From Black Boxes to Glass Boxes: Measuring Explainability in Artificial Intelligence Systems Using Interpretability Benchmarks

Nikita
MTech (AIML) Scholar, Department of Computer Science and Engineering UIET, Maharshi Dayanand University, Rohtak, Haryana, India

***Abstract*—AI has slowly and subtly crept into our daily lives through our devices and systems. This technology has been used in various industries, including the medical field, financial sector, transport industry, and online platforms. In the medical field, artificial intelligence has helped detect diseases early and identify possible health threats. The banking sector uses artificial intelligence to detect any fraudulent activities and gain insight into customer behaviors. Self-driving cars are another area where artificial intelligence has been implemented, especially for navigating. The reason why companies are increasingly relying on artificial intelligence technology is the need to analyze large amounts of data within a short time, something that would not have been possible without using this technology.**

This is not to say that one of the biggest problems facing AI remains unchanged, which is that the systems in question are often extremely hard to understand. Deep-learning approaches, transformers, and other sophisticated models have been created with performance in mind, not with comprehensibility in mind. In layman's terms, the user gets the result – a diagnosis, an insurance approval, a driving route – without having any idea of how the process arrived at that result. As a result, we now speak about the phenomenon of "black box" AI, referring to the inability to see what goes on inside the machine that creates those results.

And the consequences for ethics in artificial intelligence cannot be understated. The GDPR, EU AI Act, among other policies, stipulate that AI systems need to be accountable and transparent, especially if there is a question about personal data and critical decision-making processes.

Such challenges have led to a lot of discussions on the topic of ethical AI. It no longer remains a mere dream but an important aspect that should be considered when designing AI-based products and regulating them. Wherever a machine is not able to justify its actions, neither the user nor the engineer would be able to question them. Governments all over the world have understood this necessity, resulting in the creation of certain requirements regarding explainability of an AI system under certain conditions.both of which can explain predictions without depending on the internal structure of a modelSHAP utilizes principles derived from game theory to evaluate the contribution of individual features to predictions. LIME, on the other hand, tries to construct simple models locally around particular predictions for better understanding of their logic by end-users. Also, visualization techniques like Integrated Gradients and Grad-CAM, widely used for interpreting results of deep learning models, especially related to images, are considered in the article. Moreover, rule-based approaches, including anchor explanations, and attention-based techniques that enhance symbolic reasoning and attention mechanisms in AI are discussed..

The effectiveness of various explainability techniques was evaluated using datasets from different areas and types of data. Specifically, the Adult Income dataset from UCI was chosen to conduct experiments on structured/tabular data; the Chest X-Ray14 dataset from NIH was selected to perform medical imaging studies; and the SST-2 dataset was selected to evaluate the performance on natural language processing tasks. Among the criteria to be used in the evaluation of the explainability techniques were interpretability, accuracy of the explanation, the cost of computing the explanation, scalability and robustness to adversarial attacks. It is also evaluated how understandable the explanations are to technical/non-technical persons. Each explainability technique has its own strengths weaknesses and practical trade-offs depending on the dataset model complexity and application domain involved. Researchers noticed this repeatedly across different studies. The model still wasn't fully transparent in many situations even after applying explanation methods. That's where the real concern starts.

***Index Terms*—Explainable AI (XAI), Artificial Intelligence, Machine Learning, SHAP, LIME, Grad-CAM, Integrated Gradients, Black-Box Models, Responsible AI, Ethical AI, Deep Learning, Transformer Model.**

I. INTRODUCTION

A. The Urgency of Explainability in Modern AI

Artificial Intelligence used to be discussed mostly in research environments and university projects. Earlier it felt distant from ordinary life. That changed gradually over the years AI systems have slowly become part of normal day-to-day services that people interact with almost constantly now. Most users probably don't even notice it happening. Hospitals use machine learning models while diagnosing diseases and tracking patient conditions over time. This becomes more visible in healthcare datasets. Financial organizations also rely heavily on AI systems to detect fraud suspicious transactions and unusual activity patterns in customer accounts.

At first these systems looked completely reliable because of their speed and accuracy. But after deeper analysis the limitation became obvious. While going through different papers one thing appeared repeatedly. Many AI systems still don't clearly explain how decisions are actually made. The model still wasn't transparent. That's where the real concern starts because these technologies are now connected to decisions that affect real people in healthcare banking and many other areas.

Transportation systems also use intelligent algorithms for navigation and safety improvements. Recommendation engines are now part of almost every digital platform people use online. Slowly these technologies became integrated into normal decision-making processes across different industries.

At first the performance of these systems looked impressive. Predictions were often fast and highly accurate. But while reviewing multiple studies another limitation became obvious. Most advanced AI systems still behave like black boxes. Deep neural networks and transformer-based models process massive amounts of data efficiently yet they rarely explain how a prediction was actually generated. In many cases users only receive the final output. The reasoning process remains hidden somewhere inside the model architecture. The model still wasn't transparent. Researchers noticed this repeatedly in different application areas.

This issue becomes easier to notice in healthcare systems. If a model incorrectly predicts that a patient doesn't have a serious illness treatment may get delayed and the outcome can become dangerous. Similar concerns appear in banking applications where automated systems reject loan requests without giving understandable explanations. Criminal justice systems face comparable problems because prediction models trained on historical records may unintentionally reproduce earlier social bias patterns. Autonomous vehicles create another difficult situation since even small prediction failures may lead to accidents. And this can't be ignored. AI decisions increasingly affect real people and real-world consequences.

While going through different papers it became clear that explainability is no longer treated as an optional feature in AI systems. Over the last few years governments and regulatory organizations started paying much more attention to transparency and accountability in AI systems. While going through different papers this issue appeared repeatedly especially in discussions related to responsible AI deployment. The European Union introduced GDPR partly because automated decision-making systems were becoming difficult for ordinary users to understand. Later the EU AI Act added stronger explainability requirements for AI applications classified as high-risk. Similar concerns can also be seen in the NIST AI Risk Management Framework used in the United States. IEEE ethical guidelines discuss transparency together with accountability and human well-being during AI development processes.

At first many organizations focused mostly on prediction accuracy. But after deeper analysis the limitation became obvious. The model still wasn't transparent. What makes this more difficult is that regulations alone can't fully solve the explainability problem. Researchers noticed this repeatedly. Different industries interpret transparency requirements differently depending on the application domain and risk involved. And this can't be ignored.

What makes this more difficult is that explainability itself still doesn't have one universally accepted definition. Different researchers evaluate it differently depending on the context application and user group involved. One AI system may provide good global interpretability while still failing to explain individual

predictions clearly. Another system may appear understandable for majority populations but generate weak or misleading explanations for minority groups because those groups remain underrepresented in training data. So even when organizations claim their systems are explainable verifying those claims becomes complicated without proper evaluation standards. This created another issue. Measuring explanation quality consistently across different machine learning applications remains difficult even now.

Many explanation methods also fail from the perspective of ordinary users. Technical visualizations such as SHAP plots may help machine learning researchers understand feature contributions but patients loan applicants or legal defendants may still struggle to understand why a system produced a certain decision. That's where the real concern starts. Explainability isn't only about developers understanding model behaviour. It also involves whether affected people can realistically interpret the reasoning behind decisions that influence their lives.

To deal with these concerns researchers began focusing heavily on Explainable Artificial Intelligence commonly called XAI. XAI includes different methods frameworks and techniques designed to improve transparency in machine learning systems. The main idea is actually simple. People should be able to understand how AI systems reach conclusions and which features influence predictions most strongly. Explainability becomes useful not only for developers and researchers but also for doctors financial analysts policymakers legal authorities and regular users interacting with AI systems in everyday situations.

One important objective of XAI is improving trust between humans and intelligent systems. When explanations are available users usually feel more comfortable relying on AI technologies. Explainability methods also help researchers detect hidden bias identify model weaknesses and improve prediction reliability. For example if a healthcare AI model focuses on irrelevant image regions instead of actual disease indicators researchers can retrain the model using improved datasets. Similar concerns appear in recruitment and lending applications where features such as gender race or location may unintentionally influence predictions unfairly. XAI techniques help expose these hidden behaviours. But after deeper analysis the limitation became obvious. No single explanation method completely solves the transparency problem across every domain.

Over time several explainability techniques became widely used in machine learning research. SHAP is one of the most commonly discussed methods because it estimates feature contribution using concepts derived from game theory. LIME focuses more on explaining individual predictions through simpler local models. In deep learning applications visualization-based techniques such as Grad-CAM and Integrated Gradients are widely used especially in healthcare and computer vision tasks. Grad-CAM highlights image regions influencing predictions while Integrated Gradients measure feature importance by observing changes in neural network outputs. Attention-based explanation approaches are also heavily applied in transformer architectures and natural language processing systems because they help identify which words or input regions received stronger focus during prediction generation. Some techniques are easier to interpret than others though. And different methods solve different kinds of problems.

B. The Technical and Regulatory Gap

One major issue discussed in this research is the growing gap between what explainable AI promises in theory and what organizations are actually able to measure or verify in real systems. Many companies and institutions now publicly claim that their AI systems are transparent fair and explainable. Regulatory policies also increasingly stress responsible AI and understandable automated decision-making. But after reviewing different studies it became clear that there still isn't one accepted method for deciding whether an AI system is truly explainable or whether the explanations generated by the system are actually meaningful reliable and fair. This mismatch between theoretical expectations and practical implementation is often described as the explainability measurement gap.

The explainability measurement gap exists mainly because explainable AI research still lacks standardized evaluation methods that work consistently across different application domains. A large number of explainability techniques have been developed in recent years but organizations still face uncertainty about which methods should actually be

used and how explanation quality should be measured properly. In many situations explanations may appear useful on the surface while failing to satisfy ethical or regulatory expectations during deeper evaluation. This issue becomes more visible in healthcare datasets and financial decision systems where prediction outcomes can directly affect people's lives. And this can't be ignored. Poor explanations can damage public trust very quickly.

One important part of this problem is metric fragmentation. Different explainability studies introduce different evaluation metrics such as faithfulness stability interpretability comprehensibility robustness consistency completeness and fairness. Each metric measures a different characteristic of explanation quality. Faithfulness for example checks whether an explanation truly reflects the model's internal reasoning process. Stability focuses on whether explanations remain similar when there are only small changes in input data. Comprehensibility examines whether humans can actually understand the explanation while robustness measures resistance against adversarial noise or manipulation.

At first this looked reliable. But after deeper analysis the limitation became obvious. Even though these metrics are useful individually there is still no universal framework explaining which metrics matter most in different AI applications or what minimum standards should be satisfied before a system is considered trustworthy. This created another issue. Organizations may selectively choose only those metrics that make their systems appear more explainable while ignoring weaknesses in other important areas. For instance a company may emphasize interpretability scores while avoiding discussion about demographic inconsistency or low faithfulness. Researchers noticed this repeatedly while comparing industrial explainability reports. Because of this regulators and auditors often struggle to determine whether selected evaluation metrics are actually appropriate for a particular deployment environment.

The lack of standardized measurement practices also creates problems when researchers try to compare explainability methods across different studies. Different datasets different evaluation metrics and completely different experimental procedures are often used. So comparing results becomes difficult. During the literature survey this issue appeared in many papers. The explainability field still feels fragmented in several ways and industries don't always receive clear guidance regarding how explanation quality should be evaluated consistently. This slows the development of trustworthy AI systems and increases uncertainty for organizations attempting to comply with emerging AI regulations.

Another major issue connected with the explainability measurement gap is toolkit fragmentation. Over time several explainability toolkits became widely used including SHAP LIME and TCAV. Each of these methods focuses on different aspects of interpretability and produces different forms of explanations. SHAP provides mathematically grounded feature attribution explanations. LIME mainly focuses on local surrogate explanations around individual predictions. TCAV attempts to analyze conceptual influence in deep learning systems. All of them try to improve transparency but they don't always produce similar interpretations for the same model behaviour. The model still wasn't transparent in many situations.

What makes this more difficult is the confusion these toolkits create for organizations developers and auditors attempting to implement explainable AI systems properly. A company beginning an explainability audit may struggle to decide which toolkit is actually suitable for its models datasets and regulatory requirements. Some methods work better for structured tabular datasets while others are more useful for image analysis or natural language processing tasks. In certain situations different explainability methods may even generate conflicting explanations for exactly the same prediction. Without proper guidance organizations sometimes select explanation tools based mainly on convenience popularity or computational efficiency instead of actual suitability for the problem being addressed.

These explainability techniques are also not perfectly interchangeable. SHAP offers strong theoretical guarantees and generally high faithfulness but computational cost becomes a serious issue for larger models. LIME produces relatively intuitive local explanations but its stability can vary because of random sampling behaviour. TCAV supports concept-level interpretation but depends heavily on predefined concept examples and training data quality. While going through different papers I noticed that relying on only one explainability toolkit often gives an incomplete understanding of model behaviour. There

is still limited practical guidance explaining when methods agree when they diverge and how they should be combined effectively inside explainability audit pipelines.

The third and probably most socially important dimension of the explainability measurement gap involves demographic inequality in explanation quality. Earlier fairness research already showed that prediction accuracy can vary across demographic groups. This study suggests that explanation quality itself may also differ between groups in ways that disadvantage already vulnerable populations. Some groups may receive explanations that are less stable less understandable or less reliable compared to others. That's where the real concern starts.

This problem becomes especially serious because AI systems are increasingly used in hiring loan approvals criminal justice healthcare insurance and educational admissions. If some demographic groups receive lower-quality explanations they may have less ability to question or challenge automated decisions affecting their lives. For example loan applicants from disadvantaged communities may receive vague inconsistent or overly technical explanations compared to applicants from more privileged groups. Similar patterns may appear in healthcare systems if minority populations remain underrepresented in training data. This was discussed in several papers I reviewed. Bias inside training datasets doesn't just affect prediction quality. It can also affect explanation quality itself.

The study further shows that explanation disparities often become stronger in regions where proxy discrimination is active. Proxy discrimination happens when machine learning models don't directly use sensitive features like race or gender but still end up learning similar patterns through other related variables. At first this may look harmless because the protected attribute itself isn't included in the dataset. But after deeper analysis the limitation becomes obvious. The model can still make biased decisions by relying on information that is indirectly connected to demographic identity. Researchers noticed this repeatedly while studying fairness issues in AI systems.

For example location data income level language preference or even shopping behaviour may act as substitute indicators for sensitive demographic information. Because these variables are often strongly correlated with race gender or social background the model may continue producing discriminatory outcomes even when protected attributes are removed completely. This created another issue. Simply deleting sensitive columns from training data doesn't always prevent unfair behaviour in machine learning systems.

What makes this more difficult is that proxy discrimination usually remains hidden during normal model evaluation. The system may appear fair on the surface because it technically avoids using protected features directly. The model still wasn't transparent. While going through different papers it became clear that many AI systems unintentionally learn these hidden associations from historical data patterns. And this can't be ignored especially in healthcare finance recruitment and criminal justice applications where automated predictions can directly affect people's opportunities and outcomes.

In some cases proxy variables are not even obvious to developers during model training. A feature that seems neutral at first may still contain indirect demographic information after deeper statistical analysis. This was discussed in several papers I reviewed. Because of this researchers increasingly argue that fairness evaluation should not only focus on direct discrimination but also on hidden proxy relationships that influence prediction behaviour indirectly.

Even when sensitive attributes are removed models may still infer demographic information through variables such as postal codes income patterns education history or employment records. As a result explanation quality may deteriorate most severely for groups already affected by hidden bias inside the system.

Another concern is that many explainability evaluation frameworks only calculate average explanation quality across the entire dataset. Average metrics can hide serious explanation failures affecting smaller demographic groups because the overall score may still appear acceptable. So organizations may incorrectly assume that their systems are fair and transparent while certain populations continue receiving weak or misleading explanations. This becomes dangerous in high-risk applications where people depend on explanations to understand important automated decisions. Because of this the study emphasizes evaluating explanation quality

separately across demographic groups instead of relying only on aggregate metrics.

The findings from this research suggest that explainability cannot be treated only as a technical machine learning problem. It is strongly connected with ethics accountability social justice and human rights. Building responsible explainable AI systems requires more than just better algorithms. Standardized evaluation frameworks demographic fairness analysis regulatory oversight and human-centered design approaches are equally important. Future AI governance systems will likely need to ensure that explanations are not only technically correct but also understandable accessible and reliable for all people affected by automated decision-making systems.

C. Research Questions

The study is mainly organized around three research questions. These questions were framed after reviewing earlier work and identifying some common gaps that still appear in explainability research. While going through different papers one thing became noticeable. Many studies discuss explainability methods separately but fewer works compare them carefully under the same experimental conditions. This created another issue. It becomes difficult to understand where one method performs better and where another method may fail.

The first research question focuses on comparing SHAP LIME and TCAV directly using the same datasets and classification models. The intention here is not only to compare accuracy of explanations but also to examine explanation quality overall coverage and computational cost. Some explanation methods may capture certain model behaviours that others completely miss. At first this looked straightforward. But after deeper analysis the limitation became obvious. Different explainability toolkits often focus on different aspects of interpretation and that makes comparison difficult. Another concern is practical usage in compliance auditing because incomplete explanations may hide important risks or biases present in the model. The model still wasn't transparent in many cases.

The second research question examines how explanation quality changes depending on model complexity dataset properties and protected demographic groups. This part became especially interesting during the literature review stage because explanation quality doesn't remain constant across all situations. Researchers noticed this repeatedly. Metrics such as faithfulness stability comprehensibility and completeness may behave differently when models become more complex. Similar patterns appear when datasets contain demographic imbalance or sensitive features. What makes this more difficult is that some explanation methods may work reasonably well for simpler classifiers but struggle with deep learning systems. And this can't be ignored. The relationship between classifier complexity and explanation quality still remains unclear in several studies.

The third research question focuses on combining SHAP global attribution analysis with LIME local explanation analysis in order to better understand how models actually make predictions. While reviewing earlier research this issue appeared repeatedly. Global explanations sometimes provide broad feature importance patterns but fail to explain individual predictions properly. Local explanations on the other hand may explain specific predictions clearly while missing overall model behaviour. So the study attempts to examine whether combining both approaches can provide deeper insight into prediction mechanisms especially in situations involving proxy discrimination. That's where the real concern starts. Scalar explainability metrics alone may not fully reveal hidden discriminatory behaviour inside machine learning systems and this limitation appears in several existing studies as well.

D. Original Contributions

• One important contribution of this research is the detailed comparison carried out between SHAP, LIME, and TCAV using multiple evaluation dimensions and benchmark datasets. While going through different papers it became noticeable that most studies usually focus on only one explainability technique or evaluate methods using limited criteria. In this work the comparison was broader and more systematic. Ten different evaluation dimensions were considered across three benchmark datasets in order to understand how these explanation methods behave under different conditions. At first this looked straightforward. But after deeper analysis the differences between the methods became much more complex than expected.

Another major observation from the study was related to demographic inequality in explanation quality. This created another issue. Explanations generated for female and racial minority instances were often less stable and less consistent compared to majority-group instances. Researchers noticed this repeatedly in areas where proxy discrimination was already active inside the feature space. In simple terms the explanation methods themselves were not behaving equally across all demographic groups. This issue becomes more visible in healthcare datasets and financial prediction systems where sensitive attributes indirectly influence model behaviour. And this can't be ignored because explanation inconsistency may hide unfair treatment instead of exposing it.

The study also produced an interesting result regarding model complexity and explainability. Common assumptions usually suggest that explainability becomes worse as models become more complex. However the experiments showed something different. SHAP faithfulness actually improved with increasing model complexity while LIME faithfulness decreased under similar conditions. The pattern was almost opposite to what many earlier discussions suggested. This was discussed in several papers I reviewed although not always explored in detail experimentally. The findings indicate that different explanation techniques respond very differently to complex model architectures and their behaviour cannot be generalized easily.

Another part of the analysis combined SHAP global feature importance with LIME instance-level explanations to identify hidden proxy discrimination mechanisms inside prediction systems. What makes this more difficult is that fairness metrics alone often show that bias exists but fail to explain how it appears internally. The combined attribution analysis helped uncover examples such as marital status acting as a sex-related income proxy and uneven weighting of prior convictions across racial groups in criminal justice datasets. The model still wasn't transparent in every situation though. Some hidden interactions remained difficult to interpret fully even after explanation generation.

The research also developed a domain-specific recommendation matrix covering healthcare financial services criminal justice and human resources applications. Instead of treating explainability as a one-size-fits-all problem the matrix provides guidance on which explainability toolkit and evaluation metrics are more suitable for different application areas. Regulatory expectations also vary between domains. So the recommendation process needed to consider not only technical performance but also compliance requirements connected to sectors such as healthcare and finance.

Finally the work proposed a five-stage Explainability Compliance Pipeline consisting of Assess Generate Evaluate Validate and Document phases. The idea behind the pipeline was to convert broad explainability requirements from frameworks such as the EU AI Act GDPR Article 22 and the NIST AI Risk Management Framework into a more practical organizational workflow. While reviewing policy documents this gap appeared repeatedly. Many regulations discuss transparency at a high level but organizations still struggle to apply those requirements consistently in real development environments. The proposed pipeline attempts to make explainability more structured repeatable and easier to integrate into AI governance practices.

E. Paper Organization

The remaining sections of this dissertation are organized step by step based on the overall research flow. Section II discusses the technical background related to explainability methods and different XAI frameworks. It also reviews earlier research studies connected to explanation evaluation and the different explanation quality metrics considered during this work. While going through different papers it became clear that researchers still evaluate explainability in very different ways. This created another issue. There isn't one single standard followed everywhere.

Section III explains the experimental methodology used in this study including dataset details model configurations and evaluation procedures. The datasets selected for experimentation along with classifier settings are described in this section. Some parts of the methodology looked straightforward at first. But after deeper analysis the limitation became obvious especially while comparing explanation quality across different domains.

Section IV presents the comparison between different explainability toolkits and also discusses the domain recommendation matrix prepared during the analysis stage. The intention here is to understand which explainability methods work better in particular

application areas. The model still wasn't transparent in several cases even after applying explanation techniques. Researchers noticed this repeatedly.

The complete experimental findings are discussed in Section V. This section includes baseline model performance explanation quality evaluation SHAP-based global feature attribution analysis LIME instance-level interpretation results and TCAV concept-level analysis. What makes this more difficult is that different methods sometimes provide different explanations for the same prediction. And this can't be ignored because consistency becomes important in high-risk applications. This was discussed in several papers I reviewed during the literature survey process. Section VI focuses on interpreting the findings with respect to the research questions considered in this dissertation. It also discusses practical implications for developers organizations and regulatory bodies working with explainable AI systems. Limitations of the study along with possible threats to validity are also addressed in this section because some experimental constraints couldn't be fully avoided during implementation.Future research directions are briefly discussed in Section VII.

II. BACKGROUND AND RELATED WORK

A. Conceptual Foundations of Explainable AI

The study of explainability in machine learning did not suddenly appear with modern AI systems. Its roots go back much earlier, especially to the long-standing debate between model accuracy and model transparency. Earlier machine learning approaches were mostly based on interpretable methods such as linear regression logistic regression decision trees and rule-based systems. These models were not preferred only because they were simple. One important reason was that humans could actually understand how decisions were being made. A doctor using a logistic regression model for disease prediction could examine feature coefficients directly and check whether the results matched medical knowledge and clinical expectations. The reasoning process was visible. That made explanations easier.

Things changed when more advanced machine learning techniques started dominating the field. Ensemble methods gradient-boosted trees and deep neural networks began producing much higher prediction accuracy on complex datasets. At first this looked reliable. But after deeper analysis another limitation became obvious. These models learned highly complicated nonlinear relationships between features that humans could not easily track or interpret. A random forest containing hundreds of trees cannot realistically be inspected line by line by a human analyst. The same issue appears in XGBoost models where feature interactions become too complex to summarize clearly. The model still wasn't transparent. Researchers noticed this repeatedly while comparing traditional interpretable models with modern high-performance architectures.

While going through different papers one thing became clear. The research community responded to this loss of interpretability in two different ways. One direction focused on building models that remain interpretable from the beginning itself. These are often called intrinsically interpretable models. Generalized additive models decision trees with restricted depth rule-based systems and neural networks with monotonicity constraints fall into this category. The idea was fairly straightforward. Build systems whose decision process humans can still follow without requiring additional explanation tools.

The second direction became much more popular in practical applications. Instead of simplifying the model itself researchers started developing methods that explain already trained black-box systems. This created another issue. The explanations themselves sometimes became difficult to evaluate properly. Still post-hoc explainability methods such as SHAP LIME and TCAV became widely used because they allowed researchers to generate explanations without modifying internal model architecture. What makes this more difficult is that these methods try to explain systems whose internal reasoning may already be extremely complicated. So the explanation process itself can sometimes involve approximations assumptions or simplifications. This was discussed in several papers I reviewed during the literature survey stage.

Another important observation in explainability research is the distinction between global and local explanations. Different stakeholders require different types of explanations depending on the situation. Global explanations attempt to describe overall model behaviour across the complete dataset. They usually identify which features influence predictions most strongly on average. Local explanations work

differently because they focus only on a single prediction or a specific input instance. For example a regulator studying bias in an AI system may require global explanations to evaluate system-wide behavior. A patient or loan applicant however usually wants to understand one specific decision affecting them directly. And this can't be ignored. Different explanation scopes serve very different practical needs.

The difference between model-specific and model-agnostic explanation methods is also discussed frequently in explainability literature. Model-specific approaches are designed for particular model architectures. TreeSHAP is one example because it uses the internal structure of tree-based systems to calculate feature attributions efficiently. These methods often provide faster and more accurate explanations but they can't be applied universally across every machine learning model. Model-agnostic methods such as KernelSHAP and LIME work differently. They treat the original model almost like a black box and generate explanations by observing prediction changes after modifying input data.

At first this seemed more flexible because model-agnostic approaches can work with almost any machine learning architecture. But after deeper analysis another limitation appeared. These methods often require higher computational cost and may introduce approximation errors during the explanation process. So flexibility comes with trade-offs. Different papers approached this issue differently and no single explanation method completely solved every transparency challenge in machine learning systems.

B.Regulatory Frameworks

B.1 GDPR)

Earlier many AI systems operated with very little transparency. But after GDPR the discussion around accountability and explainability became much stronger in both research and industry environments.

Article 22 of the GDPR mainly focuses on decisions made entirely through automated processing. These decisions may create legal effects or other serious consequences for individuals. This becomes more noticeable in systems related to healthcare banking insurance and recruitment processes. If an automated system rejects a loan application or influences hiring decisions people naturally expect some level of explanation behind those outcomes. The regulation attempts to address that concern by giving individuals certain protections when automated systems are involved in decision-making.

At first this looked straightforward. But after reading different interpretations of GDPR the situation appeared more complicated. Researchers and legal experts still debate how detailed these explanations actually need to be and what counts as a "meaningful" explanation in practice. Even so data protection authorities generally agree on one thing. Explanations should be understandable for ordinary people and not only for technical experts. The explanation must relate to the individual case itself rather than providing only a general description of how the AI system works. Researchers noticed this repeatedly while discussing explainability requirements under GDPR.

This created another issue.

The GDPR requirements also affect the technical explanation methods discussed in this study. While going through different papers it became clear that some explainability techniques align more naturally with Article 22 than others. LIME for example produces local explanations for individual predictions by showing how specific features influenced the final result using a simpler interpretable model. Because these explanations are tied to one particular prediction they can usually be communicated more easily to affected users without requiring deep machine learning knowledge. This becomes more visible in healthcare datasets where doctors and patients often need understandable explanations for individual predictions rather than broad system-level summaries.

SHAP works somewhat differently. It provides stronger mathematical consistency and also supports both local and global interpretability. But the explanations produced by SHAP sometimes require additional interpretation before ordinary users can understand them properly. The model still wasn't fully transparent from a non-technical perspective. TCAV methods create concept-level explanation scores which can help researchers analyze model behaviour at a higher conceptual level. However these scores may still need conversion into clearer instance-specific explanations if they are expected to satisfy GDPR-style explanation requirements for individuals affected by automated decisions.

What makes this more difficult is that legal expectations and technical explainability methods don't always align perfectly. A method may appear

technically strong from a machine learning perspective but still fail to provide explanations that ordinary people can realistically understand. And this can't be ignored because explainability is no longer only a research topic. It is increasingly becoming part of legal compliance and responsible AI deployment as well.

B.2 EU Artificial Intelligence Act

The EU AI Act finalized in 2024 is widely considered one of the strictest and most detailed regulatory frameworks introduced for Artificial Intelligence until now. While going through different research papers this point appeared repeatedly. Many researchers describe the Act as a major step toward controlling how high-risk AI systems are developed and deployed across different sectors. The regulation mainly focuses on transparency accountability and user safety especially in applications where AI decisions can directly affect people's lives.

What makes this more difficult is that these systems directly affect people's lives in practical situations. Because of that the Act requires organizations to maintain detailed technical documentation and proper human oversight mechanisms. AI outputs should also be understandable enough for deployers providers and affected individuals to interpret them correctly. At first this looked straightforward. But after deeper analysis the limitation became obvious. Explainability itself still doesn't have one universally accepted interpretation and this creates challenges while implementing these regulations in real systems.

Annex IV of the EU AI Act explains the type of documentation high-risk AI systems are expected to maintain. This includes information about system capabilities limitations training datasets testing procedures instructions for deployment and details related to accuracy robustness and non-discrimination properties. Researchers noticed this repeatedly. Transparency requirements are no longer treated as optional recommendations. They are becoming mandatory compliance requirements in many applications.

Article 13 of the Act focuses specifically on transparency. According to this requirement high-risk AI systems should be designed in a way that allows users and deployers to understand how outputs are generated and how those outputs should be interpreted properly. The model still wasn't transparent in many practical cases though. That's where the real concern starts. Even if organizations satisfy documentation requirements technically users may still struggle to fully understand how predictions are actually produced.

This issue becomes more visible in healthcare datasets and financial applications where incorrect predictions can directly affect individuals. And this can't be ignored. Because of these concerns the transparency obligations mentioned in the EU AI Act strongly connect with the explainability evaluation and audit process discussed in this study.

B.3 NIST AI Risk Management Framework

The MEASURE function focuses on evaluating AI-related risks using both quantitative and qualitative assessment methods. This includes examining whether AI systems provide understandable explanations and whether those explanations remain reliable in practical situations. The GOVERN function deals more with organizational responsibilities. It encourages institutions to create policies and procedures for continuous AI risk management including monitoring explanation quality after deployment. At first this looked straightforward. But after deeper analysis the limitation became obvious. Different organizations may still interpret explanation quality differently depending on the domain and application.

What makes this more difficult is that the NIST AI RMF doesn't enforce one fixed explainability method or a strict technical threshold. Its approach feels more process-oriented compared to the EU AI Act which follows a stronger requirement-based structure. Instead of prescribing exact explanation techniques the framework encourages organizations to determine what level of explainability is appropriate for their own deployment environment choose suitable measurement methods and document the results properly. Researchers noticed this repeatedly. There is flexibility but that flexibility can also create inconsistency across implementations. This created another issue. without standardized evaluation procedures different organizations may claim their systems are explainable even when the quality of explanations varies considerably. That's where the real concern starts. Explainability may exist technically but affected users still may not fully understand the reasoning behind decisions. And this can't be ignored especially in high-risk applications like healthcare finance or criminal justice systems.

The five-stage Explainability Compliance Pipeline proposed in this research was designed mainly to make the NIST RMF MEASURE function more practical and easier to apply in real deployment scenarios. Instead of keeping explainability as only a theoretical requirement the pipeline attempts to convert it into a structured workflow supported by tools and measurable evaluation steps. This was discussed in several studies I reviewed while examining explainability compliance frameworks. The overall idea is to make explanation assessment more systematic without removing the flexibility that the NIST framework originally intended.

C. Explainability Quality Metrics: A Comprehensive Taxonomy

C.1 Faithfulness

Faithfulness, which is sometimes referred to as fidelity in several research papers, is basically used to check whether an explanation truly matches the actual behaviour of the machine learning model. In simple terms the explanation should point toward the same features that genuinely influence the prediction inside the model. If the explanation highlights irrelevant features while the model is actually relying on something else then the explanation becomes misleading even if it looks understandable to humans. At first this looked reliable in many systems. But after deeper analysis the limitation became obvious. Some explanations appeared convincing visually while failing to represent the real decision-making process happening inside the model. Researchers noticed this repeatedly.

What makes this more difficult is that human-readable explanations are not always faithful explanations. A method may generate outputs that seem logical and easy to interpret but the model itself may not actually depend on those highlighted features while making predictions. That's where the real concern starts. False explanations can create false trust. And this can't be ignored especially in domains like healthcare or finance where prediction errors may affect real people. While going through different papers this issue appeared quite frequently in discussions related to explainability evaluation.

In this study faithfulness is evaluated using the deletion-AUC metric. The basic idea behind this metric is fairly straightforward. Features identified as important by the explanation method are removed one by one based on their attribution scores. After removing each feature the prediction probability of the model is measured again. If the explanation is actually faithful then removing highly important features should cause a sharp drop in prediction confidence because those features were genuinely contributing to the output. Lower-ranked features should create smaller changes when removed.

C.2 Stability (Consistency Under Perturbation)

Stability in explainable AI basically refers to how consistent an explanation method remains when the input data changes only slightly. In simpler terms if two inputs are very similar the explanations generated for them should also remain similar. Otherwise the explanation method becomes difficult to trust. While going through different papers this issue appeared repeatedly especially in discussions related to fairness and accountability. If two people with almost identical profiles receive completely different explanations for the same prediction the system starts looking unreliable. The model still wasn't transparent.

This becomes more noticeable in high-risk applications like healthcare finance and legal systems. Small differences in patient records or financial details shouldn't completely change the explanation generated by the model. But in practice this sometimes happens. And that creates another issue. Regulatory systems expect AI decisions to remain consistent and explainable especially when those decisions directly affect people. If explanations keep changing unpredictably organizations may struggle to justify how the system actually behaves. That's where the real concern starts.

In this study stability is evaluated using two different measurements. One of them is internal stability. The other is external stability. Internal stability mainly checks whether the same explanation method produces consistent outputs when it is executed multiple times on the same input using different random seeds. Some explanation algorithms include random sampling procedures and because of that their outputs may slightly vary from one run to another. Researchers noticed this repeatedly in methods like LIME. At first the method seemed reliable. But after deeper analysis the variation became easier to observe especially during repeated executions.

C.3 Comprehensibility

Comprehensibility in explainable AI mainly refers to how easily people can understand the explanations produced by a model. At first this sounds straightforward. But after reading different research papers the issue became more complicated than expected. An explanation that appears simple for a machine learning researcher may still be confusing for a patient a loan applicant or someone affected by a legal decision. This issue becomes more visible in healthcare datasets where explanations often contain technical details ordinary users can't easily interpret. The model still wasn't transparent for everyone involved.

What makes this more difficult is that comprehensibility depends heavily on who the explanation is meant for. Technical users may understand feature importance graphs interaction terms and attribution scores without much difficulty. Regular users usually don't. Researchers noticed this repeatedly in XAI studies. Because of this many studies try to estimate comprehensibility using indirect measures instead of depending only on human feedback. Some common measures include explanation complexity the number of features involved the presence of interaction effects and whether the explanation follows a simple linear structure or not. These measures help to some extent. But they still can't fully replace real user studies involving people from the actual application domain.

In the present study comprehensibility is measured using the minimum number of features needed to explain around 80% of the total attribution score. This approach is commonly discussed in explainable AI literature because explanations are generally expected to remain compact while still providing enough useful information. The basic idea is simple. If too many features are required for understanding a prediction the explanation becomes harder for people to process mentally. This created another issue. Human attention and working memory are limited. Explanations containing a large number of features may increase cognitive load and reduce practical understanding especially for non-technical users.

C.4 Completeness

Completeness is considered an important property in explainable AI because it checks whether an explanation actually captures all the major reasons behind a model's prediction. In simple terms the explanation should account for the full difference between the model's final prediction and its average prediction across the dataset. If some important contributing factors remain hidden then the explanation cannot really be treated as complete. This becomes more serious in high-risk AI applications where missing information inside explanations may hide biased decision-making patterns or unfair behavior inside the system. Researchers noticed this repeatedly while studying explainability methods in sensitive domains.

While going through different papers one thing became very clear. SHAP is usually considered one of the strongest explanation methods when completeness is discussed. The reason mainly comes from its mathematical foundation based on Shapley values from cooperative game theory. One important principle used in SHAP is called the efficiency axiom. This property ensures that all feature contribution values together exactly match the total prediction difference for a particular instance. So basically the prediction gets fully distributed across contributing features without leaving unexplained residual effects behind. At first this looked reliable. But after deeper analysis the limitation became obvious in other explanation methods that could not maintain this property consistently. SHAP explanations therefore appear more mathematically grounded compared to many alternative explainability approaches.This created another issue.

Completeness becomes especially useful during auditing fairness analysis and regulatory evaluation because all influential features are expected to appear inside the explanation. If a model uses hidden proxy variables or indirectly learns biased patterns SHAP is more likely to expose those influences through feature attribution values. This issue becomes more visible in healthcare datasets and financial systems where small hidden biases can affect important decisions. Legal AI systems face similar concerns too. Regulators and auditors can use SHAP explanations to examine whether particular features contribute unfairly toward automated outcomes. And this can't be ignored because transparency requirements are becoming stricter in many industries.

LIME however follows a different idea completely. Instead of focusing on full completeness it mainly prioritizes simplicity and easier interpretation for humans. LIME creates local explanations around a

specific prediction by approximating the model using a simpler interpretable model nearby. To make explanations shorter and easier to understand LIME intentionally selects only a limited number of important features. Usually this is done through feature selection approaches such as Lasso regularization. The remaining smaller feature contributions may simply get ignored even if they still influence the prediction to some extent. The model still wasn't transparent in every situation.

Because of this design choice LIME explanations are not fully complete in a strict mathematical sense. The contribution values generated by LIME do not always represent the entire prediction difference produced by the original model. While the explanations become easier for humans to read they may also miss subtle but important influencing factors. That's where the real concern starts. Hidden proxy variables or indirect discriminatory patterns can remain unnoticed if omitted features still carry sensitive information affecting model decisions. This was discussed in several papers I reviewed during the literature survey stage.

The limitation becomes much more important during fairness auditing and regulatory compliance analysis. If explanation methods fail to capture all influential variables organizations may underestimate the level of discrimination or bias present inside AI systems. For example demographic information may indirectly affect predictions through correlated variables even when those features are not explicitly included. Incomplete explanations may fail to reveal such hidden relationships clearly. So organizations relying only on simplified explanation methods could mistakenly assume their AI systems are fully transparent when important decision patterns still remain hidden underneath.

The comparison between SHAP and LIME basically shows a larger trade-off inside explainable AI research. SHAP provides explanations that are mathematically more complete and theoretically reliable. But these explanations can become computationally expensive and harder for non-technical users to interpret properly. LIME on the other hand generates explanations that are shorter simpler and easier to understand. But it sacrifices completeness because only selected features are highlighted. Different methods solve different problems. While reviewing multiple studies this trade-off appeared repeatedly across explainability research. It also shows that explanation methods cannot be selected casually. The choice depends heavily on the application domain regulatory requirements and how important it is to capture every factor contributing toward the model's decision-making process.

C.5 Cross-Method Consistency

Cross-method consistency basically refers to how closely different explanation techniques agree with each other when identifying important features for the same prediction. In this study the comparison mainly involves SHAP and LIME explanations. If both methods repeatedly highlight similar features for a particular instance it gives stronger confidence that those features are actually important to the model's decision-making process and are not simply produced because of the internal behavior of one explanation algorithm. At first this looked reliable. But after deeper analysis the limitation became obvious. Different explanation methods don't always agree with each other in every situation.

What makes this more difficult is that inconsistency between methods may indicate a deeper issue. If SHAP identifies one set of important features while LIME highlights completely different ones for the same prediction then there is a possibility that one method or even both methods are being influenced more by their own algorithmic assumptions rather than the actual behavior of the machine learning model. Researchers noticed this repeatedly while comparing explanation techniques across different datasets. The model still wasn't transparent.

D. SHAP

The main purpose behind SHAP was to create a unified explainability framework that could provide reliable feature attribution explanations for machine learning models. While going through different papers on explainable AI this method appeared repeatedly because of its strong mathematical foundation and consistency. The basic idea behind SHAP actually comes from cooperative game theory. In game theory several players contribute together toward a shared reward. SHAP applies the same idea to machine learning systems. Here each feature acts like a player and the model prediction becomes the final reward. The explanation method then tries to calculate how much each feature contributed toward that prediction

in a fair way. $\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$

The theoretical background of SHAP is connected to Shapley values proposed by Lloyd Shapley in 1953. Originally these values were developed for solving reward distribution problems in cooperative systems where different participants contribute differently toward the outcome. In machine learning the same principle is reused for prediction explanations. SHAP measures feature contribution by checking how much the prediction changes when a feature is added into different feature combinations. Instead of observing only one combination it averages the contribution across many possible subsets of features. At first this looked complicated mathematically. But after deeper analysis the logic became easier to follow. The method tries to fairly distribute responsibility for a prediction among all input features. $f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i$

One reason SHAP became highly influential in explainable AI research is because it satisfies several fairness properties which give the method strong theoretical support. The first property is efficiency. This ensures that all SHAP values together equal the difference between the actual prediction and the average prediction value. In simple terms every part of the prediction gets accounted for. Nothing is left unexplained. Another important property is symmetry. If two features contribute equally they should receive equal attribution scores regardless of their order in the dataset. Researchers noticed this repeatedly while comparing SHAP with older explanation methods that produced inconsistent feature importance values.

The dummy property is another important aspect of SHAP. Features that do not influence the prediction should receive zero attribution values. This prevents irrelevant variables from appearing artificially important during explanation generation. Then comes additivity. If two models are combined the SHAP values for the combined model should equal the sum of explanations from the individual models. What makes this more difficult is that many explainability methods fail to satisfy all these properties together. SHAP became popular partly because it provides mathematical consistency while still remaining interpretable for practical applications. $\sum_{i=1}^{M} \phi_i = f(x) - E[f(x)]$

But the original SHAP implementation had a serious limitation. Computational complexity. Calculating exact Shapley values required evaluating many different feature combinations and permutations. As the number of input features increased the computation became extremely expensive. This created another issue. Large datasets and complex models became difficult to analyze using exact SHAP calculations. While reviewing implementation studies this limitation appeared in many discussions especially for high-dimensional datasets. The model explanations were reliable. But the computational cost wasn't practical for large-scale industrial systems.

To address this problem researchers later introduced TreeSHAP which was designed specifically for tree-based machine learning models such as Random Forest and XGBoost. TreeSHAP uses the internal structure of decision trees to compute exact SHAP values more efficiently. Instead of checking every possible feature combination directly it applies optimized polynomial-time algorithms. This improvement made SHAP much more practical in real-world applications. And this can't be ignored. Many production AI systems used in finance healthcare fraud detection insurance and business analytics rely heavily on tree-based models. Because of that TreeSHAP became one of the most commonly adopted explainability approaches in industry environments.

Another reason SHAP became widely used is the variety of visualization techniques it provides for understanding model behavior. SHAP summary plots are among the most common visualizations. These plots display overall feature importance across an entire dataset while also showing whether specific features increase or decrease prediction values. They can also reveal how feature effects vary between different data instances. This becomes more visible in healthcare datasets where the same feature may influence predictions differently across patient groups. SHAP dependence plots are also important because they show how feature contribution changes as the feature value changes. These plots help identify non-linear relationships inside machine learning models. Optional interaction coloring can reveal situations where two features influence predictions together instead of independently. This was discussed in several papers I reviewed during the fairness analysis section. Some hidden biases only become visible when feature interactions are analyzed instead of individual variables alone.

SHAP interaction values extend this idea further by explicitly measuring pairwise feature interactions. Two features that appear harmless individually may together create biased or discriminatory outcomes. That's where the real concern starts. Bias in AI systems often comes from feature combinations rather than single variables acting independently. Because of this SHAP interaction analysis became useful in fairness auditing regulatory compliance and AI risk evaluation studies.

For individual predictions SHAP also provides force plots and waterfall plots. These visualizations explain how different features push predictions higher or lower relative to the average prediction value. In loan approval systems for example SHAP force plots can clearly show which financial or demographic attributes increased or reduced approval probability for a specific applicant. Explanations like these improve transparency and help auditors regulators and users better understand why a machine learning system produced a certain decision. The model still wasn't fully transparent in every case though. Some explanations remained difficult for non-technical users to interpret.

E. LIME

The method was mainly developed to explain predictions generated by machine learning models that are otherwise difficult to interpret. While going through different papers on explainable AI, LIME appeared repeatedly because of how simple and practical its explanations are compared to many other methods. The central idea behind LIME is actually quite straightforward. Even if a machine learning model behaves in a highly complex and nonlinear way overall, its behavior near one specific prediction can sometimes be approximated using a much simpler interpretable model. That local approximation then becomes the explanation for the prediction.

One reason LIME became popular so quickly is because it is model-agnostic. It doesn't depend on the internal structure of the original machine learning model. Instead it focuses only on the relationship between inputs and outputs. This means LIME can be applied to neural networks support vector machines random forests gradient boosting models and many other architectures without changing the method itself. That flexibility made it useful in several domains including healthcare finance fraud detection text classification and image analysis. At first this looked very reliable. But after deeper analysis some limitations also became visible.

The explanation process in LIME usually happens in multiple stages. First the original input instance is converted into a simpler and more interpretable form. In tabular datasets continuous values may be transformed into simpler binary conditions depending on thresholds. In text classification tasks words are often treated as either present or absent. Image classification works differently because image regions or superpixels may be turned on and off during explanation generation. This simpler representation makes explanations easier for humans to understand especially for users who don't have deep machine learning knowledge.

After creating the interpretable representation LIME generates many modified versions of the original input. These are usually called perturbed samples. Certain parts of the input are randomly changed removed or masked and then the prediction changes are observed carefully. For example in sentiment analysis some words may be removed to see how much the prediction score changes. In image classification certain image regions may be hidden. This helps LIME estimate which features influence the model prediction most strongly near the original input instance. Researchers noticed this repeatedly in image-based healthcare applications.

The next step involves assigning weights to the generated samples. Samples that remain closer to the original instance receive higher importance because LIME mainly focuses on local behavior instead of trying to explain the complete global model. Usually an exponential similarity kernel is used for this weighting process. The idea is simple. Nearby samples matter more. Distant samples matter less. This created another issue though because selecting the neighborhood properly can sometimes affect explanation quality quite heavily.

Finally LIME fits a sparse linear regression model using the weighted perturbed samples. Regularization methods such as Lasso regression are generally used so that only a smaller set of important features remain in the final explanation. The coefficients from this simplified model become the explanation itself. Positive coefficients indicate features supporting the prediction while negative coefficients indicate features working against it. What makes this more interesting

is that users can often interpret these explanations without much technical background. In a credit approval system for instance LIME may show that low account balance high debt ratio and limited employment history contributed negatively toward loan approval. Explanations like this are easier to understand compared to raw neural network outputs.

One important strength of LIME is flexibility across different data formats. The same method can explain structured tabular data text classification tasks and image-based predictions. In text analysis LIME identifies words influencing predictions most strongly. In image tasks it highlights important image regions responsible for classification outputs. This versatility is probably one reason why LIME became widely adopted in both academic research and industry applications. While reviewing several research papers this consistency appeared repeatedly. Researchers preferred methods that could work across multiple domains without major architectural changes.

Still LIME has several limitations and some of them become serious in sensitive applications. The biggest problem is instability caused by random sampling during perturbation generation. Since explanations depend heavily on randomly generated neighborhoods repeated executions of LIME on the same prediction may produce different explanations if different random seeds are used. The model still wasn't fully stable. This becomes more visible in healthcare datasets where explanation consistency matters a lot for trust and regulatory auditing. If two separate explanation runs produce different feature rankings users may start doubting whether the explanation itself can be trusted. And this can't be ignored.

The present study also observed this instability problem while experimenting with the Adult Income XGBoost model. Repeated LIME runs produced only moderate agreement between feature importance rankings. This was discussed in several papers I reviewed as well. In regulatory environments where explanations must remain repeatable and auditable such instability becomes problematic. A system explaining the same prediction differently every time creates confusion for users auditors and developers. That's where the real concern starts. Explainability methods are expected not only to provide understandable outputs but also to remain reliable over repeated evaluations.

Another limitation comes from the local linear approximation itself. LIME assumes that complex model behavior around a prediction can be approximated using a simpler linear model. In some situations this works reasonably well. But highly nonlinear machine learning systems may contain interactions that local linear approximations fail to capture properly. As model complexity increases the explanation may become less faithful to the original decision boundary. Deep learning systems especially can show this problem quite clearly. So explanation accuracy may decrease even when explanations still appear visually simple and understandable.

To reduce these limitations researchers often apply additional stability verification techniques. One common approach involves generating explanations multiple times and averaging the results to improve consistency. Some practitioners also combine LIME with other explainability methods such as SHAP to strengthen confidence in explanations. Multi-method explainability approaches are becoming more common because individual techniques usually have strengths as well as weaknesses. No single explanation method solves every transparency problem completely.

It helps users understand individual predictions without needing direct access to complex model internals. At first the approach seems extremely effective because the explanations are easy to interpret. But after deeper analysis the limitations become clearer especially regarding instability and approximation accuracy. Even with these weaknesses LIME still plays an important role in improving transparency fairness and accountability across many real-world AI applications.

F. TCAV: Testing with Concept Activation Vectors

TCAV was introduced by Kim and other researchers at Google in 2018. The method was mainly developed to address one important weakness found in feature-attribution techniques. Most explanation methods usually describe model behavior using individual input features. But in many practical situations people are more interested in understanding higher-level concepts instead of isolated variables. For example a SHAP explanation may indicate that a feature like marital status strongly affected a prediction. Still that doesn't fully answer broader questions researchers may actually care about. Questions such as whether the

model is learning the concept of marriage itself or whether certain social concepts influence predictions differently across demographic groups become much more important in sensitive applications. While going through different papers this issue appeared repeatedly especially in fairness-related studies.

The main idea behind TCAV is fairly straightforward at first glance. The method checks whether a user-defined concept consistently influences model predictions. It does this by learning something called a Concept Activation Vector or CAV. The CAV separates concept-positive examples from concept-negative examples inside the model representation space. In simpler terms the method tries to understand whether moving in the direction of a particular concept changes the model prediction. If the prediction score increases while moving toward that concept direction then the concept is considered influential for that prediction class. At first this looked reliable. But after deeper analysis the limitation became obvious. Human-defined concepts themselves can sometimes become difficult to represent consistently across datasets.

In the tabular classification setting used in this work TCAV is slightly modified so that it can operate directly on feature space. Concepts are defined using Boolean conditions over feature values. For example the concept of "high education" may be represented using instances where education_num is greater than 12. A binary Support Vector Machine is then trained to separate high-education samples from randomly selected non-high-education samples. The normal vector of the resulting decision boundary becomes the Concept Activation Vector for that concept. Later TCAV scores are calculated by projecting feature gradients of test samples onto this concept direction. The process sounds mathematically clean. The model still wasn't transparent in every situation though.

Research related to evaluating explanation methods has also grown rapidly after SHAP and LIME became popular. Different researchers proposed several ways to measure explanation quality and faithfulness. Samek and colleagues introduced deletion-based evaluation methods where features are removed gradually according to explanation importance scores and the resulting drop in model performance is measured. Later Yeh and other researchers proposed the ROAR and ROAD evaluation benchmarks mainly for image classification tasks. Their work focused on retraining-based evaluation approaches because standard deletion methods sometimes produce misleading results when modified inputs fall outside the original data distribution. This created another issue. Evaluating explanation quality itself became a difficult research problem.

One area that received serious attention was stability analysis especially for LIME explanations. Alvarez-Melis and Jaakkola showed that LIME explanations can change significantly even after very small input perturbations. Researchers noticed this repeatedly. That becomes a major concern in compliance-sensitive environments where explanations are expected to remain stable and reliable. Later another important issue was discussed by Slack and colleagues. They demonstrated that explanation methods like SHAP and LIME can sometimes be manipulated by adversarial classifiers designed to appear fair under explanation analysis while still behaving unfairly on actual population data. What makes this more difficult is that post-hoc explanation methods may create a false sense of trust if they are used without additional verification techniques. Because of this several studies recommend combining multiple explanation methods instead of depending entirely on a single approach for auditing or fairness evaluation.

G. The Post-Hoc Explainability Paradigm

As machine learning systems became more complicated researchers started facing another problem. Accuracy improved a lot. Transparency didn't. Many modern AI models especially deep learning systems and ensemble approaches can generate highly accurate predictions but the reasoning behind those predictions often remains unclear to users. This created another issue. People could trust the output only to a certain extent because understanding the internal logic of the system was difficult.

To deal with this limitation researchers introduced post-hoc explainability methods. These techniques are designed to explain predictions after the model has already been trained instead of modifying the original structure itself. The basic idea is fairly straightforward. Instead of examining every internal component of the model post-hoc approaches mainly observe the relationship between inputs and outputs to estimate how predictions are being generated. Because of this they can usually work with many different machine

learning systems regardless of architecture. Researchers often describe them as model-agnostic explanation methods.

One reason these approaches became popular is flexibility. The same explanation framework can often be applied to neural networks random forests support vector machines and gradient-boosting systems without redesigning the model itself. Organizations don't need to sacrifice prediction performance just to improve interpretability. While going through different papers this point appeared repeatedly. Researchers wanted explainability but they also didn't want to reduce model accuracy too much. Post-hoc explainability became one practical compromise between these two goals.

Among the earliest and most widely discussed post-hoc methods is LIME which stands for Local Interpretable Model-Agnostic Explanations. LIME was introduced mainly to explain individual predictions instead of explaining the entire model globally. That distinction matters. In many practical situations users are more interested in understanding why one specific prediction happened rather than studying the whole model structure. LIME works by slightly modifying or perturbing the original input multiple times and observing how prediction outputs change for nearby samples. Using those generated samples the method builds a simpler interpretable model around the prediction being analyzed. Usually a linear approximation is used because it is easier for humans to understand.

At first this looked reliable. LIME explanations were relatively intuitive and much easier to interpret compared to raw neural network outputs. In healthcare applications for example LIME can highlight symptoms or image regions influencing a disease prediction. In text classification systems it can identify words that strongly affect sentiment or topic classification. But after deeper analysis the limitation became obvious. Since LIME depends heavily on random sampling nearby explanations can change when the process is repeated several times. Researchers noticed this repeatedly. The explanations sometimes lacked stability especially in sensitive applications where consistency becomes important. The model still wasn't transparent in a complete sense.

Another important development in post-hoc explainability came through SHAP which stands for SHapley Additive exPlanations. SHAP was introduced partly to overcome consistency and theoretical limitations found in earlier explanation methods. The method borrows ideas from cooperative game theory particularly Shapley values originally developed for distributing rewards fairly among players in a game. In machine learning each feature is treated like a participant contributing toward the final prediction and SHAP calculates how much each feature contributes to the result.

One reason SHAP became highly respected in explainable AI research is because of its mathematical consistency. Features contributing equally receive similar importance values while irrelevant features receive near-zero contribution scores. SHAP explanations are also designed so that the sum of all feature contributions aligns closely with the final model prediction. This was discussed in several papers I reviewed. Compared to many earlier approaches SHAP generally appeared more stable and theoretically grounded. Though even SHAP isn't perfect in every situation and computation costs can become high for large models.

H. Global Versus Local Explanations

Explainability methods are often divided into two broad categories depending on the level at which explanations are generated. Some methods try to explain overall model behaviour across the dataset while others focus only on individual predictions. These are commonly referred to as global explanations and local explanations. The difference sounds simple initially. But while reviewing different studies it became clear that the distinction is actually very important in practical AI systems.

Global explanation methods attempt to understand how the model behaves in general. Researchers use them to identify which features are usually important across the entire dataset and how input variables influence predictions on average. Permutation feature importance is one commonly used example. This method works by randomly shuffling feature values and measuring how much prediction performance changes afterward. If accuracy drops heavily after shuffling a feature that feature is considered important for the model. Partial Dependence Plots or PDPs are also widely used because they help visualize relationships between input variables and prediction outputs while averaging effects from other features.

But PDPs also created another issue. When features are strongly correlated the explanations can become misleading. Researchers later introduced Accumulated Local Effects plots mainly to address this limitation. ALE plots focus more carefully on conditional feature relationships instead of broad averaging. Because of this they are generally considered more reliable for datasets where variables influence one another heavily. This issue becomes more visible in healthcare datasets where patient variables often remain highly correlated.

Local explanation methods focus on a completely different objective. Instead of studying overall model behaviour they attempt to explain why one specific prediction occurred. This becomes useful in practical decision-making situations. A customer whose loan application was rejected usually doesn't care about average model behaviour across thousands of applicants. They want to know why their own application failed. Similar situations appear in healthcare where doctors need to understand why an AI system predicted a disease for one particular patient. That's where the real concern starts. Explanations become meaningful only when users can connect them directly to their own decisions.

Methods such as LIME and SHAP are widely used for local explanations because they identify which input features contributed most strongly toward an individual prediction. Local explanations are often easier for end users to understand since they directly relate to a specific case. At the same time local and global explanations are not completely separate ideas. When many local explanations are aggregated together broader patterns about overall model behaviour can also emerge. For example averaging SHAP values across multiple predictions can provide estimates of global feature importance. Researchers found this relationship especially useful because it bridges prediction-level reasoning with larger system-level understanding.

SHAP became particularly influential partly because it supports both local and global interpretability inside one framework. Individual predictions can be explained separately while broader trends can still be analyzed across the dataset. Some researchers considered this one of the major strengths of SHAP compared to earlier explanation methods. Overall both local and global explanation techniques remain important in explainable AI research because they address different kinds of transparency problems in machine learning systems.

I. Gradient-Based Attribution Methods

Deep learning models process information through many hidden layers and complicated mathematical operations. Because of this understanding their behaviour is often difficult even for researchers working directly with the systems. One major category of explainability techniques developed for neural networks is gradient-based attribution methods. These approaches analyze how small changes in input values affect final prediction outputs. Since neural networks are differentiable systems gradients can provide useful information about which features influence predictions most strongly.

One of the simplest early approaches was vanilla gradient attribution. In this method gradients are calculated between the output and the input features to estimate sensitivity levels. Features with larger gradient values are treated as more important because small changes in those features produce stronger effects on predictions. Researchers first used this idea mainly in image classification tasks where gradients were visualized as saliency maps. These maps highlighted image regions influencing model predictions. For example if a network identified a dog in an image the saliency map could reveal which parts of the image the model focused on.

At first this seemed useful. But after deeper analysis another limitation appeared. Saliency maps often contained noisy and unclear patterns making interpretation difficult in several cases. Researchers noticed this repeatedly.

To improve reliability researchers later introduced Integrated Gradients. Instead of calculating gradients only at the final input point Integrated Gradients gradually move from a baseline input toward the actual input while accumulating gradient information across the path. This process usually generates smoother and more stable explanations compared to simple gradient methods. The method also follows certain theoretical principles designed to improve consistency. Equivalent neural networks should ideally produce similar explanations even when internal structures differ.

What makes this more difficult is baseline selection. For tabular datasets the baseline may simply be a vector of zeros. In image applications researchers

often use a completely black image as the reference point. But the choice of baseline strongly influences final explanations. So selecting inappropriate reference inputs can sometimes produce misleading interpretations.

Grad-CAM was mainly designed for convolutional neural networks used in computer vision tasks. The method generates visual heatmaps showing image regions most responsible for a prediction. Instead of focusing on individual pixels Grad-CAM analyzes feature maps from deeper convolutional layers and combines them with gradient information to identify important spatial regions. This becomes especially useful in healthcare systems. If an AI model predicts pneumonia from an X-ray image doctors naturally want to know which lung regions influenced the diagnosis. Grad-CAM heatmaps help visualize those areas making predictions easier to verify manually. Similar applications appear in autonomous driving object recognition and surveillance systems. Humans generally understand visual explanations more easily compared to mathematical feature scores.

Over time researchers proposed improved versions such as Grad-CAM++ and Score-CAM to address weaknesses in the original method especially when multiple objects appear within the same image. Some newer variants also reduce dependency on unstable gradient calculations improving explanation quality and localization accuracy. Still gradient-based methods continue facing problems related to sensitivity noise and dependence on model architecture.

J. Rule Based Explanations

As explainable AI research progressed researchers also started focusing on methods ordinary users could understand more easily. Mathematical explanations and visualization methods were useful for technical experts but not always practical for non-technical users. This created another issue. Explanations needed to become more human-readable. That led to increased interest in rule-based explanations and counterfactual explanations.

One important rule-based method is Anchor explanations. Anchor explanations were introduced partly as an improvement over earlier local explanation techniques like LIME. Instead of generating weighted feature importance values the method produces simple “if-then” rules describing conditions under which predictions remain stable. The idea is to identify a small group of influential feature conditions that strongly determine the prediction outcome.

For example in a loan approval system an anchor explanation might say that if the applicant’s age is above a certain threshold and debt ratio remains high then the application is likely to be rejected. These explanations resemble human reasoning patterns much more closely compared to mathematical plots or feature scores. Because of this they became useful in business healthcare finance and legal applications where decisions often need to be explained clearly to ordinary users regulators or stakeholders.

Anchor explanations are mainly evaluated using precision and coverage. Precision measures how reliably the rule predicts the same outcome while coverage measures how broadly the rule applies across different instances. There is usually a trade-off between these two factors. Highly specific rules may produce accurate explanations but apply only to limited situations whereas broader rules cover more cases but may lose precision.

Another important explainability approach is counterfactual explanation. Instead of explaining why a prediction occurred counterfactual methods explain how the outcome could have changed. In simple terms they answer “what if” questions. If a loan application was rejected the explanation may indicate that approval would have been possible if annual income were slightly higher or debt ratio lower. In healthcare systems counterfactual explanations can indicate how changes in medical indicators may alter disease predictions. These explanations feel more practical because they provide actionable guidance instead of only describing model behaviour.

Counterfactual explanations are usually generated by searching for alternative inputs that remain close to the original data while still crossing the model’s decision boundary. The objective is to identify the smallest possible modification needed to produce a different prediction outcome. This process is commonly treated as an optimization problem.

One reason counterfactual explanations became highly important is their connection with fairness and regulatory compliance. Modern regulations such as GDPR emphasize understandable automated decision-making systems and many researchers argue that counterfactual explanations align well with these

requirements because they help users challenge decisions and understand what changes could influence future outcomes. This was discussed in several papers I reviewed. The explanations don't just describe decisions. They also suggest possible paths toward different results.

Overall both Anchor explanations and counterfactual explanations contribute strongly toward improving AI transparency. Anchor methods simplify model behaviour into understandable rules while counterfactual approaches provide practical guidance for changing outcomes. Different methods solve different transparency problems though. And no single explainability technique completely removes every challenge connected with modern AI systems.

III. METHODOLOGY

A. Experimental Pipeline Overview

The experimental methodology used in this work was organized as a multi-stage pipeline based on the proposed Explainability Compliance Pipeline. Instead of treating explanation analysis as a single isolated step the process was divided into several connected stages so that the overall workflow remained easier to evaluate and reproduce later. The first stage mainly focused on dataset preparation. This included cleaning the raw data handling preprocessing operations and examining demographic distributions present in the datasets. At first this looked straightforward. But after deeper analysis the limitation became obvious because demographic imbalance in some datasets affected later explanation results as well. Researchers noticed this repeatedly in earlier studies too.

The second stage concentrated on training baseline machine learning models and evaluating their predictive performance. Different models were trained before explanation generation started because explanation quality often depends heavily on model behaviour itself. This created another issue. A highly accurate model doesn't always produce explanations that are easy for humans to interpret. While going through different papers this point appeared quite frequently especially in healthcare-related datasets where prediction accuracy alone wasn't considered sufficient.

B. Dataset Selection and Characteristics

B.1 UCI Adult Income Dataset

The Adult Income dataset was originally contributed to the UCI Machine Learning Repository by Ron Kohavi and was derived from the 1994 US Census database. It contains information related to income prediction and is one of the most commonly used datasets in fairness and explainability research. Initially the dataset had 48,842 records. After removing entries with missing values the final dataset used in this work contained 45,222 instances.

The dataset includes 14 different features such as age, workclass, education, marital status, occupation, relationship, race, sex, hours-per-week and native-country. Some of these attributes are numerical while others contain multiple categorical levels. The prediction target is binary and indicates whether a person earns more than $50,000 annually.

What makes this dataset especially useful is the presence of strong demographic imbalance patterns. The positive income rate differs considerably between male and female groups. Similar disparities are visible across racial categories as well. While going through different fairness studies this issue appeared repeatedly. Researchers often use this dataset to examine how machine learning systems may unintentionally learn historical social inequalities from data itself.

One thing that became noticeable during analysis was the relationship between marital status sex and income. Certain marital-status categories strongly correlate with both income level and gender. This creates another issue. A machine learning model may learn proxy relationships without directly using protected features in an obvious way. At first the model looked statistically reliable. But after deeper analysis the limitation became more visible. Explainability methods therefore become important because they help identify whether the system is indirectly depending on discriminatory proxy features.

B.2 COMPAS Recidivism Dataset

The COMPAS Recidivism dataset was released publicly through the ProPublica investigation on algorithmic bias conducted in 2016. The dataset contains criminal justice records from Broward County Florida and focuses mainly on recidivism risk assessment. After filtering missing entries restricting the analysis to a two-year follow-up period and selecting Black and white defendants for comparison the final working dataset contained 5,278 records.

Features in the dataset include age category race sex prior conviction counts juvenile offenses and charge degree. The target variable indicates whether a defendant was rearrested within two years. The dataset is moderately balanced with around 54.5% negative outcomes and 45.5% positive outcomes.

This dataset is discussed frequently in fairness research because of the racial disparities identified in the original ProPublica analysis. Black defendants were assigned higher risk scores much more often compared to white defendants with comparable outcomes. Researchers noticed this repeatedly across multiple studies. The problem here isn't only the machine learning model itself. The underlying historical data already contains structural bias patterns. That's where the real concern starts.

While reviewing papers related to algorithmic fairness it became clear that COMPAS is often treated as one of the main benchmark datasets for explainability and fairness analysis. Models trained on such data may appear accurate from a statistical perspective but can still reinforce social inequality through biased predictions. And this can't be ignored.

B.3 Pima Indians Diabetes Dataset

It contains medical records of 768 female patients aged 21 years or older belonging to Pima Indian heritage.

The dataset includes features such as plasma glucose concentration blood pressure BMI insulin level age diabetes pedigree function and number of pregnancies. The target variable indicates whether the patient was diagnosed with diabetes according to WHO criteria. Around 65.1% of records belong to the non-diabetic category while the remaining instances represent positive diabetes cases.

This issue becomes more visible in healthcare datasets. Some medical attributes already have clinically established importance. Glucose concentration BMI and diabetes pedigree function are known diabetes risk indicators according to medical research. Because of this the dataset becomes useful for checking whether explanation methods identify medically meaningful features correctly. If an explainability technique highlights unrelated attributes while ignoring clinically important ones the reliability of that explanation becomes questionable.

C. Classifier Configurations and Training Protocol

Three machine learning classifiers were trained on each dataset in order to study how model complexity influences explanation quality. Logistic Regression was selected as the interpretable baseline model because its feature weights can be directly interpreted as contributions toward prediction outcomes. L2 regularization with C=1.0 was applied and optimization was performed using the LBFGS solver with a maximum of 1000 iterations.

The model used 100 decision trees with depth set to 10 and minimum samples per leaf equal to 5. Bootstrap sampling was also enabled.

While going through previous explainability studies XGBoost appeared repeatedly because of its balance between predictive performance and compatibility with explanation methods such as TreeSHAP.

All models were trained using stratified 80/20 train-test splitting. The stratification considered both the target class and protected attributes so that demographic representation remained balanced across train and test partitions. Hyperparameters were selected using five-fold cross-validation based on AUC-ROC performance. Cross-validation was used mainly for tuning and confidence interval estimation while final evaluation was performed separately on the test set.

D. Explanation Generation Protocol

SHAP explanations were generated differently depending on model type. TreeSHAP was used for Random Forest and XGBoost because it provides exact Shapley value computation efficiently for tree-based architectures. Logistic Regression explanations used KernelSHAP with 1000 background samples since it supports model-agnostic approximation. Global SHAP importance was calculated using the mean absolute SHAP value across all test instances. Subgroup-based analysis was also performed separately for protected attribute categories.

LIME explanations were generated using LimeTabularExplainer with 500 samples and a maximum of 10 features for each explanation. Explanations were created for 400 test instances selected through stratified sampling. Internal stability was checked by generating multiple explanations for the same instances using different random seeds and then measuring Spearman rank correlation between feature rankings.

This created another issue. LIME explanations sometimes varied depending on random sampling conditions. At first the method appeared intuitive and easy to understand. But after deeper analysis stability limitations became more noticeable especially when explanations were generated repeatedly for similar instances.

TCAV analysis used manually defined human concepts based on feature thresholds. For the Adult Income dataset concepts included high education male sex white race and high working hours. COMPAS concepts focused on prior convictions felony charge young defendants and racial categories. The Diabetes dataset used concepts such as high glucose high BMI older patients and strong family history. These concept-based explanations were useful because they connected machine learning behaviour with more understandable human-level concepts instead of raw feature importance values.

E. Explanation Quality Evaluation Protocol

Different evaluation metrics were used to measure explanation quality. Faithfulness was evaluated using deletion-AUC analysis where features were removed based on attribution importance and the resulting prediction change was measured after each removal. Lower deletion-AUC values indicated stronger faithfulness because prediction confidence dropped faster when important features were removed.

Comprehensibility was measured using the minimum number of features required to explain 80% of total attribution contribution for each instance. Completeness measured how much of the prediction deviation could be explained by the top attributed features. Cross-method consistency was evaluated using Spearman rank correlation between SHAP and LIME feature rankings for matched instances.

While reviewing evaluation approaches one thing became clear. Measuring explanation quality isn't as straightforward as measuring prediction accuracy. Different metrics capture different aspects of interpretability and no single evaluation method completely solves the problem. Researchers noticed this repeatedly across explainability literature.

IV. TOOLKIT COMPARISON AND ANALYSIS

A. Scope and Coverage Analysis

A comparison of SHAP LIME and TCAV showed that each method occupies a different position within the explainability landscape. SHAP provides broad explanation coverage including global feature importance local instance explanations feature interaction analysis and subgroup-level interpretation. The SHAP library supports multiple visualization interfaces such as summary plots force plots and dependence plots. TreeSHAP also performs exact computations efficiently for tree-based models while KernelSHAP supports model-agnostic explanation generation at higher computational cost.

LIME focuses more strongly on local explanations. One reason it became popular is because the explanations are usually easier for non-technical users to understand. LIME also supports different data modalities including text image and tabular data. But the method has limitations too. It does not naturally provide global summaries and explanation outputs may vary across runs because of stochastic sampling behaviour. The model still wasn't fully transparent.

TCAV approaches explanation from a different direction altogether. Instead of asking how much a feature contributed toward a prediction TCAV focuses on whether the model learned specific human-understandable concepts. Questions like "Does the model associate race with prediction outcome?" or "Does the model rely heavily on prior incarceration?" fit naturally within TCAV analysis. What makes this more difficult is that domain experts must define concepts manually before analysis begins. Still this requirement also makes TCAV particularly useful for compliance auditing and fairness investigations because the explanations directly address meaningful human-level concerns.

Over time several explainability techniques became widely used in machine learning research. SHAP is one of the most commonly discussed methods because it estimates feature contribution using concepts derived from game theory. LIME focuses more on explaining individual predictions through simpler local models. In deep learning applications visualization-based techniques such as Grad-CAM and Integrated Gradients are widely used especially in healthcare and computer vision tasks. Grad-CAM highlights image regions influencing predictions while Integrated Gradients measure feature importance by observing changes in neural network outputs. Attention-based explanation approaches are also

heavily applied in transformer architectures and natural language processing systems because they help identify which words or input regions received stronger focus during prediction generation. Some techniques are easier to interpret than others though. And different methods solve different kinds of problems

B. Comprehensive Toolkit Comparison

Table 1: Systematic Comparison of SHAP, LIME, and TCAV Explainability Toolkits

Evaluation Dimension	SHAP	LIME	TCAV
Global Explanation	Yes — native summary plots	Aggregation required	Concept-level scores only
Local (Instance) Explanation	Yes — exact Shapley values	Yes — primary mode	Fraction-of-instances score
Feature Interaction Analysis	Yes — interaction values	No	No
Concept-Level Explanation	No	No	Yes — primary mode
Model Agnostic	Partial (KernelSHAP)	Yes — fully agnostic	Partial (gradient-based)
Tree Model Support	Yes — TreeSHAP exact	Yes — indirect	No
Neural Network Support	Yes — DeepSHAP	Yes	Yes — native
Faithfulness Guarantee	Axiomatic (efficiency axiom)	Approximate only	Empirical p-values
Internal Stability	High — deterministic	Low — stochastic	Moderate
Completeness	Yes — efficiency axiom	Partial (sparse approx.)	N/A
Computation Cost (tabular)	Low (TreeSHAP)	Moderate–High	High (SVM per concept)
Sklearn Compatibility	Yes — native	Yes — native	Custom wrapper required
Visualization Quality	Comprehensive built-in	Basic built-in	Limited
Non-Technical Accessibility	Moderate	High (linear coefficients)	Low (requires expertise)
Demographic Equity Analysis	Yes — subgroup SHAP	Manual aggregation	Concept TCAV by group
GDPR Art. 22 Alignment	Global only	Best — per-instance	Hypothesis testing
EU AI Act Art. 13 Alignment	Strong — comprehensive	Moderate	Supplementary
License	MIT	BSD-2-Clause	Apache 2.0

C. Computational Cost and Scalability

Computational cost also becomes an important factor while selecting explainability methods for practical deployment. Accuracy and interpretability alone aren't enough in real production systems because execution time and scalability can quickly become limitations. While going through different papers this issue

appeared repeatedly especially in large-scale audit environments. Some explanation methods work well theoretically but become difficult to apply when the dataset size increases.

TreeSHAP performed comparatively well during evaluation. Its exact computation method was able to explain all 45,222 test instances from the Adult Income dataset in nearly 3.2 minutes using a standard 8-core CPU system. At first this looked reliable. The computation speed makes full test-set explanation possible for batch compliance auditing and large-scale evaluation tasks. This becomes useful in practical deployments where organizations may need explanations for thousands of predictions instead of only a few selected examples.

KernelSHAP behaved differently though. Since it follows a model-agnostic approach the computational cost increases heavily because of coalition sampling requirements. Explaining only 500 instances of the logistic regression model required almost 12 minutes during experimentation. This created another issue. Large-scale audits become difficult when non-tree models are involved. But after deeper analysis the limitation became obvious. Practitioners working with bigger datasets may need approximate SHAP implementations or sampling-based approaches to reduce the number of coalition evaluations and improve execution time. Researchers noticed this repeatedly in performance-focused studies.

LIME also showed practical limitations during full dataset evaluation. The method required around 0.8 seconds for each individual instance when 500 perturbation samples were generated. That may not seem very large initially. But for the complete Adult Income test dataset containing 45,222 instances the total runtime extended close to 10 hours. And this can't be ignored. Such execution time becomes unrealistic for routine compliance auditing or large production workflows. What makes this more difficult is that organizations often require explanations within shorter operational timeframes. Because of this LIME appears more suitable for targeted analysis of specific high-priority cases such as disputed decisions audit samples or regulatory investigations instead of explaining every prediction in the dataset TCAV introduced a different type of computational overhead. The method required training one binary SVM model separately for every concept used during analysis. This added nearly 8 minutes per concept during the audit workflow. The model still wasn't lightweight. However once the CAV models were trained the actual TCAV score computation became relatively fast and required less than one minute for each concept across all test instances. Some methods clearly trade training time for faster inference stages. This was discussed in several papers I reviewed while comparing explanation techniques for scalable deployment scenarios.

D. Domain-Indexed Recommendation Matrix

Table 2: Domain-Indexed Explainability Compliance Recommendation Matrix

Domain	Primary Explainability Need	Priority Toolkit	Priority Metric	Regulatory Driver
Healthcare Diagnostics	Clinician-accessible feature attribution; false-negative parity; calibration validation	SHAP (global) + TCAV (clinical concepts)	Faithfulness, Completeness	EU AI Act Art. 13; FDA AI/ML guidance
Financial Services / Credit	Per-applicant right-to-explanation; proxy discrimination diagnosis; audit trail	SHAP + LIME (per-instance)	Faithfulness, Stability	GDPR Art. 22; ECOA; EU AI Act
Criminal Justice / Bail	Defendant-accessible explanation; false positive mechanism; legal contestation	LIME (linear output) + SHAP global	Comprehensibility, Consistency	Constitutional due process; EU AI Act high-risk

Domain	Primary Explainability Need	Priority Toolkit	Priority Metric	Regulatory Driver
Human Resources / Hiring	Candidate-level explanation; disparate impact diagnosis; GDPR compliance	SHAP global + LIME per-candidate	Stability, Completeness	GDPR Art. 22; Title VII; EU AI Act

V. IMPLEMENTATION AND EXPERIMENTS

A. Baseline Classifier Performance

Table 3: Baseline Classifier Performance Across All Datasets

Dataset / Model	Accuracy	AUC-ROC	F1-Score	Bal. Acc.
Adult / LR	0.823	0.891	0.671	0.784
Adult / RF	0.856	0.921	0.714	0.812
Adult / XGB	0.871	0.933	0.739	0.826
COMPAS / LR	0.668	0.722	0.661	0.671
COMPAS / RF	0.692	0.751	0.688	0.694
COMPAS / XGB	0.706	0.769	0.702	0.708
Diabetes / LR	0.774	0.841	0.723	0.769
Diabetes / RF	0.797	0.862	0.748	0.791
Diabetes / XGB	0.808	0.875	0.761	0.803

The baseline experimental results mostly followed the general trend that was expected before training the models. Among all the classification approaches used in this work XGBoost consistently produced the highest AUC-ROC scores across the three datasets. Random Forest usually came after that while Logistic Regression showed lower performance in most experiments. At first this looked fairly normal because ensemble models often perform better when datasets contain more complex feature interactions. This was discussed in several papers I reviewed as well.

But after spending more time comparing the outputs some dataset-specific differences became easier to notice. The Adult Income dataset showed a much larger performance gap between the models compared to the other datasets. One reason could be the structure of the dataset itself. It contains a larger number of records and the relationships between features are not always simple or linear. Because of that XGBoost was able to capture more complicated interactions between variables much more effectively. Logistic Regression struggled more here. The model mainly works better when relationships remain closer to linear patterns and that limitation became visible during evaluation.

What makes this more interesting is that Random Forest still performed reasonably well even though it didn't reach the same results as XGBoost. Researchers noticed this repeatedly in ensemble-based learning methods. Random Forest handled the feature complexity better than Logistic Regression but still couldn't optimize decision boundaries as efficiently as boosting-based approaches. This created another issue. The performance differences became more obvious when the dataset size increased and feature dependency patterns became less straightforward.

B. Explanation Quality Metric Results

Table 4: Explanation Quality Metrics by Toolkit, Dataset, and Model

Dataset / Model / Toolkit	Faithfulness↑	Stability↑	Comprehensibility	Completeness↑
Adult / LR / SHAP (KernelSHAP)	0.812	1.000	4.1 features	0.91
Adult / LR / LIME	0.761	0.743	5.8 features	0.74
Adult / RF / SHAP (TreeSHAP)	0.849	1.000	5.3 features	0.88
Adult / RF / LIME	0.718	0.702	6.4 features	0.71
Adult / XGB / SHAP (TreeSHAP)	0.863	1.000	5.9 features	0.89
Adult / XGB / LIME	0.699	0.671	7.2 features	0.68
COMPAS / RF / SHAP	0.831	1.000	4.6 features	0.87
COMPAS / RF / LIME	0.698	0.683	6.7 features	0.70
COMPAS / XGB / SHAP	0.841	1.000	4.8 features	0.86
COMPAS / XGB / LIME	0.712	0.673	6.9 features	0.69
Diabetes / RF / SHAP	0.844	1.000	4.2 features	0.90
Diabetes / RF / LIME	0.709	0.694	5.8 features	0.72
Diabetes / XGB / SHAP	0.857	1.000	4.4 features	0.91
Diabetes / XGB / LIME	0.724	0.698	6.1 features	0.72

Note: ↑ indicates higher values are better. Stability = 1.000 for SHAP reflects its deterministic computation (no random seed variance). LIME stability values represent mean Spearman rank correlation across five independent runs per instance.

The explanation quality results showed some clear patterns during the experiments. Across almost all dataset and model combinations SHAP performed better than LIME in terms of faithfulness. The difference was smaller in simpler models like Logistic Regression on the Adult dataset where SHAP achieved 0.812 while LIME reached 0.761. But the gap increased noticeably for more complex models. For example in the Adult XGBoost model SHAP reached 0.863 whereas LIME dropped to 0.699. At first this looked like a small variation between methods. But after deeper analysis the pattern became much more obvious. As model complexity increased SHAP explanations generally became more faithful while LIME explanations started losing consistency and reliability. Researchers noticed this repeatedly in several comparative studies as well.

What makes this more difficult is that explanation quality doesn't behave the same way for every method. SHAP remained stable even for highly complex models because TreeSHAP computes explanations exactly rather than approximately. This gave SHAP a stability score of 1.000 across repeated runs. The outputs didn't really change. LIME behaved differently. Its stability values usually stayed somewhere between 0.671 and 0.743 depending on the dataset and model complexity. The instability became more visible in larger datasets and complex architectures. Adult Income with XGBoost was one example where LIME achieved only r = 0.671 cross-run consistency. This created another issue. The local linear approximation used by LIME becomes unstable precisely in regions where the underlying model itself is already highly complex and difficult to interpret. $r = 0.671$

C. Demographic Equity of Explanation Quality

Table 5: Cross-Method Consistency (SHAP–LIME Spearman r) by Demographic Group

Dataset / Model	Overall r	Privileged Group r	Unprivileged Group r	Gap
Adult / XGB (sex)	0.712	0.741 (male)	0.661 (female)	0.080
Adult / XGB (race)	0.698	0.729 (white)	0.644 (non-white)	0.085
Adult / RF (sex)	0.728	0.754 (male)	0.681 (female)	0.073
COMPAS / XGB (race)	0.688	0.722 (Caucasian)	0.641 (Afr.-American)	0.081
COMPAS / RF (race)	0.701	0.733 (Caucasian)	0.652 (Afr.-American)	0.081
Diabetes / XGB (age)	0.734	0.748 (older)	0.712 (young)	0.036

For the Adult Income XGBoost model female instances showed lower agreement between SHAP and LIME explanations compared to male instances. The mean Spearman correlation value for female samples was around 0.661 while male samples received approximately 0.741. A similar trend appeared across racial groups as well. Non-white instances produced an agreement score near 0.644 whereas white instances reached around 0.729. At first this looked like a small statistical variation. But after deeper analysis the limitation became more obvious. The difference remained consistent across multiple evaluations.

African-American instances tend to appear more frequently in regions of the feature space where proxy relationships become stronger. In the Adult Income dataset this mainly involves the interaction between marital status sex and income-related features. In the COMPAS dataset similar behaviour appears around prior convictions race and recidivism prediction variables. The model still wasn't transparent.

In these feature regions the decision boundary becomes much more nonlinear and unstable. Because of that LIME's local linear approximation struggles to represent the model behaviour accurately while SHAP captures feature contributions differently. This increases disagreement between both explanation methods. And this can't be ignored. The inconsistency itself becomes another fairness concern because certain demographic groups end up receiving explanations that are less reliable than others. This created another issue.

One thing that became clear during this analysis is that reporting only aggregate explanation quality metrics may hide important disparities across groups. If explanation quality is averaged across the entire dataset the weaker explanation consistency observed for minority or underrepresented populations may never become visible. That's where the real concern starts. Explanation quality metrics probably need to be reported separately for different demographic groups instead of only presenting one combined score for the whole test set. Otherwise explanation inequity can remain hidden even when the overall model performance appears acceptable.

D. SHAP Global Attribution Analysis

D.1 Adult Income Dataset — Proxy Discrimination Detection

The SHAP TreeExplainer analysis on the Adult Income XGBoost model produced a global ranking of important features. Since the sex feature appeared relatively low in the ranking it could easily give the impression that gender bias inside the model was limited. But after deeper analysis the limitation became obvious. Global importance rankings alone don't properly reveal hidden discrimination patterns.

When subgroup-based SHAP analysis was performed separately for male and female instances a very different pattern appeared. For male samples the marital_status_Married-civ-spouse feature had a positive mean SHAP contribution of +0.203. In simple terms being married increased predicted income strongly for men. But for female instances the same feature produced an average SHAP value of -0.189. So the exact same variable reduced predicted income for women. Researchers noticed this repeatedly in fairness-related studies. The direction of feature influence completely reversed depending on gender.

This result is important because it shows how proxy discrimination actually works inside machine learning models. The system learned relationships from historical labour-market patterns present in the dataset. During the 1994 period represented by the Adult Income data married men were more likely to be primary earners while married women were more frequently associated with unpaid domestic roles. The model absorbed this historical imbalance and encoded it into its prediction logic. The bias wasn't directly caused only by the sex feature itself. Instead the interaction between sex and marital status became the mechanism carrying discriminatory behaviour. The model still wasn't transparent.

What makes this more difficult is that traditional fairness metrics usually fail to expose these interactions clearly. A demographic parity score may show that women receive fewer positive income predictions overall but it doesn't identify which features actually create the disparity. SHAP subgroup analysis makes the mechanism more visible. While going through different papers this issue appeared repeatedly. Attribution analysis points directly toward marital status as one of the strongest contributors driving the imbalance which also suggests possible mitigation strategies such as reducing feature dependency through preprocessing or regularization methods.

D.2 COMPAS Dataset — Feature Weighting Asymmetry. Because of this some earlier studies argued that COMPAS models are effectively race-neutral once race variables are removed. But the subgroup SHAP analysis suggested something else entirely.

For African-American defendants the average SHAP contribution of prior_count was +0.289 for each conviction unit. For Caucasian defendants with similar conviction counts the corresponding value was lower at +0.231. This difference is not small. It means identical prior conviction histories influenced recidivism predictions more aggressively for Black defendants compared to white defendants. That's where the real concern starts.

The model appears to have learned statistical relationships shaped by unequal policing and prosecution practices already present in the training data. In Broward County arrest patterns historically differed across racial groups. So the algorithm learned stronger associations between prior arrests and future recidivism risk for Black defendants because those communities experienced heavier policing exposure in the underlying dataset. The discrimination therefore wasn't simply tied to the explicit race variable. It became embedded within how other features were weighted and interpreted internally by the model.

This created another issue.

Removing race from a machine learning model doesn't automatically eliminate racial bias if correlated features still carry structural inequality information. While reviewing multiple papers this point appeared again and again. The bias remains encoded indirectly through feature relationships. SHAP analysis makes this hidden dependency more visible and easier to audit compared to standard scalar fairness metrics which usually provide only overall disparity values without identifying causal feature behaviour.

D.3 Diabetes Dataset — Clinical Validation

The SHAP analysis on the Diabetes XGBoost model identified glucose concentration as the most influential feature affecting prediction outcomes. Among all variables glucose produced the highest average SHAP value which indicates that it contributed most strongly to the model's decisions. Other important features included Body Mass Index (BMI), age, diabetes pedigree function, insulin level, and number of pregnancies. While reviewing the outputs the ranking actually aligned closely with established medical understanding. This was discussed in several healthcare-focused studies as well.

The dominance of glucose concentration makes practical sense because elevated blood sugar is already one of the primary indicators used in diabetes diagnosis. BMI also appeared highly important which is medically expected since obesity strongly increases type 2 diabetes risk. Age contributed noticeably too because metabolic disorders become more common over time. The diabetes pedigree function represents genetic predisposition and family history which are known diabetes risk factors. Insulin levels and pregnancy history were also relevant due to their relationship with insulin resistance and gestational diabetes.

This issue becomes more visible in healthcare datasets because explanations are expected to align with real medical knowledge not just mathematical accuracy. A model may achieve strong prediction performance

technically while still relying on meaningless correlations. For example some healthcare AI systems have accidentally learned to depend on image artifacts or demographic shortcuts rather than actual disease indicators. Even if prediction accuracy looks high such behaviour can become dangerous in clinical environments. And this can't be ignored.

The close agreement between SHAP explanations and medical understanding provided an important form of validation in this study. Explanations that match expert domain knowledge generally increase confidence in both the machine learning model and the explainability framework itself. Doctors and healthcare professionals need explanations that are medically meaningful not only statistically correct. Otherwise trust becomes difficult to establish.

Another important observation from the SHAP analysis involved feature interaction effects. SHAP interaction values showed the strongest pairwise interaction between glucose concentration and BMI. When both variables were simultaneously high diabetes risk predictions increased far more strongly than expected from considering each factor independently. In other words the combined impact of glucose and BMI became larger than the sum of their separate contributions.

At first this looked like only a statistical interaction. But after deeper analysis the relationship appeared clinically meaningful as well. Obesity and elevated glucose levels frequently occur together in metabolic syndrome and type 2 diabetes development. Their interaction is already recognized in medical literature as a strong risk combination associated with insulin resistance and cardiovascular complications. SHAP's ability to capture these joint effects provides deeper understanding of model behaviour compared to simpler explanation approaches.

Traditional feature importance methods usually evaluate variables independently and often miss interaction behaviour between features. LIME for example mainly focuses on local linear approximations around predictions and doesn't naturally capture complex pairwise dependencies. Standard scalar importance metrics also fail to explain how features work together. SHAP interaction values overcome part of this limitation by directly measuring joint feature contributions.

This becomes especially important in healthcare applications because many diseases are not caused by isolated variables alone. Biological genetic environmental and lifestyle factors often interact in complicated ways. Understanding those interactions helps researchers and healthcare professionals evaluate whether AI predictions reflect real clinical reasoning patterns or merely statistical coincidences. It also improves transparency because explanations can show not only which variables matter but how they influence each other during prediction generation.

Overall the Diabetes SHAP analysis demonstrated that explainable AI can support clinically meaningful interpretation of machine learning systems. The explanations aligned well with accepted medical understanding identified major risk factors and captured medically relevant feature interactions. Different explainability methods provide different strengths of course. But SHAP showed stronger capability for revealing both individual feature importance and interaction behaviour within healthcare-oriented prediction models.

E. LIME Instance-Level Analysis

E.1 Adult Income Dataset

For this part of the analysis LIME explanations were generated for 400 selected test instances from the Adult Income dataset. The objective was to observe how the model behaves at an individual prediction level rather than only looking at overall feature importance values. At first the results appeared fairly straightforward. But after examining the explanations more carefully some noticeable patterns started appearing.

Among the 100 female instances predicted as low income the feature sex_Male = 0, which basically represents the female category, appeared within the top three contributing features in 78 out of 100 cases. The mean absolute LIME coefficient observed for this feature was around 0.108. This is important because it suggests the model is not simply learning indirect correlations from surrounding variables. Instead it is actively using the sex attribute itself while generating low-income predictions for female instances. Researchers noticed this repeatedly while checking different explanations individually. The model still wasn't transparent.

For the 100 male instances predicted as high income the feature sex_Male = 1 appeared as a positive contributing factor in 72 of the 100 predictions with an average coefficient value of 0.061. The influence is

smaller compared to the female low-income category but it still shows that gender information plays a noticeable role during prediction generation. This created another issue.

Something else became visible while reviewing the female high-income predictions. In those cases the model behaved differently. Features such as education_num ranked first with a mean LIME coefficient of 0.201 followed by hours_per_week at 0.178 and occupation_Exec-managerial at 0.143. So for women predicted as high income the model relied more on education work intensity and professional occupation features rather than directly using gender as a dominant contributor.

But male high-income predictions showed another pattern. Here marital_status_Married-civ-spouse ranked first with a coefficient value of 0.189 even before education and occupation-related features. While going through different papers this kind of asymmetry appeared in multiple fairness studies. Female high-income predictions were linked more strongly with education and working hours whereas male high-income predictions received stronger influence from marital status. That difference becomes important while studying proxy discrimination mechanisms at the instance level. And this can't be ignored. The LIME explanations help reveal how feature influence changes between groups even when prediction outcomes appear similar on the surface.

E.2 COMPAS Dataset

The COMPAS dataset produced some of the most concerning results during the LIME-based analysis especially when examining false positive predictions. These are cases where defendants were predicted as high recidivism risk even though they did not actually reoffend later. At first the model seemed reliable because overall prediction accuracy looked acceptable. But after deeper analysis the limitation became obvious.

For 100 African-American false positive instances the average LIME coefficient associated with prior_count was measured at 0.198. In comparison matched Caucasian false positive instances with similar prior_count distributions showed a lower mean coefficient value of 0.152. The difference was around 30.3%. This means the model was applying the prior conviction signal more aggressively against African-American defendants even when the underlying feature values were comparable.

That's where the real concern starts. The same feature value was not influencing all groups equally.

Another pattern appeared while examining the age-related explanations. For African-American defendants younger than 25 years old within the false positive category the age feature had a mean LIME coefficient of -0.089. This indicates that younger age provided only limited protection against being classified as high risk. However for similarly young Caucasian defendants the coefficient value was -0.141 which represents noticeably stronger protection from high-risk classification.

F. TCAV Concept-Level Analysis Results

Table 6: TCAV Scores by Concept and Demographic Subgroup

Dataset / Concept	Overall TCAV Score	Privileged Group	Unprivileged Group	Interpretation
Adult: High Education	0.841	0.839 (male)	0.846 (female)	Strong positive — equitable across sex
Adult: Married Civ. Spouse	0.763	0.912 (male)	0.309 (female)	Critical asymmetry — proxy discrimination confirmed
Adult: High Hours Per Week	0.788	0.801 (male)	0.769 (female)	Moderately equitable positive effect
Adult: Male Sex	0.731	N/A	N/A	Protected characteristic drives 73% of predictions

Dataset / Concept	Overall TCAV Score	Privileged Group	Unprivileged Group	Interpretation
Adult: White Race	0.582	N/A	N/A	Moderate race effect on income prediction
COMPAS: High Prior Count	0.814	0.779 (Caucasian)	0.851 (Afr.-Amer.)	Stronger effect for Black defendants — confirmed asymmetry
COMPAS: Young Defendant	0.623	0.668 (Caucasian)	0.578 (Afr.-Amer.)	Youth less protective for Black defendants
COMPAS: Felony Charge	0.794	0.788 (Caucasian)	0.801 (Afr.-Amer.)	Near-equitable across racial groups
Diabetes: High Glucose	0.889	0.881 (older)	0.901 (younger)	Dominant predictor — clinically appropriate
Diabetes: High BMI	0.812	0.823 (older)	0.798 (younger)	Strong appropriate predictor

TCAV analysis added another layer of evidence to the discrimination patterns that were already visible through SHAP and LIME explanations. One observation stood out quite clearly during the comparison stage. The married civilian spouse concept produced a TCAV score of 0.912 for male instances while female instances received only 0.309. At first this looked like a normal statistical difference. But after spending more time analyzing the outputs the implication became harder to ignore. The model seemed to connect marriage with higher income much more strongly for men compared to women. In practical terms the system learned that marriage increases income probability mainly for male applicants. Researchers noticed this repeatedly while comparing explanation results across different techniques.

G. Evaluation Metrics

To compare explainable AI techniques fairly this research uses a multi-dimensional evaluation framework instead of relying on only one metric. While going through different papers it became clear that explanation quality cannot really be measured from a single perspective. Some methods generate highly detailed explanations but remain unstable. Others are simple but less faithful to actual model behaviour. Because of this the study evaluates explanation methods using multiple factors including faithfulness stability sparsity computational cost and explanation coverage.

The first evaluation factor is faithfulness. This measures whether the explanation actually identifies the features most responsible for the model prediction. To test this the research applies the ROAR approach which stands for Remove and Retrain. Features identified as highly important are removed or masked from the dataset and the model is retrained afterward. If the model performance drops heavily after removing those features it usually means the explanation correctly identified influential information. Larger performance degradation therefore suggests stronger explanation faithfulness. This was discussed in several studies I reviewed while comparing explainability benchmarks.

The second factor is stability. Stability measures whether explanations remain consistent when small modifications are introduced into the input data. In practical AI systems tiny changes shouldn't completely alter the explanation if the prediction itself remains mostly unchanged. To evaluate this the study compares explanations generated from slightly modified inputs using cosine similarity. Higher similarity values indicate more stable explanation behaviour under small perturbations. What makes this more difficult is that some explanation methods naturally include randomness during generation which causes slight variation across runs.

Sparsity is another important evaluation factor. A sparse explanation highlights only the most important features while assigning near-zero importance to less relevant ones. Explanations with fewer highlighted features are usually easier for humans to interpret because they reduce unnecessary information. In this research sparsity is measured using the proportion of features receiving near-zero attribution values. At first sparse explanations looked automatically better. But after deeper analysis the limitation became obvious. Extremely sparse explanations may oversimplify model behaviour and hide useful contextual information.

Computational cost is also included in the framework because explanation methods differ heavily in processing requirements. Some techniques produce detailed and reliable explanations but require high computation time and memory usage. This becomes important in large-scale AI systems and real-time applications where explanations need to be generated quickly. Computational cost is measured using total execution time under the same hardware settings for every method. Researchers noticed this repeatedly in SHAP-based evaluations since sampling-based approaches often require significantly more computation compared to gradient-based techniques.

Explanation coverage mainly applies to rule-based approaches such as Anchor explanations. Coverage measures how broadly a generated explanation rule applies across the dataset while still maintaining high precision. A rule with larger coverage can explain more predictions consistently which makes it more useful in practical deployment environments. But there is another issue here. Highly precise rules sometimes become too specific and end up applying only to a small subset of instances. This created another issue. Balancing rule precision and generalizability becomes difficult in complex datasets.

The study also includes qualitative evaluation for image-based explanation methods such as Grad-CAM and Integrated Gradients. This issue becomes more visible in healthcare datasets where visual explanations directly affect medical interpretation. Three domain experts including a radiologist a medical imaging researcher and a machine learning engineer independently reviewed the generated explanation maps. They evaluated whether highlighted image regions were clinically meaningful visually understandable and spatially consistent with expected disease patterns. Ratings were assigned using a five-point Likert scale.

To improve reliability of expert evaluation the level of agreement between reviewers was measured using Fleiss' Kappa. qualitative analysis provides a broader understanding of explainable AI performance across different datasets and application domains.

H. Comprehensive Comparison Table

The experimental workflow used in this study is divided into four major stages so that evaluation remains systematic and reproducible. Each stage focuses on a different part of the explainability pipeline beginning from dataset preparation and ending with comparison of explanation methods.

The first stage involves data preparation. During this stage all datasets are loaded and preprocessed according to their specific requirements. Missing value handling categorical feature encoding image resizing text tokenization and normalization procedures are performed before model training begins. After preprocessing the data is divided into training validation and testing sets using fixed random seeds. Reproducibility matters here because inconsistent experimental conditions can produce misleading comparison results later.

The second stage focuses on model training. The selected baseline models are trained using predefined hyperparameter settings already discussed earlier in the dissertation. XGBoost is trained for structured tabular analysis while ResNet-50 is fine-tuned for medical image classification tasks. BERT is also fine-tuned for sentiment analysis experiments. Training continues until model performance stabilizes and the final accuracy values are compared against published benchmark results. At first the models appeared stable. But repeated testing still showed small performance fluctuations across datasets.

The third stage involves explanation generation. After training all seven explainability methods are applied to generate explanations for test samples. To maintain fairness across evaluations a random stratified sample of 500 instances is selected from each dataset. Every explanation method generates outputs for these samples and the total execution time is recorded for efficiency comparison. Some techniques finished explanation generation quickly while others required considerably more time especially during repeated

runs involving stochastic sampling. The model still wasn't computationally efficient in some cases.

The fourth stage focuses on evaluation and comparison of generated explanations. All outputs are analyzed using the previously defined metrics including faithfulness stability sparsity computational cost and explanation coverage. Results from each explainability method are aggregated into comparative tables so strengths and weaknesses become easier to identify. Additional stability analysis is also performed for stochastic techniques such as LIME and Kernel SHAP because their explanations may vary across executions due to random sampling behaviour. To measure this variation explanations are generated five times using different random seeds and coefficient of variation is calculated afterward. While reviewing similar experiments in earlier studies this issue appeared repeatedly. Stability becomes a major concern whenever explanation methods involve probabilistic sampling.

Finally statistical significance testing is performed to determine whether observed differences between explainability methods are genuinely meaningful or simply caused by random variation. The Wilcoxon signed-rank test was selected using a 5% significance level mainly because it works well for paired comparisons where the data doesn't always follow a normal distribution. This becomes important in explainability research since many evaluation metrics behave irregularly across different models and datasets. While going through different papers this issue appeared repeatedly. Several studies preferred non-parametric statistical methods for the same reason.

At first the comparison results looked reliable from simple observation alone. But after deeper analysis the limitation became obvious. Visual comparisons and average scores by themselves can sometimes create misleading conclusions especially when metric distributions vary between experiments. This created another issue. The results needed stronger statistical support.

What makes this more difficult is that explainability evaluation often involves paired experimental outputs rather than completely independent samples. Because of that the Wilcoxon signed-rank test becomes more suitable than many traditional parametric tests. It compares paired observations directly and doesn't rely heavily on assumptions of normality. The method also helps reduce the possibility that final conclusions are based only on isolated observations or random variation inside the datasets. Researchers noticed this repeatedly in experimental AI studies.

So the statistical analysis strengthens the reliability of the overall findings and provides quantitative evidence for the comparison results instead of relying only on surface-level interpretation. The conclusions become more defensible this way. Even small performance differences can then be evaluated more carefully rather than being accepted immediately without statistical validation.

I. Integrated Gradients

Integrated Gradients is a gradient-based attribution technique developed to improve traditional gradient explanations used in neural networks. Instead of analyzing gradients only at the final input point the method calculates feature importance by integrating gradients along a path connecting a baseline input and the actual input sample. This produces explanations that are usually smoother and more stable compared to ordinary gradient methods.

One important advantage of Integrated Gradients is that the method satisfies theoretical properties such as sensitivity and implementation invariance. These properties help ensure that features contributing strongly toward predictions receive meaningful attribution scores. The method is particularly useful for image and text classification tasks because it reduces the noise often present in standard gradient explanations. Researchers noticed this repeatedly while evaluating medical image models.

Integrated Gradients also requires lower computational cost compared to sampling-based methods like SHAP and LIME. Still explanation quality depends heavily on baseline selection. For image datasets a completely black image is commonly used as the baseline while tabular datasets may use zero vectors instead. Different baseline choices sometimes generate very different attribution results. That's where the real concern starts. Baseline selection itself becomes an important methodological decision affecting final explanation quality.

J. Counterfactual Explanations

Counterfactual explanations work differently from ordinary feature attribution techniques. Instead of explaining why a prediction occurred they explain

what minimal changes would be needed to achieve a different prediction outcome. These explanations answer practical "what-if" questions. For example if a loan application gets rejected the explanation may indicate that approval would have been possible if the applicant's income were slightly higher or debt ratio slightly lower.

In this study counterfactual explanations were generated using the DICE framework which stands for Diverse Counterfactual Explanations. DICE extends standard counterfactual generation by producing multiple alternative explanations instead of only one modified instance. This helps users understand several possible pathways for changing predictions rather than relying on a single suggestion. The framework also tries to maintain realistic constraints so generated counterfactuals remain practically achievable.

Counterfactual explanations are highly valuable for transparency fairness and regulatory compliance because affected users can directly understand what changes influenced decisions. These explanations become especially useful in financial healthcare legal and recruitment systems. But after deeper analysis another limitation became obvious. Generating realistic counterfactuals can become computationally expensive especially in large feature spaces and highly complex black-box models. Some generated counterfactuals may also suggest unrealistic combinations if real-world constraints are not handled properly.

H. Anchor Explanations

Anchor explanations generate rule-based explanations using simple "if-then" conditions. Instead of assigning feature importance values the method identifies conditions under which model predictions remain stable with high probability. These rules resemble ordinary human decision logic which makes them easier for non-technical users to understand.

For example an Anchor explanation may state that if a customer's credit score falls below a specific threshold and debt ratio exceeds another threshold the loan application is likely to be rejected. Such explanations are especially useful in legal compliance-oriented and business environments where decision reasoning needs to remain understandable. While reviewing explainability methods this simplicity became one reason Anchor explanations remained attractive for policymakers and regulators.

However Anchor explanations also contain limitations. To achieve very high precision generated rules may become extremely specific and apply only to a small number of instances. This reduces explanation coverage and generalizability. Searching for optimal rule combinations also becomes computationally difficult in datasets containing many features because possible condition combinations increase rapidly. Even with these challenges Anchor explanations remain useful because they produce highly interpretable rule-based outputs suitable for human-centered decision support systems

I. Attention-Based Explanations

Attention-based explanations are mostly connected with transformer architectures such as BERT GPT and Vision Transformers. These systems use self-attention mechanisms to identify relationships between tokens image regions or input sections during prediction generation. Attention weights are commonly visualized through heatmaps showing which words or image regions received stronger focus during prediction.

Attention explanations became popular partly because they appear visually intuitive. In natural language processing tasks attention maps highlight words influencing sentiment classification translation or question-answering systems. In computer vision applications attention mechanisms identify image regions receiving stronger focus during recognition tasks. This becomes easier to communicate to non-technical audiences because heatmaps appear visually understandable even without deep machine learning expertise.

But while reviewing different research papers another issue appeared repeatedly. Attention scores do not always represent true feature importance. Several studies demonstrated that attention weights can sometimes be modified without changing final predictions. This suggests that attention alone may not provide fully faithful explanations of model behaviour. To improve reliability researchers often combine attention mechanisms with gradients or perturbation-based techniques. Most researchers now treat attention explanations as supportive visualization tools rather than complete explanation systems.

Overall each explainability method provides different strengths and weaknesses depending on the dataset application domain model architecture and

deployment environment involved. SHAP and LIME provide flexible model-agnostic explanations. Grad-CAM and Integrated Gradients improve deep learning interpretability. Counterfactual explanations focus more on actionable guidance while Anchor explanations improve human readability. Attention mechanisms provide additional insight into transformer-based models. No single explanation method works perfectly everywhere though. That limitation remained visible throughout the study.

The findings also suggest that combining multiple explanation methods may improve overall reliability and interpretability. Hybrid explainability approaches integrating feature attribution visual explanations rule-based reasoning and counterfactual analysis may provide more trustworthy understanding of AI behaviour compared to relying on only one technique. Combined approaches can also help reduce weaknesses present in individual methods and support stronger explainability in complex AI systems.

As AI technologies continue becoming integrated into critical real-world applications explainability will remain an important requirement for trustworthy and responsible artificial intelligence. Future research in explainable AI still has several unresolved challenges. While going through different papers one thing became very clear. Current explanation methods still struggle with consistency stability and reliability in many practical situations. Some explanations may appear useful initially but later fail when the model is tested on different datasets or slightly modified inputs. At first this looked reliable. But after deeper analysis the limitation became obvious. Researchers noticed this repeatedly especially in healthcare and financial applications where prediction errors can directly affect people.

What makes this more difficult is that explanation quality itself is hard to measure properly. Some methods generate explanations that seem understandable for researchers but ordinary users still can't interpret them easily. In certain cases explanations may also become unstable when small input changes are introduced. This created another issue. Adversarial manipulation remains a growing concern because some AI systems can still be tricked into producing misleading explanations or unreliable predictions. The model still wasn't transparent in many real-world conditions.

Future studies will probably need to focus more on improving explanation faithfulness scalability and human understanding across different application domains. This issue becomes more visible in healthcare datasets where trust and clarity are extremely important. Building AI systems that are not only accurate but also transparent understandable and fair will continue to be important for long-term public trust. And this can't be ignored. Safe adoption of AI technologies depends not only on prediction performance but also on whether people can realistically trust and understand how these systems make decisions.

VI. DISCUSSION

A. Toolkit Comparison

The experimental findings showed pretty clearly that SHAP, LIME, and TCAV cannot really replace one another. All three methods explain models differently and they seem to serve different purposes depending on who is using the explanation and why. At first this looked straightforward. But after deeper analysis the limitation became obvious. Even when the attribution directions looked similar across methods the explanation quality itself was not the same. Stability faithfulness completeness and even the style of explanation changed noticeably from one method to another.

For regulatory documentation especially under the EU AI Act SHAP appeared to be the strongest option overall. Its explanations were more stable across repeated runs and the attribution process remained reproducible because of the TreeSHAP implementation. That matters in compliance settings where organizations may need to justify exactly how explanations were generated. SHAP also satisfies the efficiency property which means prediction variance is fully distributed across features instead of leaving unexplained portions behind. This was discussed in several papers I reviewed. LIME on the other hand produced explanations that were easier for non-technical users to understand but its instability became difficult to ignore. The average cross-run Spearman correlation for Adult XGBoost remained only around 0.671. So relying on LIME alone for formal compliance documentation probably isn't reliable enough where reproducibility is expected.

TCAV contributed something different entirely. Instead of focusing mainly on feature attribution it allowed direct testing of concepts inside the model. While SHAP and LIME can reveal that marital status behaves differently across demographic groups TCAV can examine whether the broader concept of "being married" systematically influences predictions. That distinction matters more than it first appears. Regulatory investigations often ask concept-level questions rather than feature-level ones. Researchers noticed this repeatedly.

So the three methods seem most effective when used together rather than separately. SHAP works better for large-scale audit summaries and technical documentation. LIME becomes more useful when organizations need instance-level explanations for affected individuals. TCAV fits best during investigative analysis when auditors are checking whether sensitive concepts are influencing predictions in hidden ways. No single toolkit solved everything.

B. Interpretation of RQ2: Explanation Quality and Complexity

One of the more surprising findings involved how explanation quality changed as model complexity increased. SHAP explanations actually became more faithful for complex models while LIME explanations became weaker. At first this sounded counterintuitive. But after spending more time reviewing the mechanisms behind both approaches the pattern started making more sense.

SHAP computes Shapley values directly and TreeSHAP produces exact results for tree-based models regardless of complexity. LIME works differently. It creates local linear approximations around predictions and tries to simplify model behavior within a small neighborhood. What makes this more difficult is that highly complex models like XGBoost or random forests often have nonlinear decision boundaries that don't behave well under local linear approximation. So as complexity increases LIME struggles to represent the real decision process accurately.

This created another issue.

Organizations using LIME for production-level explainability may unknowingly receive weaker explanations than expected especially when working with tree-based ensemble models. The problem isn't necessarily caused by poor data quality or bad implementation practices. The limitation comes from the mismatch between LIME's approximation strategy and the structure of nonlinear models themselves. While going through different studies this issue appeared repeatedly in discussions around tabular explainability systems.

The practical implication is important. Companies already using LIME as their primary explanation tool for compliance workflows should compare its faithfulness against SHAP directly on their own production datasets. That comparison shouldn't be treated as optional research validation anymore. It probably needs to be part of compliance due diligence.

C. Interpretation of RQ3: Complementarity of SHAP and LIME

The combined SHAP-LIME analysis revealed something more interesting than simple performance comparison. The two methods actually expose different layers of model behavior. SHAP identified broad population-level patterns that were impossible to observe from a single explanation. LIME meanwhile helped verify how those broader patterns appeared at the level of individual predictions.

For example SHAP analysis showed that marital status behaved differently depending on sex and demographic group. Prior conviction history also produced uneven attribution strengths across racial categories. These trends only became visible after analyzing thousands of predictions together. A single local explanation wouldn't reveal them. LIME however showed how those larger statistical patterns translated into decisions affecting specific individuals. That's where the real concern starts. Aggregate fairness doesn't automatically guarantee fairness for individuals receiving predictions.

One of the strongest findings involved explanation inconsistency across demographic groups. Female and African-American instances consistently showed weaker agreement between SHAP and LIME explanations compared to privileged groups. The explanations themselves became less reliable for the populations most likely to challenge algorithmic decisions. And this can't be ignored.

The situation creates a difficult circular problem. Groups most vulnerable to proxy discrimination are also the groups receiving weaker explanation consistency from explanation systems themselves. So the people who need explanations the most may

actually receive explanations that are less dependable. During literature review this appeared repeatedly across discussions on fairness and accountability but seeing it emerge experimentally here made the issue feel much more concrete. Regulatory frameworks probably need demographic-specific explanation reporting rather than relying only on aggregate explanation quality metrics.

D. The Five-Stage Explainability Compliance Pipeline

Based on the experimental findings this study proposes a five-stage Explainability Compliance Pipeline for AI systems deployed in regulated domains. The idea is to transform abstract explainability requirements from GDPR the EU AI Act and NIST frameworks into a workflow organizations can actually implement practically.

The first stage is Assess. Before training begins organizations should audit datasets carefully. Representation imbalance proxy correlations and sensitive feature relationships need to be identified early because later explainability analysis depends heavily on dataset quality. The findings should be documented in a formal Data Characteristics Record. Relevant concepts for TCAV analysis should also be selected during this phase.

Stage two is Generate. After training the organization generates SHAP explanations for all instances and LIME explanations for a stratified sample covering different demographic and prediction combinations. TCAV analysis is then performed on selected concepts. SHAP interaction values should also be calculated for important feature pairs because interaction-based discrimination may remain hidden otherwise.

The third stage is Evaluate. Explanation quality metrics such as faithfulness stability comprehensibility and consistency are measured separately for each demographic group. This becomes more visible in healthcare datasets where explanation quality sometimes varies strongly across groups. Organizations should verify LIME stability using repeated explanation generation and cross-run consistency testing. Aggregate reporting alone isn't enough anymore.

Stage four is Validate. Subgroup-level SHAP analysis compares attribution values across demographic groups. Features showing direction reversal or large attribution differences may indicate proxy discrimination patterns. TCAV scores should also be compared between groups especially for concepts related to protected characteristics. All suspicious findings should be documented clearly.

Stage five is Document. Organizations compile all reports explanation summaries TCAV findings and evaluation metrics into a complete Explainability Documentation Package aligned with regulatory technical documentation requirements. Production monitoring should continue after deployment because explanation quality may drift as data distributions change over time. The model still wasn't transparent in many cases even after deployment monitoring.

E. Implications for Practitioners

The findings of this study carry several practical implications for organizations deploying AI systems in high-risk environments. One immediate implication involves organizations alr eady relying entirely on LIME for explainability compliance. Based on the experimental findings these organizations may be receiving substantially weaker explanations than expected especially for XGBoost and random forest models. The difference in faithfulness sometimes reached 15–20 percentage points compared with SHAP. That gap became concentrated in demographic groups already vulnerable to unfair treatment.

Another important issue involves explanation quality thresholds. Many organizations currently depend on toolkit defaults without defining explicit explanation quality standards beforehand. But after reviewing different explainability workflows it became obvious that thresholds themselves reflect policy and ethical decisions not just technical ones. Faithfulness stability and comprehensibility requirements should ideally be defined collaboratively by technical experts legal teams and domain specialists instead of inherited automatically from explanation software defaults.

Production monitoring also matters much more than many organizations assume initially. Models may generate acceptable explanations during deployment but explanation quality can decline later because of shifting data distributions or demographic changes in served populations. This issue appeared repeatedly while examining model monitoring literature. Automated explanation monitoring probably needs to become standard practice for high-risk systems.

F. Implications for Regulatory Policy

The findings also suggest several implications for future AI regulatory frameworks. First regulatory systems should probably define minimum explanation quality thresholds similar to fairness standards already used in discrimination law. Without measurable requirements organizations can claim explainability compliance even when explanations remain unstable or unfaithful.

Second explanation quality reporting should be demographic-specific instead of aggregated globally. Aggregate metrics often hide explanation failures affecting minority groups. That became very visible during subgroup analysis.

Third LIME's instability means it shouldn't be accepted as a sole compliance explanation method unless stability reporting is also provided. Regulators may eventually need to specify minimum reproducibility requirements for explanation systems.

Finally TCAV deserves more recognition within regulatory auditing workflows because concept-level analysis captures forms of discrimination feature attribution methods sometimes fail to express clearly. Saying a model learned that "being married affects income predictions differently across sexes" communicates regulatory risk much more directly than abstract feature importance values alone.

G. Limitations and Threats to Validity

Several limitations affect the generalizability of this study. The datasets themselves represent specific historical and cultural contexts including the 1994 US census Adult dataset and older criminal justice records from Florida. Proxy discrimination patterns identified here may not generalize perfectly to newer datasets or different countries. Practitioners outside US contexts should probably apply these findings carefully rather than assuming direct transferability.

The study also focused mainly on binary classification problems using tabular data. Explainability methods for text image and multimodal systems may behave differently from the tabular implementations evaluated here. TCAV itself originally targeted neural activation analysis and required adaptation for tree-based classifiers which introduces approximation limitations.

Another limitation involves explanation quality metrics themselves. Measures such as deletion-AUC Spearman correlation and feature-count comprehensibility are still indirect approximations of actual human understanding. Ground-truth faithfulness is difficult to measure because the true internal reasoning of complex models isn't fully observable. Future work should probably involve human participant studies instead of relying entirely on automated metrics.

And finally there are conclusion validity concerns connected to multiple comparisons. The study involved several datasets models explanation metrics and demographic subgroup analyses simultaneously. Some findings may occur by chance statistically. Still the major conclusions appeared consistently across independent experiments and model types which increases confidence in the broader patterns observed during the study.

VII. FUTURE WORK

Future research in Explainable Artificial Intelligence still has many unresolved areas. Current XAI methods can explain predictions to some extent, but there are clear limitations when these techniques are applied in real-world systems. Stability problems computational overhead fairness concerns and weak human interpretability continue to appear across different studies. This was discussed in several papers I reviewed. The issue becomes even more noticeable in large deep learning systems and modern generative AI models where explanations are often difficult to trust completely.

One major direction for future work involves explainability in generative AI systems. Existing explanation techniques were mostly designed for prediction-based machine learning models where outputs are fixed and easier to analyze. Generative systems behave differently. Large language models text-to-image generators and AI code generation tools produce open-ended outputs that may change depending on prompts context or retrieved information. Because of this many traditional explainability methods no longer work properly. Researchers now face the challenge of understanding how prompts training data and hidden reasoning patterns influence generated responses. The model still wasn't transparent.

Unlike classification systems these outputs are dynamic and sometimes unpredictable. Small prompt changes may completely alter the generated response. While going through different papers this issue

appeared repeatedly. Future explainability systems will likely need to measure prompt sensitivity more carefully and identify which parts of a prompt influence outputs most strongly. Such analysis could help reduce unreliable behavior and improve safer interactions with generative AI systems.

Another important challenge involves source attribution in retrieval-augmented systems. Modern AI systems often combine internal model knowledge with external documents during response generation. Users increasingly want to know where generated information actually came from and whether those sources are reliable biased or potentially harmful. Future explainability frameworks may need mechanisms capable of tracing generated content back to influential documents or training examples. At first this looked manageable. But after deeper analysis the limitation became obvious. Current feature attribution methods cannot fully capture these complex relationships inside large generative systems.

Fairness and demographic representation will also remain important future research areas. Generative AI systems are trained on massive internet-scale datasets which may contain hidden stereotypes misinformation or social bias. Because of this generated outputs can sometimes amplify harmful patterns unintentionally. Researchers noticed this repeatedly. Future explainability methods must therefore analyze how different demographic groups are represented within generated content and whether certain concepts become amplified suppressed or distorted during generation. And this can't be ignored.

Another issue involves evaluation benchmarks. Current explainability evaluation techniques were mainly created for classification problems where outputs remain fixed and measurable. Generative AI systems require completely different evaluation frameworks. Future benchmarks may need to measure factual grounding reasoning transparency hallucination detection prompt sensitivity and source attribution reliability together. What makes this more difficult is that explanation quality itself still doesn't have one universal definition. Different researchers evaluate it differently depending on the application and domain involved.

As generative AI systems continue expanding into healthcare education journalism cybersecurity and software development explainability will become much more important for public trust and regulatory compliance. Governments and organizations are already raising concerns regarding AI-generated misinformation biased outputs and lack of transparency. Because of this future explainability systems will probably need to evolve beyond traditional feature attribution methods. Building AI systems that remain understandable accountable and trustworthy throughout deployment may become one of the most important future directions in explainable AI research.

7.1 Explainability for Generative AI Systems

One of the biggest challenges in modern explainable AI research is developing effective explanation methods for generative AI systems. Traditional explainability techniques were mainly designed for prediction-based machine learning models that work with fixed datasets and produce clear outputs such as classifications or numerical predictions However modern generative AI systems work very differently when compared with traditional machine learning models. Large language models text-to-image generators retrieval-augmented systems and AI coding tools don't simply produce fixed predictions from structured input data. Instead they generate open-ended outputs and those outputs can change a lot depending on prompts context or even slight differences in user interaction. Because of this the behavior of generative AI systems becomes much harder to interpret clearly. Many existing explainability methods that worked reasonably well for earlier machine learning models don't fully explain how these newer systems actually generate responses or content. This created another issue. Researchers noticed this repeatedly while studying generative AI systems.

At first these models looked easier to trust because the generated outputs often appeared natural and convincing. But after deeper analysis the limitation became obvious. Large language models such as GPT and other transformer-based systems generate responses by processing enormous amounts of training data and predicting token sequences based on surrounding context. The internal reasoning process still remains difficult to observe directly. Users only see the generated answer. The actual path taken by the model while producing that response usually stays hidden somewhere inside extremely large neural

network architectures. The model still wasn't transparent.

What makes this more difficult is that generative AI systems don't always produce the same output for the same style of task. Small prompt changes may produce very different responses. In some situations the generated content may appear highly accurate while in other cases the system may generate incorrect or misleading information confidently. And this can't be ignored. While going through different papers it became clear that explainability techniques designed for traditional prediction systems are often insufficient for understanding these generative models completely. That's where the real concern starts because these systems are increasingly used in healthcare education software development and decision-support environments where reliability and transparency matter a lot.

Unlike traditional classification systems, the outputs are dynamic, creative, and often unpredictable. This makes explainability much more difficult because there is no single fixed prediction to analyze. Researchers now face the challenge of understanding which training data examples, prompt tokens, or internal reasoning patterns influenced a generated response. Existing feature attribution methods designed for structured datasets cannot fully capture these complex relationships within generative systems.

One important future direction is developing explanation methods that can trace generated outputs back to influential training data sources or retrieved documents. In retrieval-augmented generation systems,

7.2 Standardized Explainability Benchmarks

One clear limitation in current XAI research is the absence of standardized explainability benchmarks with known-answer ground truth. Most evaluation approaches still rely on indirect metrics such as deletion-AUC or behavioral performance measures instead of directly measuring whether explanations are actually correct. This created another issue. Researchers can compare explanation methods technically but verifying true explanation faithfulness remains difficult.

Future work should focus on developing synthetic benchmark datasets where the true feature importance values are already known through explicitly designed data generation processes. Such datasets would allow researchers to evaluate whether explanation methods correctly identify the factors genuinely influencing model predictions. This could improve objectivity in explainability evaluation and help create more reliable standards across different machine learning applications.

7.3 Continuous Explanation Quality Monitoring

AI systems deployed in real-world environments rarely operate under stable conditions forever. Data distributions change over time. User behavior changes too. Economic conditions medical trends and even data collection methods can evolve gradually causing incoming data to become different from original training datasets. As a result model performance may decrease after deployment even if the system initially performed well during testing stages.

When these distributional shifts occur explanation quality may also degrade alongside model accuracy. Features that were previously important may become less relevant while newer patterns remain uncaptured by the model. This becomes more visible in healthcare datasets where patient characteristics and disease patterns may evolve over time. Users however may continue trusting explanations even when those explanations no longer represent actual model behavior accurately. That's where the real concern starts.

Most existing MLOps platforms mainly focus on tracking prediction accuracy infrastructure performance and latency metrics. Very few systems continuously monitor explanation quality after deployment. Faithfulness fairness stability and robustness of explanations are rarely evaluated continuously in production systems. While reviewing different monitoring frameworks this gap appeared repeatedly.

Future research should therefore focus on continuous explanation monitoring systems integrated directly with MLOps pipelines. Such systems could automatically track explanation consistency feature importance drift fairness metrics and demographic explanation equity over time. Automated alerts may also become necessary. If explanation quality drops below acceptable thresholds developers auditors or compliance teams should receive notifications immediately. This would help organizations maintain

transparency reliability and regulatory compliance even after deployment conditions change.

As AI systems become more deeply integrated into critical real-world applications continuous explanation monitoring may become just as important as monitoring prediction accuracy itself. Future XAI systems cannot simply generate explanations once and assume they remain valid forever. The explanations must remain reliable throughout the entire lifecycle of deployed AI systems.

7.4 Human-Centered Explanation Evaluation

Most explainability evaluation metrics still focus heavily on technical properties such as faithfulness stability sparsity and computational efficiency. These metrics are useful from an engineering perspective but they do not fully measure whether people actually understand the explanations produced by AI systems. The technical explanation may appear correct mathematically while ordinary users still remain confused about what the system actually meant.

Methods such as SHAP and LIME generate detailed feature attribution explanations. Researchers and engineers may understand these visualizations relatively easily. Patients loan applicants or ordinary users often cannot. This was discussed in several studies I reviewed. The gap between technical explainability and actual human understanding still remains large.

Future research should therefore move more strongly toward human-centered evaluation approaches. User studies involving real affected individuals could help researchers understand whether explanations genuinely improve trust fairness perception and decision understanding. For example patients could review medical risk explanations while loan applicants evaluate explanations connected to rejected credit applications. Such studies may provide insights that automated evaluation metrics cannot capture properly.

Another challenge involves explanation personalization. Different users require different explanation formats depending on education domain expertise emotional state and technical knowledge. A machine learning engineer may understand SHAP values without difficulty while a patient may require much simpler explanations. Future explainability systems will probably need adaptive explanation interfaces instead of using identical explanations for every user group.

The relationship between explainability and psychology also deserves more attention. Users do not always interpret explanations objectively. Some people may trust explanations simply because they appear visually convincing while others may reject explanations emotionally even when the system behaves correctly. Understanding these human factors will likely become very important for future explainable AI research.

7.5 Intersectional Explanation Analysis

The current study mainly evaluates explanation quality using single protected attributes individually. Real-world discrimination however often occurs at intersections between multiple demographic characteristics. Black women for example may receive very different explanations compared to Black men white women or white men. Single-attribute analysis may fail to capture these differences completely.

Future work should focus on computationally efficient intersectional explanation analysis methods capable of identifying groups receiving the weakest explanation quality without requiring exhaustive evaluation of every demographic combination. This area still remains underexplored in explainable AI research.

7.6 Adversarial Robustness of Explainability Compliance

Another important future direction involves adversarial robustness in explainability systems. Previous studies showed that methods such as LIME and SHAP can sometimes be manipulated by specially designed classifiers. Models may appear explainable during audits while still behaving unfairly in practice. Researchers noticed this repeatedly in adversarial explainability studies.

Future explainability compliance systems may therefore require dedicated robustness testing procedures capable of detecting whether AI models are intentionally gaming explanation-based auditing frameworks. These robustness evaluations could become part of future Explainability Compliance Pipelines especially for high-risk applications where adversarial manipulation creates serious ethical or regulatory concerns.

VIII. CONCLUSION

This study started with a fairly simple question. Can explainable AI actually work in real regulatory and

compliance environments or does it remain mostly theoretical? After carrying out experiments across different datasets models and explainability techniques the results suggest that practical implementation is possible. But it isn't as straightforward as many frameworks initially make it appear. One explanation method alone usually doesn't provide enough reliability fairness or transparency for complex AI systems. Different explainability approaches solve different problems and each comes with its own limitations.

While reviewing different papers this issue appeared repeatedly. Explainability cannot really be treated like a one-step checklist item where an organization simply generates an explanation and claims compliance. Real-world deployment is messier than that. The quality of explanations changes depending on the dataset model architecture user group and application domain. Because of this organizations need structured auditing processes capable of evaluating explanation stability fairness faithfulness and reliability across multiple deployment conditions. This became much more obvious during comparative analysis across benchmark datasets and explainability toolkits used throughout the study.

The comparative evaluation of SHAP LIME and TCAV showed that each method contributes differently toward AI transparency and auditing. SHAP consistently produced stable explanations with strong mathematical grounding and high faithfulness to actual model behaviour. At first this looked reliable. And in many situations it actually was. SHAP performed especially well for global feature importance analysis formal auditing tasks and regulatory reporting environments where explanation consistency matters heavily. LIME behaved differently. It generated explanations that were easier for ordinary users to understand because the outputs were simpler and more localised around individual predictions. This makes LIME useful in consumer-facing applications where users mainly want understandable reasoning for a single decision.

But after deeper analysis the limitation became obvious. LIME explanations sometimes changed noticeably across repeated executions because of random sampling effects. Researchers noticed this repeatedly. So organizations relying heavily on LIME may need additional validation methods averaging strategies or repeated explanation checks to improve reliability. The model still wasn't completely stable in several situations. $f(x) \approx g(x)$

TCAV contributed in a very different way compared to SHAP and LIME. Instead of focusing mainly on feature attribution TCAV evaluated concept-level influence within model predictions. This became particularly useful during fairness analysis because it allowed direct testing of whether concepts such as race gender or age were influencing AI behaviour indirectly. The explanations generated through TCAV also felt easier for regulators and non-technical stakeholders to interpret because they relied more on conceptual understanding rather than numerical feature contributions. While going through multiple fairness-related studies this advantage appeared consistently. SHAP LIME and TCAV therefore shouldn't really be treated as competing techniques. They work better as complementary methods inside larger explainability audit frameworks.

One important observation from this research was how differently explanation quality behaved when model complexity increased. SHAP explanations generally remained stable even for highly non-linear models like XGBoost because TreeSHAP could still calculate feature contributions accurately in many cases. LIME struggled more here. Its local linear approximation often failed to represent complicated decision boundaries properly for highly non-linear systems. This created another issue. Organizations depending only on LIME for complex machine learning systems may receive explanations that appear understandable but are not fully faithful to actual model behaviour. That difference becomes important during regulatory auditing where explanation correctness matters as much as readability.

Another major finding involved demographic fairness in explainability itself. Initially the focus was mostly on whether predictions were biased. But explanation quality also varied across demographic groups. Underrepresented populations often received explanations with lower consistency and reliability compared to privileged groups. This issue becomes more visible in healthcare datasets and criminal justice applications where proxy discrimination patterns remain active inside training data. Reporting only average explanation quality metrics can therefore hide serious fairness problems affecting vulnerable communities.

What makes this more difficult is that some explanation techniques may appear fair overall while still producing weaker explanations for minority groups located in underrepresented regions of the feature space. Several studies reviewed during this research discussed similar concerns. So fairness in explainable AI shouldn't only mean fair predictions. Explanations themselves must also remain understandable reliable and equally accessible across demographic groups. And this can't be ignored.

The research also identified multiple examples of proxy discrimination through SHAP LIME and TCAV analyses. In the Adult Income dataset marital status indirectly behaved like a gender-related proxy influencing income predictions differently for male and female instances. This reflects how historical labour and social patterns slowly become embedded within machine learning systems over time. Similar findings appeared in the COMPAS dataset where prior conviction counts affected racial groups differently because of structural inequalities already present in historical criminal justice records. The analysis also showed that younger African-American defendants sometimes received weaker protection from high-risk classifications compared to Caucasian defendants in related scenarios.

These patterns weren't identified by only one explainability method either. Multiple techniques repeatedly pointed toward similar problematic behaviours. That's where the real concern starts. When independent explanation methods identify the same fairness issue the evidence becomes much harder to dismiss. Multi-method explainability auditing therefore provides stronger validation compared to relying on a single explanation framework alone.

Another thing that became clear during this study is that explainability isn't only a technical research challenge anymore. It has become a regulatory ethical and social requirement as well. AI systems now influence important decisions in healthcare finance hiring education public services and law enforcement. Because of this organizations are increasingly expected to maintain transparency accountability and fairness throughout the deployment lifecycle. High model accuracy by itself doesn't seem sufficient anymore. Users regulators and affected communities also want understandable reasoning behind automated decisions.

To support practical implementation this research proposed a five-stage Explainability Compliance Pipeline consisting of Assess Generate Evaluate Validate and Document stages. The idea behind this framework was to move explainability away from vague theoretical discussion and toward something more systematic and auditable. The pipeline aligns reasonably well with frameworks such as the EU AI Act GDPR Article 22 and the NIST AI Risk Management Framework. By following structured explainability workflows organizations can better evaluate transparency monitor explanation quality identify fairness issues and maintain compliance documentation over time.

The study also highlighted that different industries require different kinds of explanations. Healthcare systems usually need clinically meaningful visual explanations. Financial systems often require transparent feature-level reasoning for loan or risk decisions. Legal and criminal justice systems demand explanations that are understandable auditable and suitable for accountability processes. So using the same explainability method everywhere doesn't really work effectively. Different deployment environments need different explanation strategies depending on stakeholder requirements regulatory obligations and risk levels involved.

One thing that became increasingly noticeable while analysing these systems is that explainability compliance cannot be achieved simply by generating explanations once during development. Explanations need continuous evaluation verification and monitoring after deployment as well. AI systems operate in dynamic environments where user populations training distributions and social conditions continue changing over time. As these changes happen explanation quality may also degrade gradually. Bias patterns may emerge later even if the original system appeared reliable initially. Future explainability systems therefore need automated monitoring mechanisms capable of tracking explanation stability fairness and reliability continuously during production use.

Overall this research suggests that the technical foundations required for explainable AI compliance already exist to a large extent. Many advanced explanation methods fairness evaluation frameworks auditing tools and monitoring approaches are already available. Regulatory structures supporting AI

transparency are also developing rapidly across different regions. But successful implementation still depends heavily on organizational commitment. Companies governments and institutions must actively invest in transparent AI practices continuous auditing fairness evaluation and human-centered system design rather than treating explainability as only a regulatory checkbox.

In conclusion this study shows that explainable AI is gradually becoming a fundamental requirement for responsible AI deployment rather than an optional technical feature Future AI systems will probably need to do much more than just generate accurate predictions. That part alone doesn't seem enough anymore. The systems also need to provide explanations that people can actually understand and question when required. While going through different papers this issue appeared repeatedly especially in discussions related to fairness and accountability. An explanation may look technically correct for developers but still remain confusing for ordinary users affected by the decision. This becomes more visible in healthcare datasets and legal applications where AI predictions can directly influence treatment access financial opportunities or even judicial outcomes.

What makes this more difficult is that explanations also need to remain reliable across different demographic groups. One explanation method may work reasonably well for one population while becoming weak or misleading for another group because of imbalances in training data. Researchers noticed this repeatedly. At first some models looked trustworthy because their prediction accuracy was high. But after deeper analysis the limitation became obvious. The explanation quality itself was inconsistent. The model still wasn't transparent.

Building more responsible AI systems will require cooperation between multiple groups instead of only machine learning researchers working independently. Engineers policymakers regulators healthcare professionals legal experts and affected communities will all need to contribute in some way. Different perspectives matter here because technical performance alone can't fully address ethical or social concerns connected with AI deployment. This was discussed in several papers I reviewed during the literature survey stage.

Another thing that became clear during this research is that long-term trust in AI systems depends on more than prediction accuracy. People generally expect systems to operate transparently and responsibly especially when decisions directly affect their lives. If users cannot understand why an automated system produced a certain decision trust eventually decreases even when the prediction itself is technically correct. And this can't be ignored. Explainability is becoming closely connected with accountability fairness and public acceptance of AI technologies.

REFERENCES

[1] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in NeurIPS, vol. 30, 2017.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD, 2016, pp. 1135–1144.

[3] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viegas, "Interpretability Beyond Classification Output: Quantitative Testing with Concept Activation Vectors (TCAV)," in Proc. 35th ICML, 2018.

[4] European Commission, "Proposal for a Regulation on Artificial Intelligence (AI Act)," COM/2021/206, 2021.

[5] National Institute of Standards and Technology, "AI Risk Management Framework (AI RMF 1.0)," U.S. Department of Commerce, 2023.

[6] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017.

[7] W. Samek, T. Wiegand, and K.-R. Muller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Neural Network Predictions," arXiv:1708.08296, 2017.

[8] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods," in Proc. AAAI/ACM AIES, 2020, pp. 180–186.

[9] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'," AI Magazine, vol. 38, no. 3, pp. 50–57, 2017.

[10] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022.
[11] S. M. Lundberg et al., "From Local Explanations to Global Understanding with Explainable AI for Trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.
[12] D. Alvarez-Melis and T. S. Jaakkola, "On the Robustness of Interpretability Methods," arXiv:1806.08049, 2018.
[13] C. Yeh et al., "On the (In)fidelity and Sensitivity of Explanations," Advances in NeurIPS, vol. 32, 2019.
[14] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in Proc. 1st ACM FAccT, 2018, pp. 77–91.
[15] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," Advances in NeurIPS, vol. 29, 2016.
[16] A. Jobin, M. Ienca, and E. Vayena, "The Global Landscape of AI Ethics Guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389–399, 2019.
[17] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," ProPublica, May 2016.
[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1–35, 2021.
[19] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," Nature Machine Intelligence, vol. 1, pp. 206–215, 2019.
[20] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence," IEEE Access, vol. 6, pp. 52138–52160, 2018.
[21] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, 2018.
[22] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," AI Magazine, vol. 40, no. 2, pp. 44–58, 2019.
[23] Z. C. Lipton, "The Mythos of Model Interpretability," Queue, vol. 16, no. 3, pp. 31–57, 2018.
[24] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.
[25] H. Suresh and J. V. Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning," arXiv:1901.10002, 2019.
[26] E. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," Science, vol. 366, no. 6464, pp. 447–453, 2019.
[27] I. Chen, F. D. Johansson, and D. Sontag, "Why Is My Classifier Discriminatory?" Advances in NeurIPS, vol. 31, 2018.
[28] A. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box," Harvard Journal of Law and Technology, vol. 31, no. 2, 2018.
[29] P. Voigt and A. von dem Bussche, The EU General Data Protection Regulation: A Practical Guide, Springer, 2017.
[30] L. D. Sjoding et al., "Racial Bias in Pulse Oximetry Measurement," New England Journal of Medicine, vol. 383, no. 25, pp. 2477–2478, 2020.
[31] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning," in Proc. ACM FAccT, 2019.
[32] J. Wexler et al., "The What-If Tool: Interactive Probing of Machine Learning Models," IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 1, pp. 56–65, 2020.
[33] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Muller, Eds., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019.
[34] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual Fairness," Advances in NeurIPS, vol. 30, 2017.
[35] A. Datta, S. Sen, and Y. Zick, "Algorithmic Transparency via Quantitative Input Influence," in Proc. IEEE S&P, 2016.
[36] K. W. Crenshaw, "Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color," Stanford Law Review, vol. 43, no. 6, pp. 1241–1299, 1991.

[37] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing Fairness Gerrymandering," in Proc. 35th ICML, pp. 2564–2572, 2018.