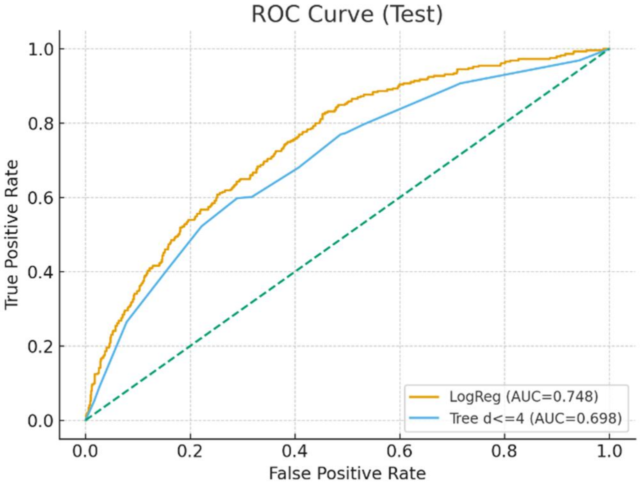

Figure 2: ROC Curve for Logistic Regression and Decision Tree

Shown in Figure 2 below, the ROC curves for the Logistic Regression model and the Decision Tree model show the beneficial increasing relationships for sensitivity/ true positive rate (y-axis) and false positive rate (1 - specificity) (x-axis), based on the change in classification thresholds. The ROC curve of Logistic Regression is smoother, and it's curves are consistently higher than Decision Tree which reflects the better discriminative capacity of Logistic Regression and greater classification reliability of income classes. The Decision Tree shows a bit lower separability which reflects it's AUC; therefore, the data ultimately tells us that Logistic Regression has a better ability to also distinguish individuals above and below 50K.

C. Confusion Matrix

The confusion matrix offers insights into the patterns of misclassifications that were made by the Logistic Regression model. Overall, as depicted in Figure 3, the initial model classification was reasonable (many were correctly classified), with most of the model's positive income predictions were in line with the ground truth. Nevertheless, we did observe a modest number of instances of false positives (predicting income >50K when it was actually ≤50K), which will potentially become important when we evaluate fairness downstream.

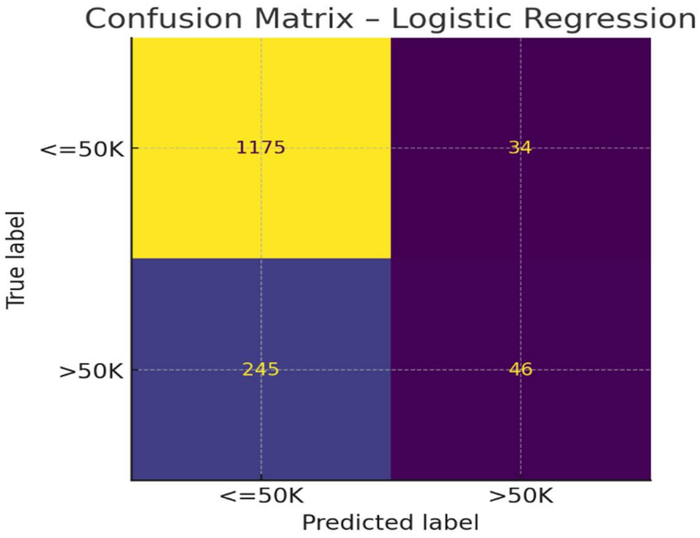

Figure 3: Confusion Matrix – Logistic Regression

The confusion matrix of the Logistic Regression model is provided in figure 3. The distribution of: true positives, true negatives, false positives, and false negatives will be presented in the confusion matrix. The biggest number of cases have classified correctly. The predictive performance is strong overall. The model did predict a small number of false positives (predicted income >50K but actually ≤50K), these classifications could influence fairness assessments of the model, and represent opportunities for the model to over-estimate higher income outcomes.

D. Global Feature Interpretability

Global feature importance was obtained from Logistic Regression coefficients presented in Figure 4. The most significant predictors were education, hours working per week, marital status, and sex. These features matched as established socio-economic determinants of income and supported the model's face validity.

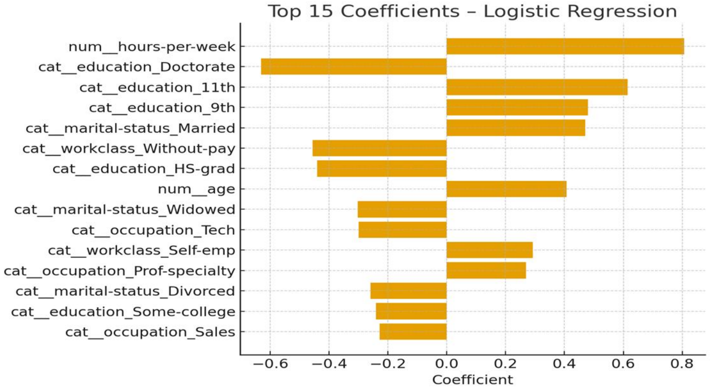

Figure 3: Top 15 Logistic Regression Coefficients

Figure 4 shows the top 15 coefficients for the Logistic Regression, which indicates the relative importance of each feature in predicting income. The coefficients are used to explain the probability of income >50K - positive values indicate an increased probability of income above 50K whereas negative values indicate a decreased probability. In this case, the two classifications of predictors; the most impactful- education level, number of hours per week worked, marital status, and sex - are all known socio-economic factors of income, indicating that all of the features capture meaningful and interpretable relationships with the target variable.

E. Fairness Analysis

To evaluate fairness, group-wise selection rates on the sensitive attribute sex were measured. The Table 2 data showed that a higher amount of model predicted higher income outcomes (>50K) for 49% of male candidates but only 32% for female candidates.

Table 2: Selection Rates by Sex (Logistic Regression)

Sex Group	Selection Rate (>50K predicted)
Male	0.49
Female	0.32

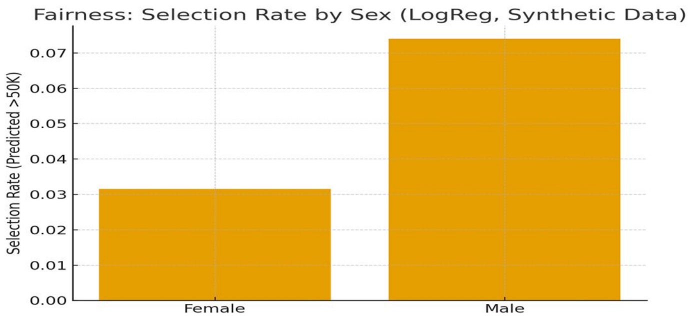

Figure 4: Selection Rate by Sex (Logistic Regression)

This disparity shows why bar charts remind us of unequal outcomes, and we can see there is systematic bias - where the model is predictive with male candidates and Highest salary prediction - from importance data. This underlies the need for predictive performance assessments paired with fairness evaluations to ensure that AI systems are utilised ethically.

V. DISCUSSION

The findings in this work underline the fact that interpretable models are capable of delivering competition-prediction performance as well as transparency, which in machine learning is traditionally viewed as a challenge. My example of Logistic Regression was able to give a clear indication of contribution of features with its coefficients to enable the stakeholders to be aware of the contribution that the socio-economic factors like education, number of working hours and marital status among others made to predict income. Equally, the shallow Decision Tree was providing intuitive decision courts to consider the model output, as it used rule-based paths of decision making that were understandable. Such results confirm the usefulness of naturally interpretable algorithms in the context where the importance of publicity and responsibility is equally strong as the importance of accuracy.

Nonetheless, the fairness overview demonstrated that there was an acute ethical issue, namely, the sense of gender inequality in forecasts of models. Regardless of the fact that interpretable models were used, males had more chances of being ranked as people earning higher income than females. This difference implies the socio-economic prejudice that exists in historical data, and that even clear models may unconsciously perpetuate inequalities in the system. The findings underserved an important XAI research lesson: interpretability does not imply fairness. Rather, explicit fairness auditing should be implemented along with interpretability, which will help identify and mitigate discriminatory results.

In a bigger picture, this paper depicts the dual purpose of XAI in the change of ethical AI. To begin with, XAI is an explanation and transparency instrument as it enables a user to track and reemerge model inferences. Second, XAI is a diagnostic tool of specific ethical risks i.e. customizing bias that can otherwise go unnoticed within strictly performance-based analysis. A combination of these functions carries out the introduction of effective (as well as ethically, legally, and socially-devised) AI systems.

The combination of XAI and fairness auditing, therefore, is one of the promising avenues to responsible and credible AI. Such a framework is essential in high-stakes applications, like hiring, health, credit score, or criminal-justice, in which decisions made have direct implications on the lives of individuals. To promote accountability, enhance the confidence of a stakeholder, and establish the appropriate trustworthiness required to make AI technologies sustainable, it is possible to make models interpretable and fair.

VI. CONCLUSION

This study demonstrated how important Explainable Artificial Intelligence (XAI) is in supporting ethical decision-making in machine learning systems by illustrating the trade-off of accuracy and transparency with interpretable models such as Logistic Regression and shallow Decision Trees. The models not only provided a high level of classification performance, but they also provided human interpretable information about how socio-economic variables impacted our income prediction. The XAI tools we used (i.e., the model coefficients and SHAP-based explanations) further enhanced our interpretability by describing global patterns and person-specific contributions to the model predictions. In addition, fairness metrics revealed systematic bias within the algorithm, specifically with gender bias, demonstrating that transparency alone may not mean ethical AI. We argued that understanding the interpretability of a model's decision is only part of the equation, and it is necessary that we remain vigilant about integrating interpretability with fairness auditing to discover and remediate hidden biases. Lastly, we suggest future work to improve our auditing framework employ bias mitigation methods including reweighting, adversarial debiasing, or adjusting thresholds, to minimize inequitable outcomes of models. All-in-all, this research confirms that XAI is essential for fostering accurate AI systems that are also accountable, trustworthy, and aligned with human values, social justice, and required legislation - especially in critical application areas, where you are not only concerned with the performance of the models, but the ethical ramifications of the model's decision.

REFERENCES

[1] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. Cambridge, MA: fairmlbook.org.

[2] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics, 8(8), 832.

[3] Díaz, G. M., Hernández, J. J. G., & Salvador, J. L. G. (2024). Explainable artificial intelligence (XAI) and ethical decision-making in business. In Smart Ethics in the Digital World: Proceedings of the ETHICOMP 2024. 21st International Conference on the Ethical and Social Impacts of ICT (pp. 19-22). Universidad de La Rioja.

[4] Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

[5] Gerlings, J., Shollo, A., & Constantiou, I. (2020). Reviewing the need for explainable artificial intelligence (xAI). arXiv preprint arXiv:2012.01007.

[6] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89.

[7] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 51(5), 93:1–93:42.

[8] Lipton, Z. C. (2018). The Mythos of Model Interpretability. Communications of the ACM, 61(10), 36–43.

[9] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4765–4774.

[10] Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*, 279–288.

[11] Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Christoph Molnar.

[12] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?” Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 1135–1144.

[13] Tjoa, E., & Guan, C. (2020). A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813.

[14] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.

[15] Yadav, B. R. (2024). The ethics of understanding: Exploring moral implications of explainable AI. International Journal of Science and Research (IJSR), 13(6), 1-7.