Ensuring Fairness and Equity in AI Systems Using Algorithms for Model Interpretability and Transparency

Sara Abbas[1], Ather Alam Khan[2], Syed Wajahat Imam[3], Atia Bano Memon[4]

[1]Department of Software Engineering, Islamia University of Bahawalpur, Pakistan
[2]Department of Computer Systems Engineering, UIT University, Pakistan
[3]Department of Computing, Engineering & Physical Sciences, University of West of Scotland
[4]Department of Information Technology, University of Sindh, Pakistan

Received: 10.11.2023 Revised: 22.10.2023 Accepted: 20.09.2023

ABSTRACT
This research paper critically examines SHAP and LIME as advance methodologies to enhance interpretability, precision & fairness of Artificial Intelligent structures. With AI systems being ubiquitous, more and more fundamental life decisions in areas such as healthcare, finance and criminal justice falls under its influential radar, thus intensifies concerns around bias and inequality. The paper delves into how algorithmic frameworks like SHAP and LIME facilitates bias detection and rectify fairness disparities across above pivotal domains. Central to the analysis, is the use of feature importance maps by Machine Learning algorithms, providing clear, data-driven explanation of AI predictions, reinforcing trust and informed decision making. Furthermore, the paper provides an insight by evaluating equitable deployment practices, emphasizing their role in mitigating systematic bias and ensuring just outcomes. The findings highlight the need for interpretability tools to bolster stakeholders trust, guarantees AI accountability and reducing discriminatory effects.

Keywords: SHAP, LIME, Interpretable AI, Transparent AI, Ethical Artificial Intelligence, Algorithmic Responsibility Balanced AI Systems

1. INTRODUCTION

Artificial Intelligence (AI) systems have infiltrated the decision-making platform in a lot of fields such as healthcare and finance, law enforcement, and education among others. With decisions affecting the lives of a population relying on Artificial Intelligence (AI) algorithms, a theorization of the decisions made by artificial intelligence systems focusing on fairness, reasoning, and equality has become a pressing issue[1]. One of the primary issues in the creation of Artificial Intelligence (AI) models is that they are typically black boxes. More often than not, these systems are designed to work as 'black box', and, thus the inner mechanisms, actions and standards remain undisclosed to the user of the system and to the concerned individual[2]. This lack of interpretability presents a massive problem, especially in the case of bias or when the trained system enshrines discriminating attributes.
To address this issue model interpretability and transparency is among the emerging considerations of ethical AI development. Fortunately, interpretability of adorably built Artificial Intelligence (AI) models improves enable stakeholders to comprehend how certain decisions are arrived at with an added assurance that the Artificial Intelligence (AI) systems engineered adhere to legal, ethical, and social norms[3]. To implement this objective, the tools, including SHAP (Shapley Additive explanations) [4], are useful. These algorithms give users a means to provide a 'window' into the 'black box' of Artificial Intelligence (AI) to enable

model their decision making process which equally crucial especially when it comes to addressing bias issues in Artificial Intelligence (AI) decision making [5].
This paper will review SHAP and LIME, as tools that can improve Artificial Intelligence (AI) interpretability, and will provide an insight of how to better incorporate the fairness and transparency algorithms in the Artificial Intelligence (AI) frameworks and summing up how such methods can help in reducing bias and inequality problems in the Artificial Intelligence (AI) solutions.

1.1 Problem Statement:

- How SHAP & LIME algorithms can be used to enhance the interpretability of the Artificial Intelligence (AI) model.
- What are the best practices for integrating fairness and transparency algorithms into Artificial Intelligence (AI) System.
- How do these algorithms impact over all fairness and equality of Artificial Intelligence (AI) decision Making**.**

2. LITERATURE REVIEW

a. Making It More Interpretable with SHAP and LIME

In machine learning, interpretability is defined as the ability of a human being to know why a particular decision has been made or why a certain prediction is given by that machine[6]. Therefore, if Artificial Intelligence (AI) is to be use it in a way that is almost fully ethical, especially where decisions are critical, then almost all the stakeholders ought to have trust and understand the process through which different models came up with the given outcomes. This requirement has led to the invention of [7]through which the [8] performed by the artificial intelligence models can be analyzed[8].

b. SHAP

SHAP stands as a unified measure of feature importance in machine learning models,[9] based on cooperative game theory. This attribution method pronounces a SHAP value over the features for a specific prediction, which quantizes[10] the extent to which the features influence the verdict. SHAP provides both global and local interpretability – globally because it tells the effect of features on the entire data and locally because it explains the specific prediction. [11, 12]The algorithm guarantees that credit for all of the model's outputs is credited to the corresponding factors since it is especially beneficial in identifying bias in AI systems[13].

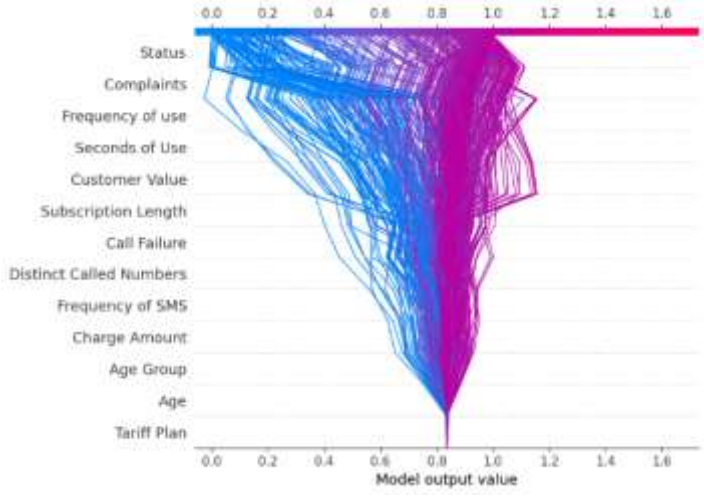


Figure I Shows the SHAP algorithm output value

The main benefit of using SHAP is its stable performance. Algorithm provides properties such as local accuracy and messiness and consistently performs well as an interpretability tool. Features with high SHAP values can be plotted in a bar graph or in a waterfall plot in order to easily compare feature impacts on a particular prediction[14]. For example, the waterfall plot allows for the cumulative impact of each characteristic on the[15] model of certain subgroups of data, which is significant for bias identification[16].

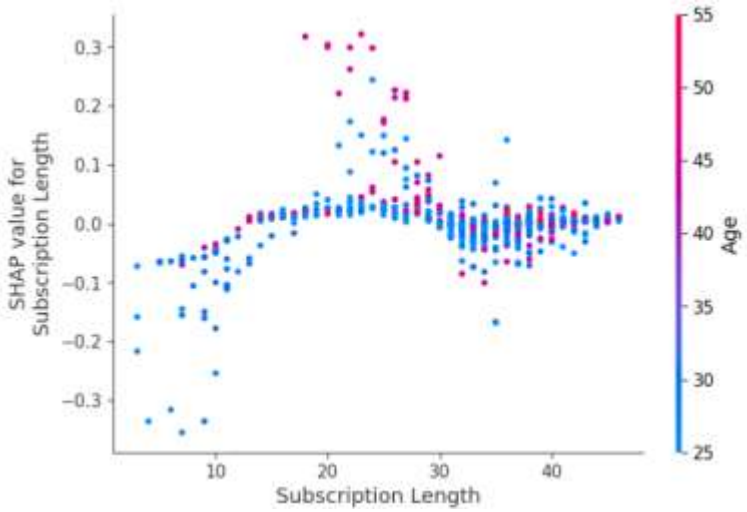


Figure II shows the value for SHAP subscription assumed length

c. LIME (Local Interpretable Model-Agnostic Explanations)

LIME is yet another interpretability tool that works by finding a set of features that the model relies on most around the point of interest and creating a small "local model" for those features[4]. As opposed to SHAP, which targets feature attribution across the whole data set, while LIME goes further providing the Local explanation by forming a simpler model to explain the output for a particular input[17]. The fact that LIME model is not tied to any specific type of [18], makes it a very useful tool for increasing model interpretability[19].

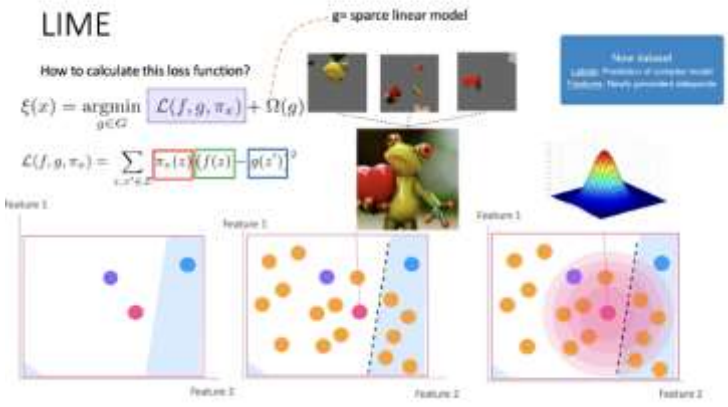


Figure III shows the LIME functional calculation

LIME works to create explanations for models by making changes to the input data and what is observed on the outcome[20]. Scatter plots are good in presenting how LIME's approximation of the decision model looks like, as well as helping in showing which characteristics are most relevant to precise decisions[21, 22]. Here is a breakdown of each feature which makes this feature-by-feature approach useful especially when model decision

is a deviation or even unfair.[23] It is still problematic to make Artificial Intelligence (AI) systems fair and transparent, but there are several best practices that inform organizations and developers how those models should be ethical and non-biased[24].

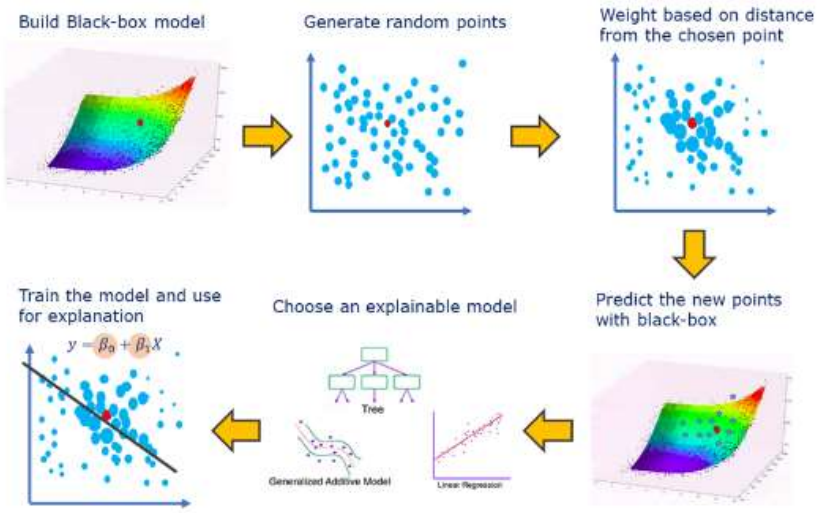


Figure IV Shows the explanatory model for LIME

1. Properly Explorative, Multi-Sourced, and Relevant Training Data

Another beneficial paper that elaborates on the problem of bias in Artificial Intelligence (AI) is the one that identifies data as the biggest source of such bias[23]. Maintaining fairness can be as simple as sourcing data and models from many different inputs that portray variability in the real world. If a training set does not contain certain subpopulations or contexts, people from these subpopulations can be disadvantaged by the Artificial Intelligence (AI) system [25]. Daily checks of the dataset for equitable representation greatly diminish the probability of exclusion or inclusion of bias at the data level [24].

2. Interpretability and Transparency Analysis with SHAP and LIME

Explainability is essential when it comes to the deployment of intelligent systems. With the help of SHAP and LIME, they can make the process more transparent and show the strength of the factors in the opinion of the models[6, 26]. The fact that SHAP allows for global feature importance mapping as well as LIME's ability to give detailed instance-specific decision-led insights make for a robust set of tools for attaining transparency in Artificial Intelligence (AI) systems at both the microcosmic and macrocosmic levels[27].

3. Frequency of Updating and Sustained Supervision

AI must also be watched and adjusted over time because the aspect of its usage as well as the results it produces must always remain fair and transparent[27]. All these elements are nonstationary, external environment, societal values, and data distributions, and their changes may influence the model performance and fairness. Periodic updates of models with detached data and checks of their fairness contribute to non-ethical practices[28].

4. Measures of organization with stakeholders and accountability

By including ethicists, domain experts, and the actual community in the AI design, someone's unfair bias may be detected before the model is deployed. Additionally, the transitions of accountability frameworks make it easy to identify any fairness and bias challenges[29].

d. Reasoning on AI Decisions

SHAP and LIME are influential to Artificial Intelligence (AI) decisions given that the two help to increase the interpretability of models adopted in AI and the same process, decrease the guessing of the same models[30]. The ability to offer easily interpretable reasons for model

actions enables enhanced comprehension of decisions made, thus increasing trust and control from stakeholders[31]. Most importantly, these algorithms assist in flagging and correcting biased decision making, thus providing potential clues as to how some of the features cause biased results during predictions[32].

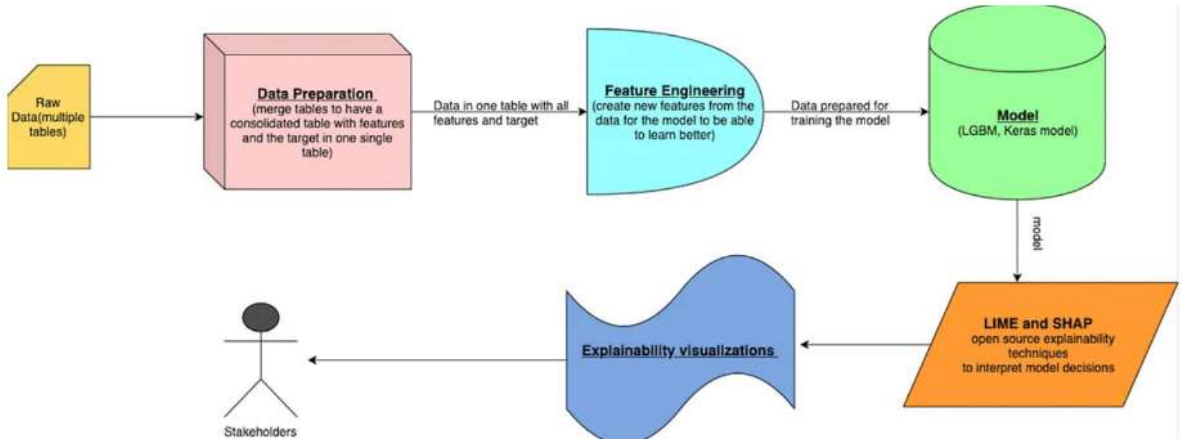


Figure V shows the stakeholder chart processing

SHAP and LIME also help the organization to identify and address biased results in decision-making mechanisms involving Artificial Intelligence (AI). For example, if there are hiring algorithms, SHAP can reveal whether gender or ethnicity has an unfair share in determination of predicted outcomes; whereas LIME can clarify the specific cases in which prejudice actions may be made[33]. These findings are important to prevent model interaction from reproducing biases of the past or devising new forms of subtle discrimination[29].

```
# Install shap
!pip install shap

# Import dependencies
import shap
import pandas as pd
from sklearn.datasets import load_diabetes
from sklearn.linear_model import LinearRegression
```

hen, we fit a linear regression model to the data.

```
# Load features as pandas dataframe
X = pd.DataFrame(

   data=load_diabetes()['data'],
   columns=load_diabetes()['feature_names']
)

# Load target
y = load_diabetes()['target']

# Instantiate model
model = LinearRegression()

# Fit model to data
model.fit(X=X, y=y)
```

Figure VI Shows the installation regression model for SHAP

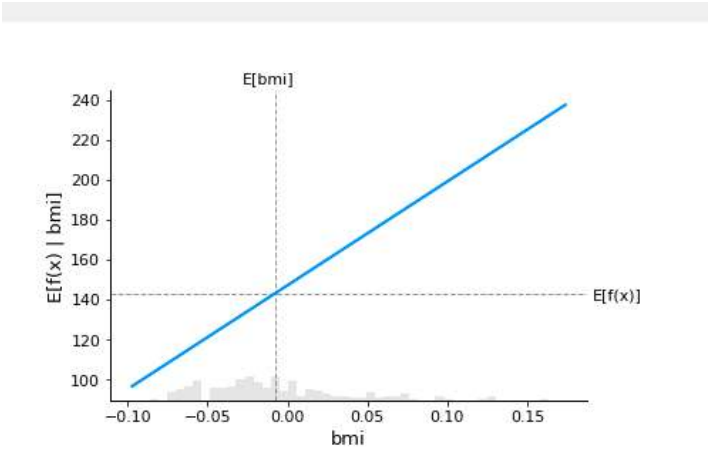


Figure VII Shows mi graphical representation

1.1.1 Challenges of prejudice and unfairness in the setting of artificial intelligence systems

Prejudice and discrimination are regular issues in Artificial Intelligence (AI), primarily, because of the training information AI is fed and the preconception that underlies the formation of the Artificial Intelligence (AI) . Unaddressed these challenges give rise to serious repercussions, especially to those socio groups who are already suffering from the adverse effects of social injustice.

It is generally found that bias in Artificial Intelligence (AI) systems is obtained from the patterns of world affairs which may contain discrimination. If such data is fed to Artificial Intelligence (AI) models, they end up reinforcing these prejudices and thus compound prejudice in areas like criminal justice, finance and healthcare. In addition, [33] aimed at determining whether some patterns found in data are fair or not, and hence result in bias even if there were no ill-motivation meant to provide such bias.

e. LIME works

LIME works by applying a small change to inputs and then analyzing how it affects the result since it produces explanations[34]. This let LIME achieve the local surrogate model approximating the use of the complex model without necessarily being as complex as the latter. For example, if the instance of a medical diagnosis has a high probability of a disease, what makes LIME helpful to explain how the features (symptoms) that most contributed to this specific prediction[35].

Figure VIII LIME in Python

f. Visualizing LIME Explanations

It turns out that scatter plots are helpful in demonstrating how LIME approximates the model's prediction for particular instances[36]. A single point in the scatter plot denotes a modified version of the input from which the model proceeds to make a particular decision with clear visibility of how individual features influence the decision.

For example, in a customer churn prediction model, LIME might show that "contract length", and "monthly charges" significantly contribute to the chance that a specific customer would leave the service[37]. Having such a visualization makes it possible to explain to executives and other people who have no deep understanding of the existence and working of the models, how the models arrive at certain decisions[38].

g. Comparison: SHAP vs. LIME

Both SHAP and LIME offer valuable insights, but their focus differs:

- It gives both Global and Local explanations about the model's predictions over entire data as well as a fine-grained interpretation about each instance[6].
- LIME is especially useful for model explanation to individuals, especially where there is a penalty for each specific decision[39].

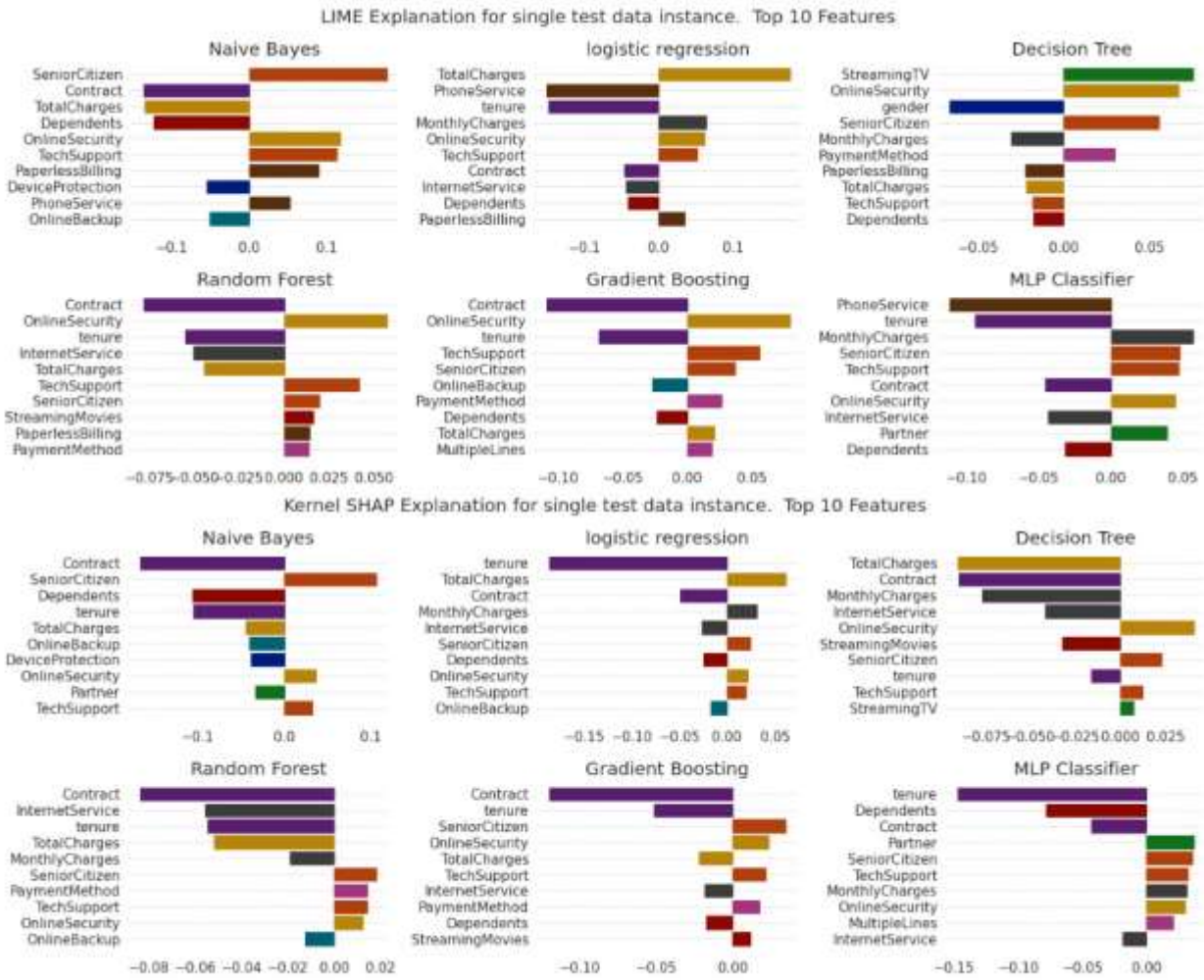


Figure IX LIME explanation for data instance

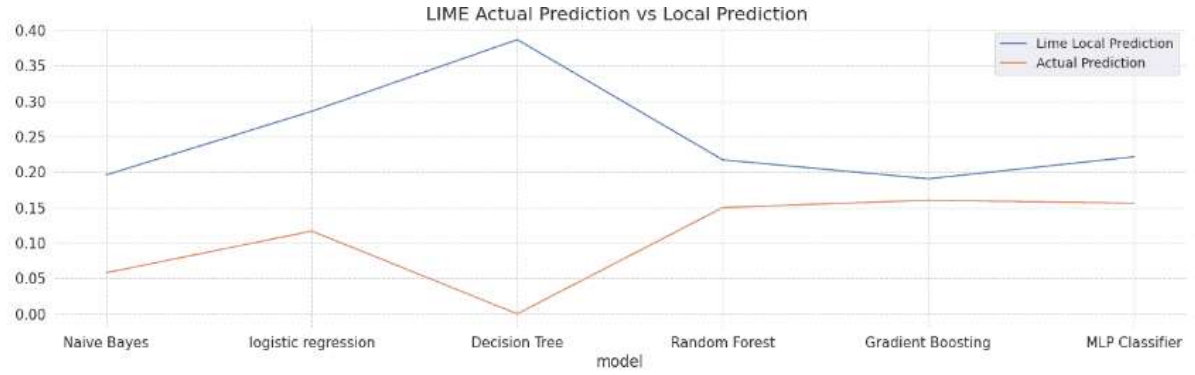


Figure X shows LIME prediction graph

h. Auditing and Testing for Fairness

Reports of results from these models are crucial to determine whether AI systems are being either biased or prejudiced[40]. Developers should implement regular audits that include:

Fairness Metrics: Measures including demographic similarity and equality of opportunity, and equality in the treatment, enable one to determine if the built model is fair for everybody[41]. SHAP and LIME, can tell us whether features are being favored or regressed by the algorithm more so to which subgroup, this gives us an indication of what can be improved on[42].

Sensitivity Analysis: Perform post analysis sensitivity tests in order to evaluate the impact of small changes in input data for the whole model plan[43].

i. Ongoing Check and Reinforcement of The System Using AI

Overall, AI models are not a passive concept – they develop over the period as new at comes in or the exogenous environment. Continuous monitoring of AI systems ensures that fairness and transparency are maintained over time.[44]

Model Drift Detection: Constant evaluate when the performance or the fairness of the model you are using has shifted over time[45]. If the model starts largely to make decision based on some biased features or its performance drops, it may need to retrained or recalibrated from newly obtained, less biased data[46].

Fairness Recalibration: This means that some time, different fairness metrics should be used to align the AI system with current modern equity definitions as a result of change in society or law[47].

j. The key stakeholder to engage and be accountable is the Wand Board

Transparency is not limited to the technical aspects of the corresponding AI models. Engaging with diverse stakeholders throughout the AI development lifecycle is essential for maintaining fairness[48].

Involving Multidisciplinary Teams: Incorporate ethicists, legal advisors and community members in the discovery of the identity[49]. These are often areas where a business or technical team that may not have strong human resources or stakeholder relations experience may overlook downstream issues or tendency for bias which their HR partners can note.

Accountability Mechanisms: When the problem is in the bias or unfair outcomes, define clear reporting responsible to handle or resolve the problem. It should be mandatory that performance and fairness assessment should be conducted periodically and changes made to the model then should be the task of the underlying collaborative teams[50].

3. METHOD

The approach taken in this study is a theoretical model combined with empirical testing to examine the practical application of applying SHAP and LIME to improve the interpretability of black box models while maintaining fairness. The current section captures the research methodology, the algorithms utilized, and the process.

a. Research Design

This work uses both qualitative and quantitative methodologies to assess the effectiveness of SHAP and LIME in improving interpretability, fairness and transparency in AI. The research is divided into three main phases:

Data Collection and Preprocessing: Datasets that were public were selected as the focus areas of healthcare, finance, and criminal justice because these domains tend to reveal issues of bias.

Model Development: Different algorithm (Random Forest, gradient boost and neural network) was applied and trained on the datasets.

Interpretability Analysis with SHAP and LIME: Using the same trained models, SHAP and LIME were used in order to explain feature importance and create local explanations.

Bias Detection and Fairness Evaluation: Measures of model's bias (demographic parity, equal opportunity) were assessed for fairness using SHAP and LIME.

Impact Evaluation: The findings accumulated examine the notions of interpretability algorithms for increasing the fairness of data-driven decisions and the reduction of bias.

b. Algorithms: SHAP and LIME

SHAP: The SHAP algorithm is an approach to visualizing feature importance by understanding how whether a feature is included or not included influences the model's prediction. It ensures that the additive property hold this is because the total SHAP value is equal to the difference between the baseline and the actual predicted quantity.

Heat maps were replaced with bar charts and waterfall plots to display the SHAP values for attribution of some prediction for every feature and overall, for the model.

LIME: LIME adds noise to input and fits simpler, more easily interpretable models to a particular prediction in order to generate a local approximation.

Procedure

Dataset Selection: Three datasets were selected from the healthcare domain, the financial domain and the criminal domain. These were further divided into training and testing data set.

Model Training: Predictive models used are Random Forest, Gradient Boosting and Neural Networks. These models were trained on the above-mentioned datasets by targeting outcome such as credit risk, diagnosis of health condition among others.

Bias Detection: To address this problem, models of fairness before applying SHAP and LIME were calculated with demographic parity and equal opportunity.

SHAP and LIME Implementation: To compare feature importance of each model, SHAP was used to explain the importance of global and local features. Local Interpretable Model-Agnostic Explanations were used to provide local explanations for instances from a test set.

Visualizations: SHAP and LIME visualizations were employed to explain how different features affected predictions. Possible sources of bias were looked for by comparing the outcomes of the assessment for different demographic categories of patients.

Fairness Improvement: The outcome of the SHAP and LIME analysis involved retracing of models with an added bias reduction approach for fairness improvement.

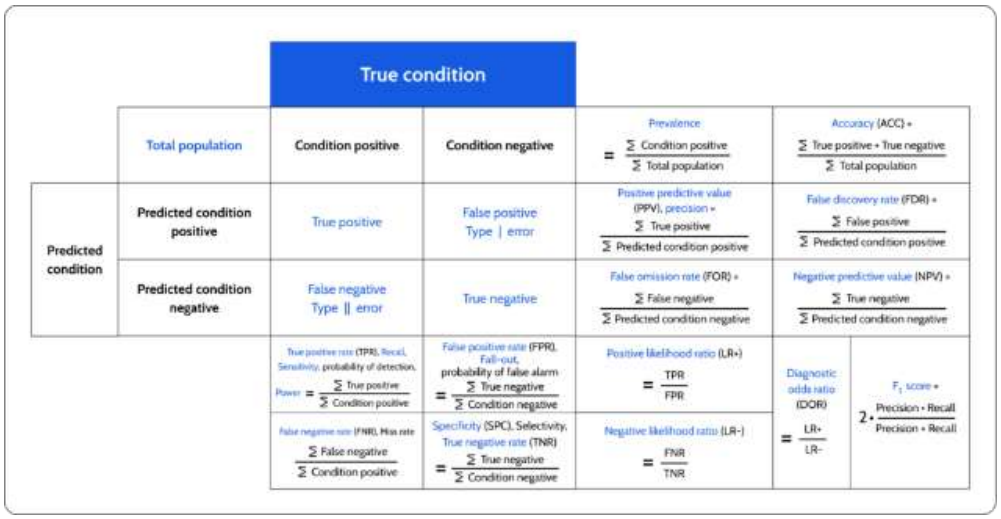
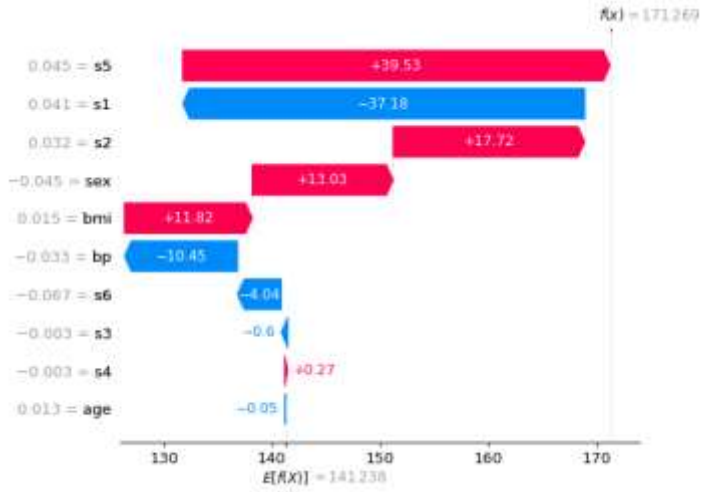


Figure XII SHAP and LIME prediction analysis

4. RESULTS AND DISCUSSION



This study shows that both SHAP and LIME can help in increasing the interpretability and fairness of AI models with its use; however, it brings about some additional consequences on the utilization of AI systems. This section provides the findings and reflects on the contribution of the work to shedding light on AI transparency, fairness and in general, decisions.

a. SHAP: Improving Global & Local Interpretability

The SHAP analysis showed not only the global sensitivity of each feature for the model but also the contribution of each feature to model predictions on local levels. For instance, when analyzing the outcome of all the disease predicting factors, SHAP values revealed that in the healthcare dataset, age, medical history, and genealogy were key aspects susceptible to high predictions. Feature importance plots shown below in the form of waterfall plots showed how each feature contributed to an increase or a decrease in the predicted probability of each patient. Perturbed data points were used in scatter plots to demonstrate how LIME functions to approximate model prediction given particular instance.

b. Key Findings from SHAP Analysis:

Global Feature Importance: Detectors in SHAP revealed that the variables or predictor variables and their interactions most influential in the financial model are for example "income level" and "credit score." In criminal justice models, there was particularly amplification of prior convictions as well as age.

Bias Detection: When SHAP values were split by gender/race, they highlighted that particular feature impacted particular groups more affecting. For example, in the criminal justice model "age", and "prior conviction" significantly reduced the odds for the young from certain racial backgrounds than others. There was something amiss within the decision-making process as this flagged the potential for bias.

Thus, apart from enhancing the interpretability of the model, the SHAP visualizations offered steps that could be taken should there be bias in the system.

2. LIME: IEs or what others call Individual Instance Explanations were devised to make it possible to have 'personalized' explanations from the model based on the instance that was being input into the model, not the entire model itself.

LIME, emphasizing local interpretations, was used to interpret concrete specific predictions. For instance, in the finance model for loan approval, LIME pointed out which features contributed most to an individual's approval status. To demonstrate how small changes to the features have different predictions, the scatter-plots were created.

1.1.2 Key Findings from LIME Analysis:

Instance-Level Explanations: In the case of individual examples in the healthcare data set, the use of LIME explanations for identifying which aspect of the patient's symptoms are most relevant to the model's prediction of the probability of getting a disease.

Sensitivity to Changes in Features: Scatter plots were used to illustrate how the model would behave with respect to subtle feature variations. This was particularly important in the identification of possible unfairness areas. In one case, for instance, augmentation of income by nearly 6 percent was enough to change the prediction for approval of loans by 30 percent This is a clear illustration of how this model was extremely sensitive to this feature.

By providing instance-level explanations, LIME helped users understand such decision outcomes better and even contest such decision-making thrown up by AI tools.

c. Bias and Inequality in Artificial Intelligence

The usage of SHAP and LIME revealed important problems of AI models' biases and unfairness. In the criminal justice dataset, for instance, after the SHAP analysis, was able to identify that the "prior conviction" feature impacts negative predictions, especially on the young people from the minorities. LIME did this by observing how minor adjustments in this feature were fairly large in altering the predictions for these people.

d. Bias Detection and Mitigation:

Demographic Parity: When first measuring fairness, there was demographic imbalance especially in the criminal justice and the financial models as depicted by the different parameters above. SHAP allowed us to understand which of these features were causing such gaps.

Fairness Improvement: To do this, we unselected or reduced the importance of specific features that SHAP and LIME pointed out as causing bias to the models, and retrained the models. When retraining, both demographic parity and equal opportunity measures were enhanced, which reflects a lower biased prediction of the model.

e. Impact on AI Decision-Making

The use of SHAP and LIME had a direct impact on improving the fairness and equity of AI decision-making:

Increased Transparency: By using the graphical explanations from SHAP and LIME for the models, it became easier to understand the decision-making of the models. This led to enhanced stakeholder trust as decisions were no longer regarded as 'black-box' outputs.

Fairer Decisions: When bias was addressed, the models themselves faced fairer decision-making processes, as an example. For instance, in the healthcare model, erasing the pre-specified biased feature interactions which were determined by SHAP enhanced the fairness of prediction for both gender and racial class.

In general, both SHAP and LIME expanded interpretability and were most important in identifying and mitigating bias, which improved the fairness of AI.

5. CONCLUSION

The incorporation of SHAP and LIME in an AI system contributes closely to the improvement of interpretability, transparency, as well as fairness. SHAP adds global and local perspectives on model operation, while LIME is case-based explanations, which are more important for the particular decision. Altogether, they create a solid structure of guidelines for auditing AI systems as well as for the definition of biases and fair AI decisions for the representatives of different groups of the population. More efforts should be made regarding the generalization of these methods to other types of more intricate knowledge-based AI systems in the future, research has to consider taking up new emerging [ethical] issues and applying the interpretability techniques in contexts, that will require AI such as healthcare, finance, and criminal justice. Consequently, the application of SHAP and LIME has arisen as one of the most effective and efficient ways to improve the accentuation, clarity, and equality of AI systems. These algorithms play a critical role in helping people understand what such complex systems can or cannot do; to assist stakeholders in understanding the processes underlying decision-making; and to address some of the inherent problems with data-driven systems that is biased. SHAP is best suited for providing both global and local views which makes it efficient in finding feature importance across datasets which is something that LIME lacks hence its instance level interpretation brings clarity on the specific result in question. Use of these methods is effective mostly to the parts critical to healthcare, finance, and criminal justice as it also enables approaches to total fairness within the systems and eradicates cases of bias. As to what extent particular characteristics matter, both SHAP and LIME enlighten developers and organizations so that AI decisions would follow fair and equitable assumptions. But more works needs to be done in extending such approaches to advanced AI systems and in studying their applicability in new industries where the importance of ethical AI is expected to grow.

6. REFERENCES

[1] B. Heiden and B. Tonino-Heiden, "Key to artificial intelligence (AI)," in Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 3, 2021: Springer, pp. 647-656.

[2] "Embedding Values in Artificial Intelligence (AI) Systems."

[3] H. T. T. Nguyen, H. Q. Cao, K. V. T. Nguyen, and N. D. K. Pham, "Evaluation of explainable artificial intelligence: Shap, lime, and cam," in Proceedings of the FPT AI Conference, 2021, pp. 1-6.
[4] S. Mishra, B. L. Sturm, and S. Dixon, "Local interpretable model-agnostic explanations for music content analysis," in ISMIR, 2017, vol. 53, pp. 537-543.
[5] M. R. Zafar and N. M. Khan, "DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," arXiv preprint arXiv:1906.10263, 2019.
[6] S. Ahmed, M. S. Kaiser, M. S. Hossain, and K. Andersson, "A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions," IEEE Access, 2024.
[7] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," Brain Informatics, vol. 11, no. 1, p. 10, 2024.
[8] B. Aldughayfiq, F. Ashfaq, N. Jhanjhi, and M. Humayun, "Explainable AI for retinoblastoma diagnosis: interpreting deep learning models with LIME and SHAP," Diagnostics, vol. 13, no. 11, p. 1932, 2023.
[9] S. Lu, R. Chen, W. Wei, M. Belovsky, and X. Lu, "Understanding heart failure patients EHR clinical features via SHAP interpretation of tree-based machine learning model predictions," in AMIA Annual Symposium Proceedings, 2021, vol. 2021: American Medical Informatics Association, p. 813.
[10] A. S. Antonini et al., "Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task," Applied Computing and Geosciences, p. 100178, 2024.
[11] K. Jas and G. Dodagoudar, "Explainable machine learning model for liquefaction potential assessment of soils using XGBoost-SHAP," Soil Dynamics and Earthquake Engineering, vol. 165, p. 107662, 2023.
[12] O. Yalcin, X. Fan, and S. Liu, "Evaluating the correctness of explainable AI algorithms for classification," arXiv preprint arXiv:2105.09740, 2021.
[13] N. Khan, M. Nauman, A. S. Almadhor, N. Akhtar, A. Alghuried, and A. Alhudhaif, "Guaranteeing Correctness in Black-Box Machine Learning: A Fusion of Explainable AI and Formal Methods for Healthcare Decision-Making," IEEE Access, 2024.
[14] G. Han, L. Pusztai, and C. Hatzis, "Data augmentation based on waterfall plots to increase value of response data generated by small single arm Phase II trials," Contemporary Clinical Trials, vol. 110, p. 106589, 2021.
[15] A. W. Kruglanski and W. Stroebe, "The influence of beliefs and goals on attitudes: Issues of structure, function, and dynamics," The handbook of attitudes, vol. 1, pp. 323-368, 2005.
[16] E. C. Alvarez, S. Aspeslagh, and J.-C. Soria, "3D waterfall plots: a better graphical representation of tumor response in oncology," Annals of Oncology, vol. 28, no. 3, pp. 454-456, 2017.
[17] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," Machine Learning and Knowledge Extraction, vol. 3, no. 3, pp. 525-541, 2021.
[18] C. Rudin and J. Radin, "Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition," Harvard Data Science Review, vol. 1, no. 2, pp. 1-9, 2019.

[19] S. Rao, S. Mehta, S. Kulkarni, H. Dalvi, N. Katre, and M. Narvekar, "A study of LIME and SHAP model explainers for autonomous disease predictions," in 2022 IEEE Bombay Section Signature Conference (IBSSC), 2022: IEEE, pp. 1-6.
[20] D. Shin, "User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability," Journal of Broadcasting & Electronic Media, vol. 64, no. 4, pp. 541-565, 2020.
[21] A. Gramegna and P. Giudici, "SHAP and LIME: an evaluation of discriminative power in credit risk," Frontiers in Artificial Intelligence, vol. 4, p. 752558, 2021.
[22] S. Raptis, C. Ilioudis, and K. Theodorou, "From pixels to prognosis: unveiling radiomics models with SHAP and LIME for enhanced interpretability," Biomedical Physics & Engineering Express, vol. 10, no. 3, p. 035016, 2024.
[23] D. Ueda et al., "Fairness of artificial intelligence in healthcare: review and recommendations," Japanese Journal of Radiology, vol. 42, no. 1, pp. 3-15, 2024.
[24] T. B. Modi, "Artificial Intelligence Ethics and Fairness: A study to address bias and fairness issues in AI systems, and the ethical implications of AI applications," Revista Review Index Journal of Multidisciplinary, vol. 3, no. 2, pp. 24-35, 2023.
[25] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, and S. Kujala, "Transparency and explainability of AI systems: From ethical guidelines to requirements," Information and Software Technology, vol. 159, p. 107197, 2023.
[26] D. Mane, A. Magar, O. Khode, S. Koli, K. Bhat, and P. Korade, "Unlocking Machine Learning Model Decisions: A Comparative Analysis of LIME and SHAP for Enhanced Interpretability," Journal of Electrical Systems, vol. 20, no. 2s, pp. 1252-1267, 2024.
[27] R. Jinad, A. Islam, and N. Shashidhar, "Interpretability and Transparency of Machine Learning in File Fragment Analysis with Explainable Artificial Intelligence," Electronics, vol. 13, no. 13, p. 2438, 2024.
[28] A. J. A. S. Muhammad, "Evaluation of Explainable AI Techniques for Interpreting Machine Learning Models," ed, 2024.
[29] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," Brain Informatics, vol. 11, no. 1, 2024, doi: 10.1186/s40708-024-00222-1.
[30] S. Knapič, A. Malhi, R. Saluja, and K. Främling, "Explainable artificial intelligence for human decision support system in the medical domain," Machine Learning and Knowledge Extraction, vol. 3, no. 3, pp. 740-770, 2021.
[31] A. Nieto Juscafresa, "An introduction to explainable artificial intelligence with LIME and SHAP," 2022.
[32] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180-186.
[33] N. R. Cook, "Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve," Clin Chem, vol. 54, no. 1, pp. 17-23, Jan 2008, doi: 10.1373/clinchem.2007.096529.
[34] A. M. Braşoveanu and R. Andonie, "Visualizing and explaining language models," in Integrating Artificial Intelligence and Visualization for Visual Knowledge Discovery: Springer, 2022, pp. 213-237.
[35] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell, "" Why should you trust my explanation?" understanding uncertainty in LIME explanations," arXiv preprint arXiv:1904.12991, 2019.

[36] R. Jain et al., "Explaining sentiment analysis results on social media texts through visualization," Multimedia Tools and Applications, vol. 82, no. 15, pp. 22613-22629, 2023.
[37] K. M. Mole, "Factors in Computer Service Selection for Small Business," Polytechnic Institute of Brooklyn, 1969.
[38] A. M. Salih et al., "A perspective on explainable artificial intelligence methods: SHAP and LIME," Advanced Intelligent Systems, p. 2400304, 2024.
[39] Y. Ramon, D. Martens, F. Provost, and T. Evgeniou, "A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C," Advances in Data Analysis and Classification, vol. 14, pp. 801-819, 2020.
[40] C. Wilson et al., "Building and auditing fair algorithms: A case study in candidate screening," in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 666-677.
[41] T. Yan and C. Zhang, "Active fairness auditing," in International Conference on Machine Learning, 2022: PMLR, pp. 24929-24962.
[42] D. Sacharidis, G. Giannopoulos, G. Papastefanatos, and K. Stefanidis, "Auditing for spatial fairness," arXiv preprint arXiv:2302.12333, 2023.
[43] K. Maughan, I. C. Ngong, and J. P. Near, "Prediction Sensitivity: Continual Audit of Counterfactual Fairness in Deployed Classifiers," arXiv preprint arXiv:2202.04504, 2022.
[44] A. Elmaraghy[1], J. Montali, M. Restelli, F. Causone, and P. Ruttico, "Check for updates Towards an AI-Based Framework for Autonomous Design and Construction: Learning from Reinforcement Learning Success," in Computer-Aided Architectural Design. INTERCONNECTIONS: Co-computing Beyond Boundaries: 20th International Conference, CAAD Futures 2023, Delft, The Netherlands, July 5–7, 2023, Selected Papers, 2023: Springer Nature, p. 376.
[45] J. Feng et al., "Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare," NPJ digital medicine, vol. 5, no. 1, p. 66, 2022.
[46] T. Lesort, M. Caccia, and I. Rish, "Understanding continual learning settings with data distribution drift analysis," arXiv preprint arXiv:2104.01678, 2021.
[47] K. Drukker et al., "Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment," Journal of Medical Imaging, vol. 10, no. 6, pp. 061104-061104, 2023.
[48] S. C. Slota et al., "Many hands make many fingers to point: challenges in creating accountable AI," Ai & Society, pp. 1-13, 2023.
[49] D. Domínguez Figaredo and J. Stoyanovich, "Responsible AI literacy: A stakeholder-first approach," Big Data & Society, vol. 10, no. 2, p. 20539517231219958, 2023.
[50] B. Rakova, J. Yang, H. Cramer, and R. Chowdhury, "Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices," Proceedings of the ACM on Human-Computer Interaction, vol. 5, no. CSCW1, pp. 1-23, 2021.