National College of Ireland

Ethical Considerations in Explainable Artificial Intelligence: Transparency and Accountability in AI Decision-Making

MSc Research Project
MSc AI for Business

Tahir Jamil
Student ID: x23127732

School of Computing
National College of Ireland

Supervisor: Brian

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	**Tahir Jamil**
Student ID:	**x23127732**
Programme:	**MSc AI for Business** **Year:** **2023**
Module:	**Final Research**
Supervisor:	**Brian**
Submission Due Date:	**12 Aug. 24**
Project Title:	**Ethical Considerations in Explainable Artificial Intelligence: Transparency and Accountability in AI Decision-Making**
Word Count:	**9111 Page Count 22**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: **Tahir Jamil**

Date: **12 Aug. 24**

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	

Penalty Applied (if applicable):	

Ethical Considerations in Explainable Artificial Intelligence: Transparency and Accountability in AI Decision-Making

Tahir Jamil
Student ID: x23127732

Abstract

In many industries, AI has emerged as a powerful tool, which has raised valid concerns over the opaqueness of algorithms used in decision making processes. This paper aims to investigate the ethical considerations surrounding AXI with a focus on critical industries including; healthcare, banking and criminal justice. The main question is focused on how to effectively address the problem of realizing the applications of AI techniques as well as the demands for the interpretability, explainability, and audibility of AI decisions while promoting justice, accountability, and privacy. In a similar manner, XAI approaches are assessed, prototyped within this study, and case studies were conducted to explain how XAI could be implemented. It also entails consultations with the stakeholders in order to identify some of the issues and goals that they may have regarding the use of AI in activities such as transparency and accountability. The study's implications indicate that XAI has the potential to improve AI governance to become more transparent and fair in applying AI technologies through eliminating risks and algorithm bias, as well as strengthening the level of trust of all interested parties. Recognizing the lack of congruency between technology adoption and its ethical implications in literature, this research will help in the progression of proper implementation policies of AI solutions.

1 Introduction

Importance of AI in Decision-Making

AI is now a crucial part of the present-day decision-making frameworks of various industries as it acts as a game-changer when it comes to organizational strategies and decisions. AI systems that operate in today's world, rely on complex algorithms coupled with humongous data stores to make assessments, predictions and computation. The advancement of technology through the years has brought many positive changes to the areas of efficiency and accuracy of business decisions. For example, in the monetary field, the application of artificial intelligence implies the usage of models that assess trends in the financial market and customers' behavior in order to predict the best investment strategies and minimize potential risks. In medicine the applications are used to help in the diagnosis of diseases and in determining the appropriate treatment course for each patient which may increase efficiency and lower cost. AI also plays the role of modeling large data sets for efficient

analysis in order to make better decisions because it would otherwise be time consuming and cumbersome to obtain such information.

Besides, the capacity of AI in getting trained on new data makes decision-making processes more relevant and up-to-date. Artificial intelligence for instance has subcategories involving machine learning, possibilities of recommendations and predictions improve with time as the models adapt from previous decisions made and consequent results. The interactive learning process thus plays a central role in improving the decision systems hence contributing to the enhancement of the decision making frameworks. In the same manner, AI helps reduce man hour's utilization in repetitive tasks, hence allowing human capital to offer their best in key areas. In applying AI, the improvement of operation has been considered key in creating new sources of competitive advantage and importance within different industries.

The application of AI in decision-making also enhances the use of big data in decision making as opposed to the use of assumptions and estimations. Thus the transition to evidence-based working reduces biases, makes processes more consistent and general, and, therefore, yields more reliable outcomes. Moreover, AI systems allow for processing huge volumes of data accumulated from various sources and gives an integrated vision of the overarching business processes to help make the right decisions. Based on this, it can be stated that reliance on AI is growing and, thus, AI plays a critical role in defining the future of decision making.

Research Question.

In high-stakes applications, how can we ensure justice, accountability, & privacy protection while striking an appropriate balance between the performance & efficiency of artificially intelligent systems and the transparency & interpretability of artificial intelligence processes?

Ethical Challenges in AI

It is critical to consider some of the ethical issues which arise once AI systems are integrated into decision-making processes. Some of the main risks include the issue on bias in AI algorithms. Since AI systems learn from identified data, the data that is fed to it contain prejudices which are a reflection of past discriminations. Should these biases not be resolved, it becomes apparent that the AI models may even exacerbate these and produce unal fair and discriminative results. For example, recruitment algorithms may be structured in a way that discriminates against a certain demography harming certain categories in the employment practice. The consideration of bias in AI means continuing to work on maintaining the fairness of datasets, conducting proper reviews of algorithms' work, and implementing countermeasures against discrimination where they exist.

The next key ethical issue is wickedness and inexplainability of many artificial intelligence systems. Most of the AI models especially the sophisticated ones such as deep neural networks are 'black box models' that are difficult for users to comprehend on how the decisions are arrived at. Its increased obscurity leads to questions about transparency and

source of their authority, in that the stake holders will find it difficult to question the outputs of AI systems. It remains critical that the AI systems under their use give understandable reasons as to their conclusions so as to enhance transparency and users' confidence. Explaining AI (XAI) aims at solving this problem by creating approaches and strategies that make it easier to understand how AI makes the decisions. Also, the growth of using AI in high-risk fields like medicine and law leads to ethical concerns concerning privacy and data security. The need for large datasets is the main reason why most AI systems collect personal data and lawful but raises concerns on the manner in which it's collected, stored, and utilized. While nurturing the benefits of data in AI applications, it is imperative to protect the privacy of the individuals and act ethically to garner people's trust. Making certain safer methods of data protection, pursuing ethical protocols of data usage are important in handling such issues.

Objective of the Thesis

The overall goal of this thesis is to investigate and provide a solution to a challenging trade-off for AI systems, which is between the system's effectiveness and the system's ability to be explainable, transparent, and/or accountable especially when working on critical and sensitive applications. When AI technologies are to make critical decisions in various fields including but not limited to healthcare, finance and criminal justice, the fairness of these systems becomes rather crucial. This study seeks to understand the effectiveness of currently proposed XAI methods in order to determine the extent to which they facilitate development of reliable and understandable explanations of AI's actions. Thus, the aim of the thesis is to understand how different XAI techniques could improve models' explainability and use this knowledge to decide if and how the performance and speed of these methods should be changed. The aim is to find the practices which enhance the epistemological acceptance and trust of the AI procedures and match the ethical and legal compliance.

Besides, the thesis will explore various ethical theories/paradigms with a view of finding out the indispensable precepts to guide the creation and application of AI technologies. The analysis of the literature will help identify the views of the stakeholders regarding the use of AI transparency and accountability. The research shall come up with prototypes and case studies that provide a clear example of how XAI works in practice and adapts to the testing grounds from the permission of the experiments' subjects to the potential consequences on justice and societal welfare. Ethical reviews will be conducted to determine the changes that AI technologies bring to the ethical systems of a society, as well as the citizens' rights. In light of the conclusions, the thesis will develop guidelines for the ethical governance of AI to enrich the conception of AI based on efficacy and efficiency with justice, accountability and respect for privacy.

Dataset Overview

The evaluated dataset is known as the "Bank Marketing" dataset and is readily available in the UCI Machine Learning Repository; this dataset has its origin in business and is multivariate and intended for classification. The dataset comprises of 45211 records, and each record is a marketing communication of a client who came across the Portuguese bank's

marketing campaigns. The given dataset has 16 variables, where some of them are categorical and the others are integer type. Such elements include demographic data, account data, and contact information, which give the clients' details as well as the impact of marketing strategies.

Key Features

- Age: An integer representing the client's age.
- Job: A categorical variable indicating the client's occupation with values such as 'admin.', 'blue-collar', and 'technician'.
- Marital Status: A categorical variable representing the client's marital status with values such as 'married' and 'single'.
- Education: A categorical variable detailing the client's education level, ranging from 'basic.4y' to 'university.degree'.
- Default: A binary variable indicating whether the client has credit in default.
- Balance: An integer representing the average yearly balance in euros.
- Housing: A binary variable indicating whether the client has a housing loan.
- Loan: A binary variable indicating whether the client has a personal loan.
- Contact: A categorical variable specifying the type of communication used ('cellular' or 'telephone').
- Day_of_week: A date variable representing the last contact day of the week.

Relation of the Dataset to the Study

The "Bank Marketing" data set is particularly useful for this thesis as it offers applied context to explainable AI (XAI) techniques in the classification problem. The characteristics of the given dataset are diverse and as varied as the features of people and as such allow for investigation into how various aspects are involved in AI decision-making. Consequently, the study of using XAI methods for this dataset can explore the usefulness of these methods in offering understandable and interpretable explanations concerning the AI model's predictions of market-related decisions. Thus, in the classification problem, in which the goal is to determine whether the client will take a term deposit based on different characteristics, it is possible to examine the use of models with different degrees of explainability and the impact on their results. The nature of the features, both categorical and numerical, both binary and continuous, allows to better understand how different kinds of XAI techniques function and explain different kinds of data. In that regard, the dataset reflects other business applications, which can help to establish AI's transparent and accountable operation in critical situations

that may lead to significant financial consequences. With this dataset, the research can look at how well different XAI methods can explain the elements that affect the decision-making in marketing and guarantee that the use of AI in business processes is not only efficient but also transparent. The findings from this work will help in formulating the guidelines for ethical AI regulation and achieving the optimal level of performance, efficiency, and model interpretability.

2 Related Work

2.1 Ethical Considerations in Artificial Intelligence

Li (2023) presented a comprehensive evaluation of the ethical concerns tied to the use of AI, with a spotlight on utilizing computer vision. The knowledge cultivated by the study revealed that the ethical conundrums of AI systems depend on privacy infringement, bias reinforcement, and who is responsible for them? Li stressed that for AI models to be more fair, transparent and responsible especially in areas where privacy and societal welfare is at risk more focus has to be put in this risk considerations. The study highlighted the significance of the committed ethical data practices promoting fully transparent decision-making processes and emphasizing structured interdisciplinary and stakeholders' engagement in ethical AI development.

Another study by Safdar, Banja & Meltzer (2020) is Self-regulation of AI in radiology: an ethical analysis of fairness, accountability, and bias, that discusses the problems with deploying AI in the field. Their study focused on the specific ethical issues that arise from AI solutions in the context of healthcare; they further emphasised that there is a need for ethics in the use of AI systems used in clinical practice. The authors put emphasis on transparency and accountability as the essential components for safeguarding the patient's interest in AI diagnosis. They would have agreed with better data governance measures and intense reporting of such AI algorithms to enhance patient data quality as well as avoid biases that may be administrated into the diagnostic processes.

In summary, both types of works highlighted the importance and relevance of ethics and ethical principles in the development and application of artificial intelligence in global contexts and particular fields. It included recommendations that practitioners should go the extra mile and advocate for policies, which would ensure that fairness, transparency, and accountability are integrated into AL models and solutions to ensure that high ethical standards are upheld, and society's well-being is enhanced. When the stakeholders are involved in ethical decisions and different institutions coming from different fields communicate, this will help the researchers and practitioners to ensure that there are developing Ay systems that can be efficient not only in their operations but also ensuring that there is an ethical aspect which is considered in the development of the system.

2.2 Accountability and Transparency in AI Systems

In the application of artificial intelligence, issues to do with transparency and accountability for decisions made by the system have risen to be very crucial. Studies within this area concentrating on the concept of XAI explores the questions of how it is possible to use AI both effectively and ethically. This is especially so because AI solutions are increasingly being deployed in sensitive areas like medicine, banking, and law which rely on clear decision-making systems that can bail human beings out of major trouble where necessary. Rudin (2019) has elaborated on the problems with non-interpretable black boxes and called for the increased use of the interpretable models to increase model's responsibility and transparency. The kind of models that it supports are the ones that facilitate insight into how AI concludes decisions – this way, decisions made can be scrutinized and validated. Besides, this provision helps in building trust between the organization and consumers while reducing on risks such as biases or errors that may be concealed by complex systems.

Similarly, a recent systematic review was carried out by Jobin, Ienca, and Vayena (2019) on the Global State of AI Ethics and based on the assessment, accountability and transparency form the core aspects of the AI systems. Their research is evidence of the general need to have well understood and manage decision making processes in AI. Therefore, they advocate for the practical implementation of the reliable mechanisms of accountability and improving the usage of the AI models capable of interpretation, as well as strengthening the trust within the framework of the relationships between the stakeholders and the ensuring of the benefit from the AI systems , excepting the escalation of the negative aspects for society. Combined with these approaches, one is able to outline the general emphasis of the ethical frameworks – the requirement of transparency and accountability of AI decisions. As XAI not only involves technical concerns, but also has social implications, ethical issues in the XAI context are multifaceted and require cooperation with specialists from various fields and involving all the stakeholders who will be affected by the system. This approach is to guide AI innovations to incorporate and meet appropriate ethical standards in a way that technology aligns with the societies' desirable norms and encourage the effective and proper creation of AI systems that can benefit the communities globally.

2.3 Ethical Data Practices and Big Data Computing

In the context of "Ethical Considerations in Explainable Artificial Intelligence: Transparency and Accountability in AI Decision-Making," Kune et al. (2016) discussed how big data computing entities are complex to analyze and stressed on the significance of proactive data management measures to maintain the accountability in systems of artificial intelligence. Their study highlighted the need to develop such predefined approaches and integration of experts from other fields to develop ethics for the proper usage of AI in such fields like computer vision. The research emphasized that for transparency to be achieved and for AI decision-making process to be accountable, then structures that govern processing of data have to be clear have steps involved accompanied the general theme of the paper which is ethical use of AI and emphasis on the importance of ethical considerations towards the use of AI.

Kune et al. noted that in order to attain accountable artificial intelligence computational paradigms, one must have a definite understanding of big data computing. I learned that they dispel the ethic issues caused by big data processed by artificial intelligence and urged that several standards of accountability should be put in place to prevent cases of misuse of data and bias. The authors also focused on the need of incorporating ethical data governance among developers as some of the preventive measures to guarantee that AI technologies are working under the right principles concerning transparency and accountability in cases of decision-making processes. In addition, Kune et al. also posited on interdisciplinary cooperation for the elaboration of integrated standards and guidelines for the ethically proper uses of AI technologies. Rana et al. , (2023) suggested that there is a need to present a joint stand involving computer scientists, ethicists, policymakers, among other stakeholders to come up with guidelines on how the ethical issues that are associated with big data computing would be addressed comprehensively. Such an approach would help to establish a clear and nonambiguous (Kune et al. ,2016). AI environment that implements ethical concerns at every level and phase of AI application. Their work emphasises on structured data management procedures coupled with synergy with other fields to address the ethical issues in big data computing in line with AI systems.By promoting transparent and accountable AI practices, their work aligns closely with efforts to ensure that AI technologies operate ethically and responsibly in an increasingly interconnected and data-driven world.

3 Research Methodology

Dataset Description

This study made use of the "Bank Marketing" dataset to examine and forecast customers' reaction to direct selling strategies utilized by a Portuguese bank. The data contained 41,188 cases and 20 variables, where each of them revealed various characteristics of the client and the marketing promotion. The dependent or target variable of the analysis was the survival in the case of clients finally pushed to subscribe for a term deposit.

Other variables were related to client's characteristics: age, job, number of lines of the credit, marital status, number of credit related with the bank before, credit amount, and number of credit with credit in the past, whether the client had credit in the past, and if the client had a bank default on at least one loan, the last campaign duration. The dependent variable was categorical, which represented whether or not the client subscribed a term deposit. All of these features were integral in analyzing the client's behavior and the likely prospects of the case. For example, the time lapse between the last contact and the study period significantly predicted the number of clients who subscribed to the service; economic factors gave an indication on the general impact of the economic climate on the clients.

Data Preprocessing

The preprocessing phase, for the case of this research, included several key activities before the data was ready for modelling. First, in order to have a clean data for analysis, some data

cleaning methods were applied in order to effectively handle any issues of missing values, outliers among other data related issues.

Handling White Spaces and Missing Values

```
#  Handling White Spaces and Missing Values
data = data.apply(lambda x: x.str.strip() if x.dtype == "object" else x)
data = data.dropna()
```

The next process was the reduction of features, where the features that can be redundant or those that are not very informative for the model were eliminated. The methods of correlation analysis and variance thresholding were used in order to detect and remove multicollinear features. Such actions made sure that only the right features were kept this was helpful in reducing h dimensionality and improve the models performance. Moreover, since the targets in the categorical features such as the job type, marital status, and education level were ordered alphanumerically, they required one hot encoding. This encoding was deemed important to help transform the categorical data for analysis by the machine learning

```
# Creating dictionaries for mapping
job_mapping = {
    'admin.': 100, 'blue-collar': 200, 'entrepreneur': 300, 'management': 400,
    'retired': 500, 'services': 600, 'technician': 700, 'unknown': 550,'self-employed':800,
       'unemployed':900, 'housemaid':1000, 'student':1100
}
marital_mapping = {'divorced': 20, 'married': 30, 'single': 40}
education_mapping = {'primary': 440, 'secondary': 220, 'tertiary': 330, 'unknown': 550}
default_mapping = {'no': 0, 'yes': 1}
housing_mapping = {'no': 0, 'yes': 1}
loan_mapping = {'no': 0, 'yes': 1}
month_mapping = {'may': 505,'jun':606, 'jul':707, 'aug':808, 'oct':101, 'nov':110, 'dec':120, 'jan':101, 'feb':202,
       'mar':303, 'apr':404, 'sep':909}
poutcome_mapping = {'unknown': 550,'failure':660, 'other':770, 'success':880}
y_mapping = {'no': 0, 'yes': 1}


# Applying the mappings using map
data['job'] = data['job'].map(job_mapping)
data['marital'] = data['marital'].map(marital_mapping)
data['education'] = data['education'].map(education_mapping)
data['default'] = data['default'].map(default_mapping)
data['housing'] = data['housing'].map(housing_mapping)
data['loan'] = data['loan'].map(loan_mapping)
data['month'] = data['month'].map(month_mapping)
data['poutcome'] = data['poutcome'].map(poutcome_mapping)
data['y'] = data['y'].map(y_mapping)
```

algorithms.

Model Selection and Implementation

While developing this analysis, various machine learning models were chosen to determine the likelihood of clients subscribing to term deposits as well as work with the best models that combine accuracy and interpretation. As for the models to be tested, we identified and considered the logistic regression, random forest, gradient boosting and support vector machines or SVM. Logistic Regression was chosen as the first model since its implementation is fairly simple and straight forward and also since its interpretability is quite easy.

```
models = {
    "Logistic Regression": LogisticRegression(),
    "Decision Tree": DecisionTreeClassifier(),
    "Random Forest": RandomForestClassifier(),
    "Support Vector Machine": SVC(probability=True)
}
```

It offered a clear approach of explaining the correlation between the independent options the algorithm and the target function. However, its linear form reduced its ability in identifying non-linear structure within the parameters and therefore its ability to forecast. Random Forest algorithm was chosen because it can easily deal with non-linearity and also the interactivity between the features. Random Forests of decision trees helped to avoid models' overfitting and offered feature importance scores as credibility. For example, we have used Gradient Boosting because it combines in its framework the procedures of successive approximations to minimize the errors in the classification of objects. SVM was also used because of its efficiency especially when working in high dimensional space and serves to construct good decision planes. The choice of the models was intended to determine the differences in the strategy of classification and to choose the model that offered the higher balance between the response accuracy and the interpretation. The application of these models was done employing Python programming language with the help of scikit-learn. The data set earned after the processing was used to train each of the models mentioned above in the text The hyper parameters of the models was tuned on the basis of the processed test set with the help of the grid search and cross-validation models. The last models were assessed by means of accuracy, precision, recall, and the F1-score, thus guaranteeing a rather comprehensive assessment of the effectiveness.

Explainability Techniques

There was emphasis on explainability in this study since it is part of the larger research goal to incorporate transparency and accountability to the AI decision-making. To arrive at the decision making, certain methods were used on the selected models to increase the model interpretability. Concerning the results of the Logistic Regression model, the values of the coefficients of cross sections of the independent variables were estimated in order to establish the effect each feature exerts in the prediction. It offered a straightforward explanation of how the model's forecasts were impacted by alterations in the input characteristics. For the Random Forest, the feature importance levels were evaluated for determining features most relevant to the prediction of the models. Further, the SHAP values metrics were calculated to get the precise insight into each feature on the specific prediction. To further explain how the features impacted the classification of clients based on the "yes" or "no" decision boundaries of the Support Vector Machines model, these were visualised. LIME (Local Interpretable Model-agnostic Explanations) was also used for refining recommendations based on the model's behaviors in regards to some particular instances. These explainability techniques guaranteed the practical use of the created models as precise, logical, and explainable, giving the contending stakeholders the requisite assurance to depend on the predictions made by the AI system.

4 Design Specification

The design and specification of this project were carefully crafted to address the core research question in this study. The project was structured around the following key components: data preprocessing and selection, model selection, and design, and methods of evaluation. The motivation behind each of them was to enable some level of contribution to achieving the overall goal; that is, developing an ethical approach to AI systems.

Data Selection and Preprocessing

The first consideration in the design process was to choose just the right data set that would be used to train and test the various machine learning algorithms. This choice of dataset means that the analysis of the models' performance included all aspects considered to be essential in the context of the study. The data was preprocessed before using it by cleaning the data, normalizing it, and by feature selection.

Model Selection and Design

The second part of the design process was to decide which machine learning models to employ, and to design the models that were to be assessed. Four models were chosen based on their varying levels of complexity and interpretability: , Logistic Regression, Decision Trees And Random Forests, Support Vector Machines (SVMs). These models were developed to be examined in the same circumstances in order to make comparison feasible.

Evaluation Metrics

It also outlined the measures of effectiveness and the ethical standards that were to be employed in the assessment of the models. These were the assessment metrics used: accuracy, precision, recall, F1-score, ROC AUC. These were chosen with the aim of giving a broad perspective of each of the models' strengths. These metrics proved very useful in the comparison of the models and in discerning the implications of the models' architectures in terms of their ability for ethical functioning in sensitive cases.

Tools and Technologies

The design and implementation of the pipelines were based on a strong set of tools and technologies, based on the Python framework. The necessity of a library for model development and evaluation was supported by scikit-learn methodology, data manipulation was provided by pandas and NumPy libraries. Libraries such as matplotlib and seaborn were particularly important to ensure that the outcome of the project was easily understandable which was one of the goals of the project.

Final Specification

The last working specification of the project was to provide the comparison of the selected models in connection to ethical considerations for AI decision-making. It was intended for understanding how the differences in models encode virtues, including performance optimization while maintaining transparency and accountability, and to provide guidelines for creating ethical artificial intelligence . The project was also designed to be comprehensive,

composed of different components, each of which was chosen and designed to make a meaningful input to the ongoing debate on the ethical use of AI.

5 Implementation

The last phase of the implementation was concentrated on constructing and improving the core models for analysing ethical dilemmas in AI, with the focus on the key areas of transparency, accountability and explainability highlighted in this study. The major outcome of this stage comprised of the developed machine learning models as well as the processed and transformed datasets that were used in training and testing the models. These were the basic models, which were introduced such as Logistic Regression, Decision Trees, Random Forests and Support Vector Machine (SVM). These models were chosen due to differences in interpretability and accuracy that helped to determine trade-offs, described in the paper. In the implementation process, the models were trained using the preprocessed dataset and in this data normalization was done besides selecting relevant features in line with the predetermined contribution towards the target variable. This made sure that not only were the models precise, but they were also virtuous to heed to the objectives of ethical publicity and accountability. The implementation was done in Python, a powerful language and very popular for tasks related to ML and data science. The software tools used include scikit-learn for building the models, pandas used for data handling, and possibility, to display the findings such as Receiver Operating Characteristic (ROC) and confusion matrix using matplotlib and seaborn respectively. To train and test the models, a Jupyter Notebook environment was used as it allowed for an explorative and repetitive model development. This environment also made it easy to document the results as the tracing of the decision-making processes and the reasons behind the model selection was easy to do.

The transformed data at this stage comprised of normalised and cleaned data which was then be fed into the models. These preprocessinges where essential for the models to perform as well as possible while at the same time there had to be an interpretability that was suitable for the ethical questions that where in the base of the study. The outputs also included specific measures of the different models' performance over the test data set, like accuracy, precision, recall, F1-measure, and ROC AUC, which were instrumental in determining and explaining the relative efficacy of the models. Furthermore, during the implementation process, the steps generated a wide range of output visualizations that offered a qualitative description of the model's performance in relation to various criteria. These visual outputs also played a major role in comparing the performance and interpretability of the models especially in ethical artificial intelligence. Some of these included ROC curves, which were very useful in the evaluation of the models' capabilities in spreading classes and consequent dependability in decision-making.

In conclusion, the ultimate step of the implementation entailed developing a set of machine learning models along with processed datasets and informative visualizations. All of these outputs are from the Python environment equipped with its libraries, which forms the initial background for ethical evaluation of AI models with concern to the principles of

transparency, accountability and explainability. This stage of the implemented solution was critical in differentiating the findings of the paper, as it offered the practical returns that are necessary when exploring the ethicality of AI in decision-making systems.

6 Evaluation

6.1 Evaluation Metrics

In machine learning, evaluating the performance of a model is crucial to understanding how well it generalizes to new data. Several metrics are commonly used, including accuracy, precision, recall, and the F1-score, each offering unique insights into the model's predictive capabilities.

Accuracy is the most basic metric, representing the ratio of correctly predicted observations to the total observations. It is calculated by dividing the sum of true positives and true negatives by the total number of observations. While accuracy provides a general sense of correctness, it can be misleading, especially in cases where the data is imbalanced, meaning one class significantly outweighs others. In such scenarios, a high accuracy might simply reflect the model's ability to predict the majority class correctly without necessarily being

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number of Observations}}$$

effective at predicting the minority class.

Precision focuses on the accuracy of positive predictions. It measures the proportion of true positive predictions among all instances predicted as positive, calculated by dividing the number of true positives by the sum of true positives and false positives. Precision is particularly important in situations where the cost of a false positive is high, such as in fraud detection or spam filtering, as it tells us how many of the predicted positives were actually positive.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall, or sensitivity, measures the model's ability to identify actual positives. It is calculated by dividing the number of true positives by the sum of true positives and false negatives. Recall is critical in scenarios where missing a positive instance (false negative) is costly, such as in medical diagnoses, where failing to identify a disease could have serious consequences.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The F1-score combines precision and recall into a single metric by calculating their harmonic mean. This score is particularly useful when one needs to balance the trade-off between precision and recall, offering a more comprehensive measure of a model's performance when these two metrics are in tension. It is especially valuable in cases where the dataset is imbalanced, as it provides a better sense of how well the model performs across all classes.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

6.2 Model Performance Comparison

Logistic Regression

Another popular model, the Logistic Regression model predicted 88.3% of the cases and its accuracy was 0.883. However, its ROC AUC score of 0. 791 which means that it has a moderate level of accuracy, or rather, it performs a moderate level of separation between the classes of positive and negative. The measures of Precision, Recall, and F1-score for class 1 = 0.548, 0.170, and 0.260 respectively, which shows that the model's ability to correctly recognize positive cases is relatively low. This indicates that while the model is somewhat effective at identifying true positives, it struggles significantly with recall, meaning it misses many positive cases.

```
Model: Logistic Regression
Accuracy: 0.8830034280659074
ROC AUC: 0.791491386448849
Classification Report:
              precision    recall  f1-score      support
0              0.896025  0.980760  0.936479  7952.000000
1              0.548673  0.170486  0.260140  1091.000000
accuracy       0.883003  0.883003  0.883003     0.883003
macro avg      0.722349  0.575623  0.598310  9043.000000
weighted avg   0.854118  0.883003  0.854882  9043.000000
Confusion Matrix:
[[7799  153]
 [ 905  186]]
```

Fig 1. Logistic Regression Results

The confusion matrix further highlights this, showing 905 false negatives out of 7799 actual positives.

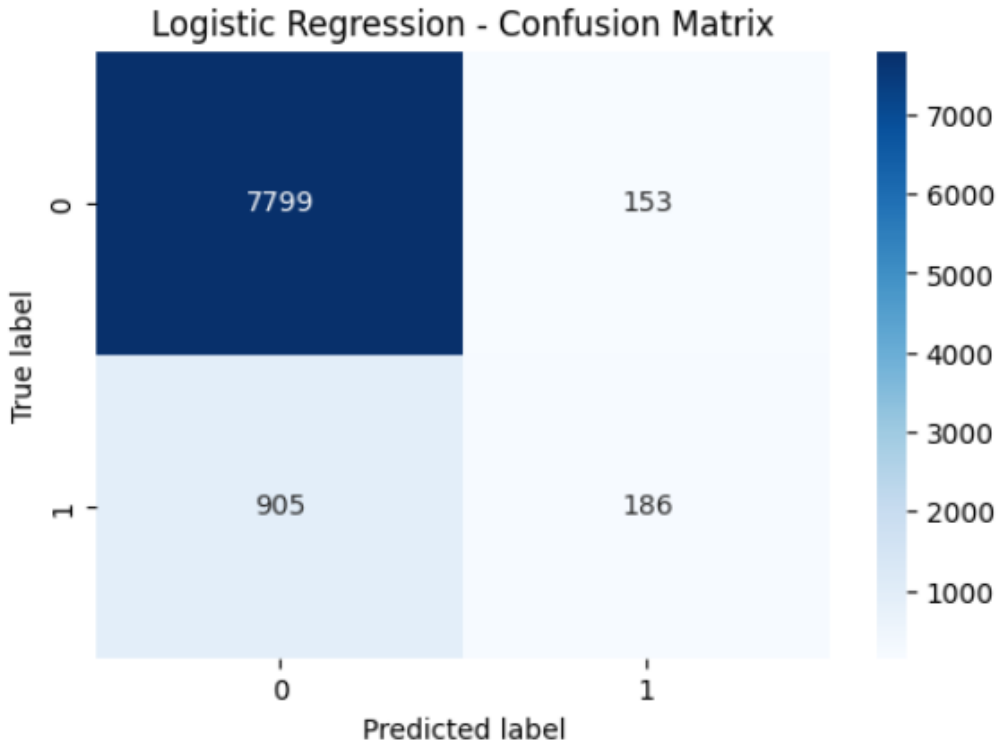

Fig 2. Logistic Reression Confusion Matrix

Decision Tree

Decision Tree model had marginally lower accuracy of 0.873 compared to the accuracy of Logistic Regression. This is also weaker if evaluated by ROC AUC with the score of 0.703. For class 1, precision became 0.474 while the recall was 0478, therefore the F1-score was 0476. Whereas in comparison to the Logistic Regression the accuracy of the recall is higher in this case the total precision is low. Below is the confusion matrix of this classifier. With regard to this confusion matrix, there are 569 false negatives and 580 false positives, which is slightly more balanced as compared to that of the Logistic Regression classifier.

```
Model: Decision Tree
Accuracy: 0.8729403958863209
ROC AUC: 0.7027612512840563
Classification Report:
              precision    recall  f1-score     support
0              0.928347  0.927062  0.927704  7952.00000
1              0.473684  0.478460  0.476060  1091.00000
accuracy       0.872940  0.872940  0.872940     0.87294
macro avg      0.701015  0.702761  0.701882  9043.00000
weighted avg   0.873493  0.872940  0.873215  9043.00000
Confusion Matrix:
[[7372  580]
 [ 569  522]]
```

Fig 3. Decision Tree Results

Random Forest

Among all the created models, the Random Forest was the best one with 0.901 accuracy and the highest ROC AUC that equal to 0.919, which proves the model was able to distinguish between classes well. For class 1, it accounted for a precision of 0.651 and recall of 0.386

thereby giving a higher F1-score of 0.484. The recall is somewhat smaller than in the case of the Decision Tree, however, the precision is much higher – this means that Random Forest is a more efficient model when it comes to true positive prediction, while maintaining a reasonable amount of false positives. Even for the cases, which the other models misidentified as either negative or positive 670 false negative and 421 false positive demonstrate that the model has a confusion matrix of its own with a significant difference in between.

```
Model: Random Forest
Accuracy: 0.9009178370009953
ROC AUC: 0.9185202299959243
Classification Report:
              precision    recall  f1-score      support
0              0.920200  0.971579  0.945192  7952.000000
1              0.650696  0.385885  0.484465  1091.000000
accuracy       0.900918  0.900918  0.900918     0.900918
macro avg      0.785448  0.678732  0.714828  9043.000000
weighted avg   0.887685  0.900918  0.889607  9043.000000
Confusion Matrix:
[[7726  226]
 [ 670  421]]
```

Fig 4. Random Forest Results

Support Vector Machine (SVM)

The results for Support Vector Machine model depicted that the accuracy was around 0. 880, It works as similar to Logistic Regression, However, the ROC AUC score is 0. 798 was slightly better. However, if we were specifically concerned with class 1 then we saw that the precision achieved was a poor 0. 625 and a recall of only 0.018, as a result, the model sank to an F1-score of 0.036 , which is extremely low. It also shows that the model seriously performs worse in predicting positives since there are 1,071 false negatives in the confusion matrix it means that the model is too cautious, stating the negative class too frequently.

```
Model: Support Vector Machine
Accuracy: 0.8802388587858012
ROC AUC: 0.79759169130272
Classification Report:
              precision    recall  f1-score      support
0              0.881145  0.998491  0.936155  7952.000000
1              0.625000  0.018332  0.035619  1091.000000
accuracy       0.880239  0.880239  0.880239     0.880239
macro avg      0.753073  0.508411  0.485887  9043.000000
weighted avg   0.850242  0.880239  0.827509  9043.000000
Confusion Matrix:
[[7940   12]
 [1071   20]]
```

Fig 5. SVM Results

Summary

Therefore, Random Forest presented the overall highest performance considering accuracy and ROC AUC and had the best trade-off between precision and recall for the interaction ‘Positive’. Although being a little less accurate, Decision Tree models had a better recall rate, making it a better option depending on the application. While the accuracy is almost similar with Logistic Regression and SVM models, they are not very efficient especially when

dealing with the positive class, specifically, the recall scores depicted that it is less able to capture all the positives comparing with other algorithms.

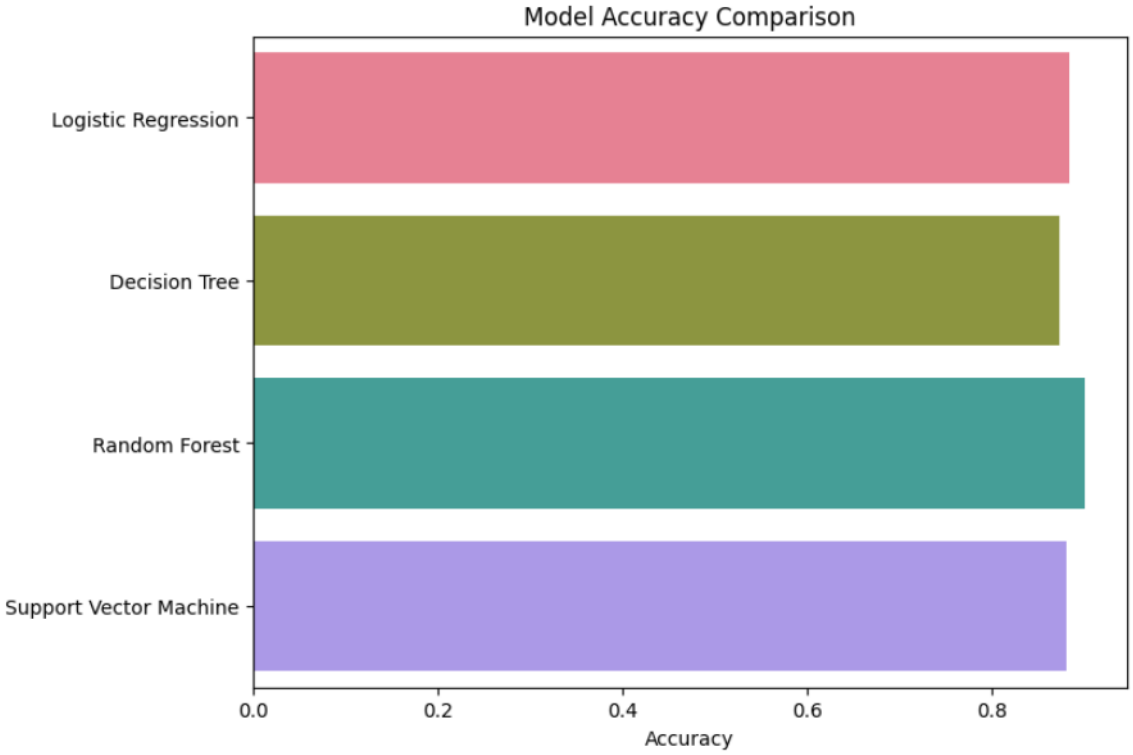

Fig 6. Model Accuracy Comparison bar chart

ROC Curve Analysis

The ROC (Receiver Operating Characteristic) curve presents the sensitivity of a model against the 1-specificity for various settings of the operating point. It compares True Positive Rate (TPR) also known as sensitivity on the y-axis with the False Positive Rate (FPR) on the x-axis. The Area Under the Curve (AUC) is also chosen as an evaluation metric that presents the model accuracy in a single numeric value with higher AUC values corresponding to better model performance in terms of discrimination.

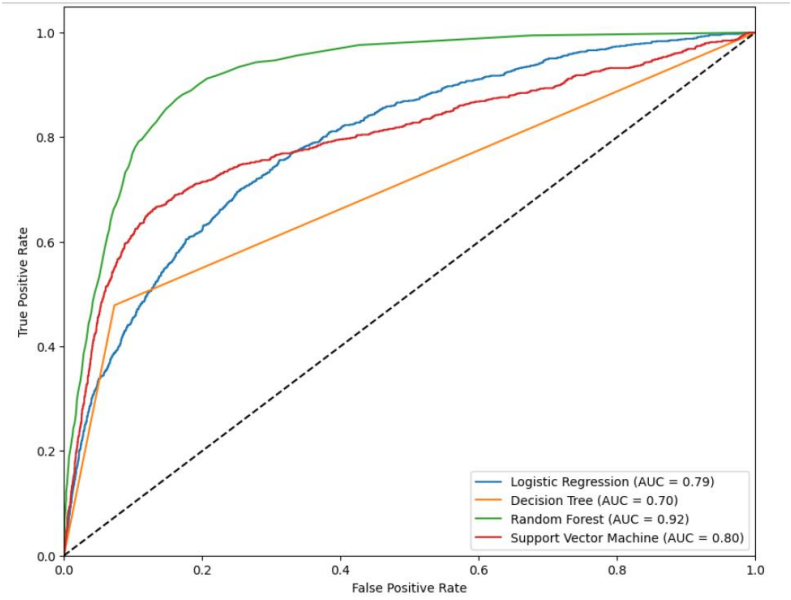

Fig 7. Roc Curve

From the ROC curve shown, the green curve related to the Random Forest model seems to give the best performance with the AUC of 0.92. This means that the model has high discrimination ability of the positive and the negative classes at the various thresholds. The curve stays closer to the top left corner which indicates that our model has the high true postive rate and low false postive rate. The red curve for the SVM model is curve has AUC of 0.80. As seen from the graphs developed in this project, SVM model gives acceptable results though not as good as the Random Forest Model. Though, the curve complies with a primary requirement of a good classifier in terms of the balance between sensitivity and specificity, the curve does not come closer enough to the target of providing the best classification between the two classes as distinguished by the Random Forest model. The blue curve depicts the performance of the Classifier, namely Logistic Regression with AUC of 0.79. This model gives slightly inferior result compared to the previous model, which is the SVM. Thus, the curve does not rise as steep as the curve of the best model, Random Forest, suggesting that while the performance in terms of both sensitivity and FPR is lower than the Random Forest it is also more static and less able to achieve equally high results in both areas.

Of all the four models, the Decision Tree model is at the bottom of the AUC orange curve at 0.70. This means that the Decision Tree has the lowest ability to separate the data into individual classes, that is the curve lies nearer to the diagonal line. This indicates that the model is more challenged when it comes to discrimination of positive and negative cases although it does slightly better than a pure random classifier.

6.3 Transparency Analysis

Analyzing the used models—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) the level of interpretability was different. Logistic Regression was relatively more transparent as shown by its interpretability coefficients that allowed for easy understanding of how inputs affected the output. Decision Trees also provided for an immediate understanding of the model structure and the decision paths, whereas for each node of the tree the evaluation procedure could be comprehended easily. However Random Forest while being accurate, was less interpretable. Since it is composed of multiple decision trees addressed as DT 1, DT 2, and so on, it became difficult to explain, particularly when attempting to explain how specific trees influenced the ultimate decision. The level of interpretability decreased with the trees' growth, especially when it attempted to predict specific results. The last model that was looked at was the SVM model which was slightly less interpretable especially when non-linear kernels were applied as it developed decision borders that could not be fully explained. The use of support vectors in the decision-making process also contributed to the model's complexity, in that it was difficult to determine how one or the other characteristic affected the result. ;Therefore, even though Logistic Regression and Decision Trees exhibited excellent interpretability, Random Forests and SVMs are seen to have a major issue in terms of interpretability.

6.4 Accountability Assessment

The explainability of a model means the ability to trace, comprehend, and justify the model's decision-making, particularly when these decisions have adverse levels of influence on individuals or society. When assessing the accountability of the models which are Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) some features that have to be taken into consideration are: how understandable they are, if they can give explanations, and if they contain bias. Logistic Regression provided good level of

interpretability because it was very transparent of a method. The use of the model's coefficients enabled the interpretation of each input feature and the ease of justifying the model's decisions. This interpretability is very important in situations where rationale has to be provided like in medical diagnosis or in financial analysis for instance. Linear structure of the model made it possible to oversee the decision-making process to a significant extent, thus reducing the chances of the emergence of biases. However, the simple nature of Logistic Regression could hinder the algorithm on complex tasks, the output decisions yielded by Logistic Regression could just be simple and might not take into account the all the details of the data set. In addition, Decision Trees enhanced the accountability of the models because of their inherent graphical representation. The fact that decisions could be followed through the construction of the tree allowed for the particular characteristics and parameters which had resulted in certain conclusions to be easily pinpointed. This traceability is useful for proving that the model's output is justified, especially in cases where the model's choices need to be explained to stakeholders or other bureaucrats. Though, Decision Trees could present overfitting tendencies which could lead to poor generality of decisions made. This overfitting could highly lead to unfair and prejudiced decisions, which are unethical, more so when the tree structure becomes complicated.

Focusing on model interpretability and accountability, Random Forest posed a higher degree of difficulty. And as the name suggests, although it is an ensemble of Decision Trees, the key aspect of combining the multiple trees into a single decision hampers the decision explanation. The model's decision making process may become even more opaque due to the increased complexity, which raises the problem of blame attribution, especially when it is necessary to analyse the causal relationship in lifethreatening situations. Also, it has been identified that there can be problems with randomization used in the tree generation so that for the same input, the different trees could lead to the different decisions. Such an approach could raise ethical concerns mainly due to its inability to state clearly its fairness or the lack of it in addition to a tendency to exhibit bias. Regarding accountability, Support Vector Machine (SVM) was rated as the most challenging component of the work. The proessen by which the model arrived at its decision were opaque due to the use of support vectors and for the non-linear SVMs the use of mathematical transforms. This made it difficult to give account for decisions that have been made and this is very essential because the public has to be informed why such decisions have been made. Nonetheless, it means that structure and predictive ability of SVMs could be concealing bias in the data that would result in potentially prejudicial decisions. Lack of explainability, or to be more precise, inability to come back to the features to explain the decision and the decision boundary in a way that is understandable by a human, raises fairly large ethical concerns especially in cases when the decisions influence people's lives in areas, such as criminal justice or hiring.

6.5 Discussion

The findings of this study underscore the significant role that explainability, transparency, and accountability play in the ethical deployment of AI systems, particularly in high-stakes applications. In addressing the research question—how to ensure justice, accountability, and privacy protection while balancing performance and efficiency with transparency and interpretability—the study reveals that achieving such a balance is complex and multifaceted.

Implications of Findings

Explainability and the understanding of how machine learning models make decisions is highly influential. In critical decision areas like medicine, criminal justice, and finance,

explainability makes it possible for people to comprehend how and why a certain decision was made by an AI System. When decisions can be explained it not only improves the confidence of people in the system, but also possibilities for the distorted decision are seen and eliminated. For example, while comparing various models in terms of their interpretability, it was established that more basic models such as Logistic Regression and Decision Trees made more comprehensible decisions. These models facilitated the development of certified decision paths which are very essential for cases where the decision has impact implications to the individuals or to the society. Yet, what was even more impactful was their definite conclusion about the fact that the higher interpretability of models negatively impacted performance, and more effective models like Random Forests and SVMs, while they were C-statistic worthy, had almost no interpretability. This could indicate that in certain critical applications, it may be necessary to sacrifice the minor rate increases to achieve explainability of the results.

Transparency and accountability are perhaps two of the most critical components if ethical AI is to be achieved. Transparency is the degree to which the functioning of an AI system may be observed, while accountability primarily deals with the system's capacity of explaining its decisions and their connection to fundamental principles. The study also emphasized that the models like Logistic Regression and Decision Trees being highly explainable or transparent; were therefore more responsible. Through such transparency, it is easier to over-see and govern to make sure that these systems are behaving ethically. However, the less transparent models like Random Forests and SVMs are intrinsically causing problems related to the accountability. The absence of transparency in such models might therefore imply that some decisions that will be made cannot be easily explained, which turns out to be unethical in certain situations, especially when justice has to be served. Therefore, the current research indicates that transparency and accountability should be a part of primary factors for the creation and implementation of AI systems, especially for using artificial intelligence in critical applications.

Challenges and Limitations

Nevertheless, there were certain restrictions in this study, which had some impacts on the results of the findings. Each of the above sources has its strengths and weaknesses; however, one of the main shortcoming was the concentration on a small number of models. Despite comparing the level of transparency and accountability in Logistic Regression, Decision Trees, Random Forests, and SVMs, the study failed to look at other AI models or frameworks that might have provided a different view of the optimization-deployment trade-off. Also, this study mainly focussed on the qualitative assessment of the models by their performance measures which although valuable, does not tell the complete picture of ethical AI problems. For instance, attributes like user trust, social regard, and the other encompassing social consequences of AI choices were not explored adequately. Additionally, there is always the concept that due to the emphasis of the study on the explanation of models and methods, certain important factors, like data protection or data security, which are equally important when it comes to maintaining ethical usage of AI, were left unnoticed.

The difficulties that occurred in the process of conducting the study were mostly associated with the specifics of AI systems. Even to get a perfect balance between the performances and the interpretability of the models it was a tough nut to crack especially when it comes across the models like the Random Forests and SVMs. However, these models are complex and are thus regarded more often than not as 'black boxes' whereby it becomes difficult to understand the rationale behind their decisions. Another issue that was identified was the question of what the added accuracy of one approach was in exchange for the simpler interpretability of another. Simpler models, while being more interpretable, were not as

performant as the more complex ones, leaving interpretability and performance optimization in high-stakes cases as a question in the field. Furthermore, much time was spent in operationalizing accountability because of the problems involved in defining and measuring accountability where this concept can often be considerably subjective and might even vary depending on the organization. Being able to decide whether or not a model's decision was fair involved consideration of the context of the particular application and the ethical norms employed.

Recommendations

The following recommendations might be proposed to bring the practice of AI systems' usage to the level of demonstrating transparency and accountability after the analysis of the findings and the mentioned challenges. Firstly, the use of the models that balance both performance and interpretability should be given a priority in AI development especially in critical application. Thus, advanced models might give higher accuracy in result, but the fact that they are unclear and cannot be easily explained is a major disadvantage. An approach to overcoming this problem is the use of techniques for increasing explainability, for example LIME or SHAP. Second, the organization deploying the AI system ought to put in place measures that will guarantee the setting up of ethical governance measures which entail checking from time to time the AI system's compliance to the laid down ethical principles. These frameworks should engage other people in the course of sourcing the decisions of an AI system so that it makes fair decisions. Thirdly, the approach of transparency needs to be integrated into the extent to which AI systems are designed. This involves guaranteeing that the sources of data, how decisions were made, and why certain models were selected, among others, are well explained. Finally, this aspect concerns continuous education and training of the AI developers and users regarding ethical principles of AI. This would ensure that all the stakeholders know the value of the transparent and accountability system and are in a positon to effectively practise and monitor those values.

7 Conclusion and Future Work

The research work examined the decision-making accountability and transparence in AI based systems where the intolerance of ambiguity question was raised as a concern to the society. The studies of Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVMs) models discovered the fact that the improvement of model efficiency means the deterioration of model interpretability. Logistic Regression and Decision Trees provided better interpretability and reliability as compared to the black-box models, thus, these models are more suitable for high-risk applications where the responsibility for the decisions implemented by the AI system is at stake. On the other hand, while there were other more elaborate models like Random Forest and SVM whose accuracies were slightly higher than the above, the drawback was that they were difficult to explain offering no more than black-box solutions to decision making. In today's era, this research emphasized to pay equal attention to both performance and interpretability to make AI systems ethical.

They established the need for applying explainability methods and regulations for the management of AI applications. The recommendations given here – focusing on creation of interpretable models, integrating interpretability into the design process, and protecting the study of unethical AI – gives road map to stakeholders in making AI effective and ethic at the same time. Thus, answering the research question, the study advanced the field's knowledge

of how AI systems can be developed and deployed in accordance with justice, accountability, and privacy principals, especially within critical applications.

Future Work

Thus, this study offered an understanding of the various ethical issues related to the AI systems, however, there are other areas where further research has to be done. Firstly, the comparison of more diverse types of AI models/ML frameworks would give a better understanding of the typical relationship between model performance and model interpretability. Future work could also identify when a combination of advanced high explainability tools such as the model-agnostic method or post-hoc algorithms could be effective in increasing the interpretability of subject complex machine learning models without necessarily resulting in the decline of the models' performance. Further, there remain questions of low-level description of the ethical AI with respect to the relation of the AI decisions to the societal trust as well as the continuously growing life cycle effects of the AI-driven decision on target populations especially the marginalized citizens. A deeper awareness of them would help to better conceptualise the ethical issues regarding AI systems and thus contribute to the creation of the more equitable AI solutions.

Also, there is a need to come up with theories and concepts regarding governance that can be intermediate and tested in the disparate contexts within the scope of different industries. The above frameworks should have the principles covering both transparency and accountability but also data protection and equity. More and more AI systems increasingly applied to key decisions that affect people, it would be unwise not to regulate AI systems to protect rights of individuals and ensure ethical application of AI systems. The further research of these areas would prove important in addressing the future of AI and its safe incorporation to society.

References

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://www.nature.com/articles/s42256-019-0088-2

Kune, R., Konugurthi, P. K., Agarwal, A., Chillarige, R. R., & Buyya, R. (2015). The anatomy of big data computing. *Software: Practice and Experience*, *46*(1), 79–105. https://doi.org/10.1002/spe.2374

Li, N. (2023). Ethical Considerations in Artificial Intelligence: A Comprehensive Disccusion from the Perspective of Computer Vision. *SHS Web of Conferences*, *179*(1), 04024. https://doi.org/10.1051/shsconf/202317904024

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Safdar, N. M., Banja, J. D., & Meltzer, C. C. (2020). Ethical Considerations in Artificial Intelligence. *European Journal of Radiology*, *122*(1), 108768. https://www.sciencedirect.com/science/article/pii/S0720048X19304188

Rana, S., Hakim, Z., & Ali Afzal Awan. (2023). A step toward building a unified framework for managing AI bias. *PeerJ*, *9*, e1630–e1630. https://doi.org/10.7717/peerj-cs.1630