10.48047/jocaaa.2026.35.01.59

Operationalizing Explainable AI in Insurance Fraud Detection: From Model Transparency to Decision Justification

Harender Bisht
Independent Researcher, USA

Abstract

Artificial intelligence and machine learning increasingly have their place in the insurance sector as a tool for detecting fraud, but there is a big disparity between what an algorithm spits out and what data combination fraud detectives, compliance officers, and regulators actually need. Model transparency is not sufficient in high-stakes situations of fraud detection, where this understanding requires action by multiple stakeholders instead of technical interpretability. The difference between model transparency and decision justification is an important conceptual step of understanding that explanations should go beyond the disclosure of algorithmic mechanics to include more general communicative and evidentiary needs. The conceptualization of decision justification as an operational artifact to address the needs of particular stakeholders throughout the lifecycle of fraud detection seems the right place where explainable AI can be implemented in insurance applications. In the generation of scalable explanations, patterns in architecture supporting it have to happen in a way that is seamlessly integrated with human-in-the-loop processes that are typical of claims operation. Ethical implications of fairness, detection of bias, and obligation to transparency in relation to policyholders require constant attention as the fraud detection systems are becoming more advanced and consequential. Guidelines on practical design considerations that focus on the association between the level of explanation and the level of decision severity can give concrete guidance to organizations interested in responsible AI implementation in fraud detection situations.

Keywords: Explainable Artificial Intelligence, Insurance Fraud Detection, Decision Justification, Human-AI Collaboration, Algorithmic Transparency

1. Introduction

The application of artificial intelligence and machine learning technology in the detection of fraud in the insurance industry has brought about a great change in the operations of this sector. Contemporary fraud detection systems make use of advanced algorithms that have the ability to process vast amounts of claims data to find suspicious patterns that are actually invisible to human operators operating at an operational level. Nonetheless, the application of sophisticated analysis has posed an underlying functioning difficulty relative to explainability differences among algorithmic results and the decisionmaking necessities of fraud investigators, compliance officers, and regulatory authorities.

The increasingly complex machine learning algorithms have resulted in what researchers refer to as the black-box problem, where the processes that take place in it to generate its decision-making remain obscure to human stakeholders who are obligated to act in line with its suggestions [1]. The study on explainable artificial intelligence has highlighted that the explainable versus performance trade-off poses a considerable challenge to high-stakes applications where accuracy and explainability are key factors. The survey done on XAI methodologies establishes that although many tools have been developed to deal with issues of

10.48047/jocaaa.2026.35.01.59

transparency, there are still enormous gaps between technical interpretability and practical expectations of domain practitioners who need actionable insights into algorithmic prescriptions.

The boundary between the model transparency and decision justification is a vital conceptual leap towards the concept of explainable artificial intelligence in insurance. A thorough discussion of techniques of explaining black-box models has created taxonomies between the various types of explanations, such as model explanations, which expose the inner logic, and outcome explanations, which explain particular predictions [2]. This study establishes that explainability best entails the consideration of the audience of the explanation, the nature of the problem under consideration, and the method by which the explanations should be conveyed to the different stakeholders whose level of technical literacy and information requirements are dissimilar.

2. Explainability in Insurance Fraud Detection: Conceptual Foundations

The theoretical underpinnings of explainability in insurance fraud detection must pay careful attention to the various concepts, taxonomies, and stakeholder needs that define the manner in which artificial intelligence systems should be built and implemented in this inductive field. The consequences of insurance fraud detection decisions have high stakes to a variety of stakeholder groups, which means that explainability requirements are far beyond those of lower-stakes analytical settings where the award of algorithmic errors has less severe consequences and can be more readily mitigated.

A large body of research analyzing explainable AI concepts and taxonomies has confirmed that responsible use of AI in the system with high stakes demands focusing on the numerous aspects of transparency beyond the technical interpretability level [3]. As highlighted in this analysis, explainability transcends the ability to comprehend the mechanics of models and extends to the wider scope of accountability, fairness, and trust that will regulate interactions between organizations using AI-based systems and the subjects of the algorithms. The study establishes the various stakeholder views which need to be dealt with using the explanation design, bearing in mind that developers, operators, and affected people have various informational requirements and technical abilities that define their ability to interact with different types of explanations.

Implementation of explainability is especially complicated by the insurance fraud detection context due to the adversarial nature of the area and its high stakes on policyholders whose claims are reported to investigation. Explanations that explain certain factors that lead to suspicion determination are necessary for fraud analysts to enable them to efficiently devote investigative resources and develop suitable interview strategies. Compliance teams require the presentation of documentation that shows that the recommendations of the algorithms are in conformity with the regulatory requirements and company policies that regulate how the policyholders can be treated fairly during the claims procedure. When policyholders are made susceptible to increased scrutiny on their claims, they have valid interests as to why they are being investigated, especially in light of the reputational and financial costs that might come with fraud claims.

Studies that consider the rigorous science of interpretable machine learning have explored the manner in which the explainability requirements ought to be measured and approved in a given setting of application [4]. This literature highlighted the fact that interpretability is not a concept that has a single face, but instead, it has many dimensions that have to be characterized in relation to specific tasks, stakeholders, and operational limitations. The analysis has identified the need to develop evaluation systems that determine whether explanations are being used for the purpose they are intended or are being used to meet shallow technical standards. In the case of insurance fraud detection, this would mean that the explanation systems

10.48047/jocaaa.2026.35.01.59

should be checked against their ability to assist effective and correct analyst decision-making as opposed to being rated on the basis of technical fidelity measures that are not necessarily related to operational utility.

Stakeholder	Primary Need	Explanation Format	Decision Support
Fraud Analyst	Evidentiary factors	Contextual narratives	Investigation prioritization
Compliance Team	Policy alignment	Audit documentation	Regulatory verification
Policyholder	Basis for scrutiny	Plain language	Appeal preparation
Regulator	Fairness assurance	Systematic reports	Oversight validation

Table 1: Stakeholder Explainability Requirements in Fraud Detection [3, 4]

The time aspect of fraud detection also makes the explainability requirements more difficult in a manner that cannot be dealt with by the usage of a static classification framework. Contrary to the applications where it is possible to make a decision once and consider it final, fraud determination often changes as an investigation unfolds, additional evidence is discovered, and conditions change during the claims lifecycle. Explanations need to have adaptability, not necessarily just starting with the initial assessment of the algorithm, but also the evolving evidentiary terrain that typifies complex fraud investigations where ground truth can be unknown over prolonged durations.

3. Limitations of Model-Centric Explainability

Modern explainable artificial intelligence studies have yielded a multiplicity of methods intended to shed light on the inner working of machine learning frameworks but these model-focused methods have proven to have tremendous weaknesses when used in realistic fraud detection settings where actionable decision support is needed as opposed to technical transparency. The awareness of these limitations is critical to the development of the explanation frameworks that would really fulfill the interests of fraud analysts and other stakeholders who need to interpret the algorithm outputs into the relevant investigative steps.

Studies that give a single method of explaining model forecasts by use of Shapley values have laid mathematical rigorous grounds on how to attribute feature contributions to model predictions [5]. This approach is based on cooperative game theory, offering stable and locally valid explanations, which are calculated by calculating the marginal contribution of each feature of all possible feature coalitions. The method shows how complicated models may be broken down into understandable parts that identify which input variables influenced specific predictions. Nonetheless, the conversion of Shapley value attributions into actionable advice to fraud analysts poses some basic problems that can not be addressed using purely technical methods. An analyst who has been given feature attribution values that show that some characteristics of claims had positive effects on the fraud scores is only loosely guided about the importance of the characteristics in the particular investigative situation, or how the observations should be applied in further evidence collection.

Another important addition to the model-centric explainability literature is the development of local interpretable model-agnostic explanations [6]. This study shows that complex black-box models can be locally approximated by interpretable surrogate representations that explain model behavior around local predictions. The approach also stresses being faithful to the underlying model but retaining interpretability to be reviewed by humans. Although these contributions exist, the inherent weakness of this kind of approach is that they tend to emphasize algorithmic mechanics as opposed to the justification of decisions

10.48047/jocaaa.2026.35.01.59

that are befitting high-stakes situations of fraud detection. A technically complete description of how a given model handled a claim with no insight of how necessary it was to flag a claim to be investigated offers no information as to whether flagging a claim to get investigated is a necessary, proportional, or fair response to the overall setting of organizational policies, regulatory standards, and ethical considerations that govern the operations of fraud detection.

The technical explanations provide a cognitive load that is hard to overcome in operational difficulties in the world of fraud detection, where the analysts are time-constrained and must identify a large number of claims in a minimum duration. The interpretation effort to explain a feature attribution is not always feasible in the case where the analyst is examining hundreds of claims per day and has to decide on the triage of investigation priorities. The structure of technical explanations, which is usually represented in numerical figures or visualizations with statistical literacy, might not correspond with the modalities of reasoning that fraud analysts use when implementing domain knowledge to make decisions about the investigation. There are further practical constraints of scalability since producing detailed model descriptions of all flagged claims might be computationally expensive beyond operational resource capacity, yet offer little marginal value to claims on which algorithmic evaluations are simple and unambiguous.

Technique	Core Approach	Strength	Operational Limitation
SHAP	Shapley value attribution	Mathematical rigor	Limited actionability
LIME	Local surrogate models	Model-agnostic	Fidelity variability
Feature Attribution	Variable importance ranking	Computational efficiency	Context absence
Attention Maps	Weight visualization	Neural network insight	Domain translation gap

Table 2: Model-Centric XAI Techniques and Operational Limitations [5, 6]

4. From Model Transparency to Decision Justification

The shift in model transparency into decision justification is a paradigm shift in the way explainable AI needs to be thought about and implemented into insurance fraud detection systems that work under regulated and high-stakes settings. Justification of decisions is not limited to exposing algorithmic mechanisms to include more overall requirements of communication and evidence of stakeholders who need to take, review, or challenge fraud determinations on the basis of contextual knowledge, as opposed to technical model knowledge.

Studies on explainable AI needs of medical decision support systems are instructive guides to insurance fraud detection since the stakes are similar and the regulatory conditions in both areas are similar [7]. This discussion highlights the fact that effective explanation systems of high-stakes applications should deal with the contextual needs of domain experts who hold much professional knowledge that should be combined with the output of algorithms to come up with appropriate conclusions. The study provides the main requirements, such as causability, which captures how well explanations can allow human beings to understand the causal relationships, and interactivity, which involves the ability of the user to interact dynamically with the explanations and not receive fixed outputs. In the context of insurance fraud detection, the above requirements mean that the explanations should relate the algorithmic observations with known

10.48047/jocaaa.2026.35.01.59

fraud typologies and investigation processes in a manner that makes an analyst's reasoning easier and not more cumbersome.

The systematic layers of translation needed to map model explanations to business-relevant reasoning contextualize the technical outputs in domain semantics relevant to the fraud investigators and compliance reviewers. When a model detects time-related irregularities in the patterns of claims submissions, the model's decision justification presupposes describing how the identified irregularities correspond to the identifiable fraud indicators, what investigation steps they imply, and what other possible explanations should be considered before an escalation. Such a translation process requires close interaction between technical teams that are aware of the model behavior and domain experts that are aware of the fraud investigation needs and constraints.

In-depth review of the concept of model interpretability has discussed the various concepts of transparency and interpretability, which should be differentiated when developing explanation systems intended to be used in practical scenarios [8]. This study shows that transparency can be interpreted on various levels, such as algorithmic transparency with respect to the mechanics of models, decomposability with respect to understanding the components, and simulatability with respect to the ability of humans to mentally trace through the reasoning of models. The analysis indicates that the various types of transparency are used to serve varying purposes and that a mix-up of these causes a creation of systems of explanation that are not able to satisfy the needs of the stakeholders in spite of the attainment of technical transparency goals. In the case of fraud detection, this means that explanation designs should be based on explicit knowledge of what transparency dimensions are in practice needed to support certain operational choices instead of aiming to achieve transparency as an abstract technical objective.

Dimension	Model Transparency	Decision Justification
Focus	Algorithm mechanics	Stakeholder understanding
Output	Feature weights	Contextual narratives
Audience	Technical users	Domain practitioners
Temporality	Static prediction	Dynamic investigation
Validation	Fidelity metrics	Operational utility
Goal	Model interpretability	Actionable reasoning

Table 3: Model Transparency versus Decision Justification [7, 8]

Temporal context is an important aspect of decision justification that is normally overlooked by modeldriven approaches, as they are based on a notion of single predictions but not on a process of investigation that evolves. The investigation of fraud cases has a long-term nature, with new evidence discovered, the original suspicion justified or dismissed, and the basis of the decision made by the investigators changing significantly as the team learns more. Reasoning behind decision-making should thus take into consideration time-based reasoning, in which one would be able not only to present the present-day assessment by the algorithms but also to clarify the relation between the present-day assessment and the preceding observation, and what needs to be included in deciding the way of future investigation as the cases evolve through review and escalation procedures.

10.48047/jocaaa.2026.35.01.59

5. Architectural Patterns and Human-in-the-Loop Workflows

The explainable AI operationalization in fraud detection systems has to be architecturally patterned to enable a scalable, high-performance explainable generation and a seamless fit with current analysis processes and human inspection of explanations that define the insurance claims business. Its efficient implementation would require the consideration of both the technical infrastructure and the human factors that will make the explanations in question actually enhance the quality of the decisions made in the operational situation.

The studies that analyze definitions, methods, and applications of interpretable machine learning offer guidelines on the way explanation systems are to be developed and implemented in real-life scenarios [9]. This discussion shows that interpretability should be interpreted in the context of any contextualised circumstances of generating and consuming explanations and not as an abstracted characteristic of different models or methods of explanation. The study finds different application situations, such as in scientific discovery, where explanations are used to generate hypotheses, and in high-stakes decisionmaking, where explanations are used to promote accountability and proper human control. The detection of insurance fraud lies predominantly in the high-stakes category, which means that explanation architecture should support the human judgment and responsibility higher than the other goals, which could be valuable in other application scenarios.

The architectures of fraud detection specify the integration points at which explanations are generated and delivered to human reviewers whodecide onf an investigation. To minimize latency in the investigative process, event-driven explanation generation with events such as claim scoring or flagging can be used to generate explanations in advance of analyst review, before investigation begins. Nonetheless, this method must also be mindful of the necessity of allocating the computational resources since it might be expensive in terms of infrastructure development to create detailed explanations of all the flagged claims. The organizations often apply the tiering of explanation with light explanations produced during the first screening and detailed explanations produced on-demand during claims that enter the intensive investigation, where the extra effort in the computation is justified by the importance of the decision.

Human-in-the-loop workflows characterize the operational integration of explanations in the process of reviewing and making decisions by analysts, which finally defines the results in fraud. Studies looking at explainable AI metrics have dealt with the issue of measuring whether or not explanations are effective in serving their intended goals in assisting human decision-making [10]. This discussion highlights that to measure the quality of the explanation, they can not be measured using technical fidelity measures but rather must be measured in relation to human factors such as comprehension, degree of trust, and quality of the decision made. The study suggests assessment constructs that deal with aspects such as good explanations when it comes to quality testing, user satisfaction in terms of subjective experience, and performance in terms of the improvement of the task outcome. In the case of fraud detection applications, these dimensions of evaluation would mean that the explanation systems are to be tested with operational testing that shows a practical increase in the accuracy and efficiency of the decisions made by the analyst, and not only based on technical standards.

The problem of explanation fatigue deserves some specific discussion when it comes to high-volume fraud detection settings where analysts have to go through large volumes of claims daily. When analysts get desensitized to explanations that seem repetitive or that do not contribute substantive content to fraud scores in their own right, this may take place. The logic of escalation should involve an explanation-based criterion that marks the claims in which the work of an algorithm may be of interest to senior consideration because of the failure to explain, or because of some contribution of an unusual feature, or some seeming conflict

10.48047/jocaaa.2026.35.01.59

with known patterns of fraud that may suggest that the model itself is limited in some way and needs to be examined by human hands.

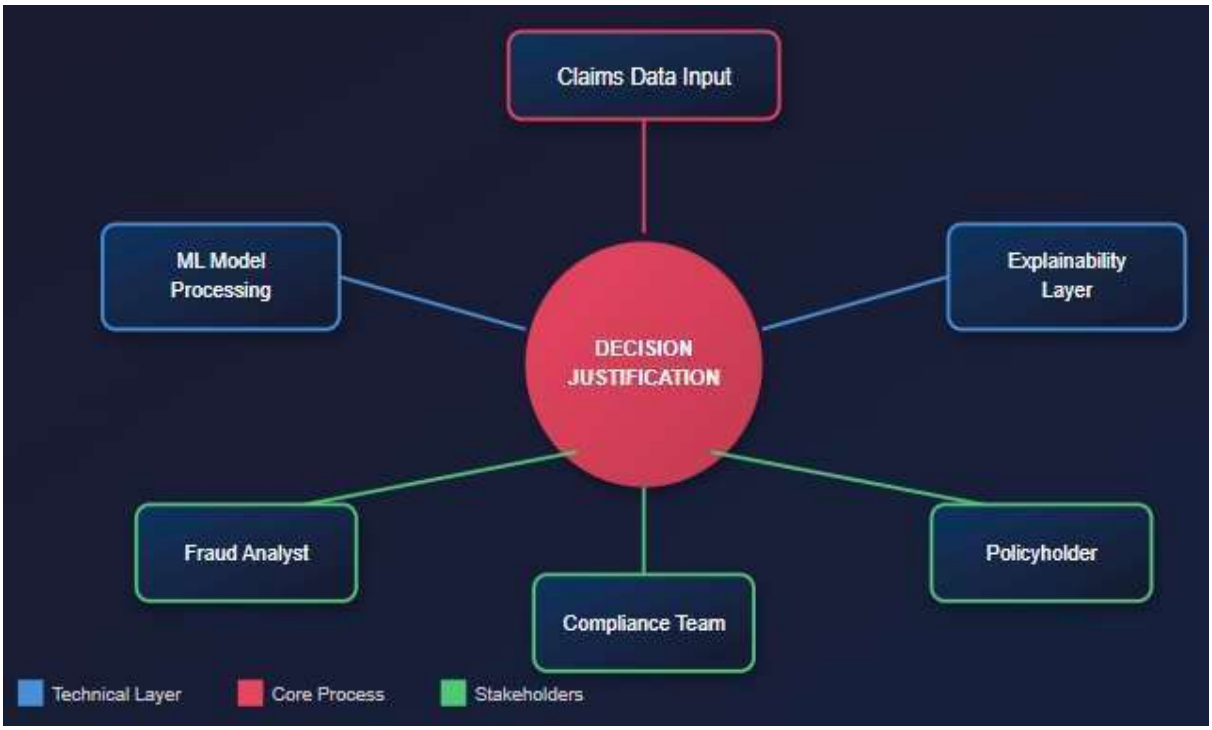

Fig 1: XAI Framework in Fraud Detection Lifecycle [7, 8, 9, 10]

6. Ethical Implications and Practical Design Guidelines

Operationalizing explainable AI in insurance fraud detection is associated with significant ethical consequences beyond technical performance aspects to include inherent inquiries of intercession, responsibility, and the just utilization of algorithmic authority over policyholders and claimants. These dimensions of ethics should be utilized to develop effective design principles that can guarantee the use of the systems of explanation to achieve the goals of accountability and assist in the functionality of the fraud detection activities.

Probably one of the most important elements of explainable fraud detection systems in terms of consequences is fairness and bias detection due to the possibility of algorithms trained on previous data continuing or increasing discriminatory patterns existing in previous claims handling and investigation decisions. Explainability is an important mechanism of bias detection and reduction, which allows auditors to look at whether a model explanation provides a disproportionate use of demographic proxies or features that are associated with a protected attribute and potentially breach anti-discrimination constraints. Explanation systems that are developed based on fairness goals should help to conduct a systematic analysis of model behavior according to demographic groups in order to detect possible differences that need to be corrected.

Transparency to policyholders whose claims are put under investigation is a field of growing regulatory and ethical interest as more of their business is automated in insurance activities. This principle that persons bearing the consequences of consequential automated decisions yield rights to an effective explanation and review has become very popular in regulatory systems worldwide. Applied use of policyholder-facing

10.48047/jocaaa.2026.35.01.59

explanations must take into account a strong emphasis on communication design to be able to provide justifications in a way that would be easily understood by laypeople, without disclosing information that could be used to commit fraud or otherwise undermine investigative procedures that rely on differences in information to perform effectively.

The possibility of misleading or simplistic explanations is highly dangerous, and they need to be mitigated by designing it carefully, validating it, and continuously monitoring it. When there are confident explanations of the explanations, which provide reasons rooted in weak or unsure evidence, the development of inappropriate trust can take place, and the results can be negative, such as wrongful accusations of fraud, not detecting the actual fraudulent claims. Likewise, those explanations that reduce complex algorithmic reasoning to oversimplistic forms might meet the requirements of transparency at a superficial level without actually offering a meaningful understanding of the basis of the decisions that would support both functions of effective human review and proper accountability.

Design principles that are applicable in practice to the development of usable explanations in the case of insurance focus on the correspondence between the granularity of the explanation and the severity of the decision, as more detailed justifications are needed to the decisions with more severe consequences in the case of policyholders. Accountability and compliance: The explanations should be documented and auditready, which implies that they necessitate unremitting storage and retrieval systems that would allow retrospective access to justifications that were made to certain decisions during their life cycle in operation.

Conclusion

The given framework introduces the essential difference between model transparency and justification of decisions as the basis of successful explainable AI usage in insurance fraud detection algorithms. Despite being useful in interpreting the mechanics of algorithms, technical interpretability methods do not provide enough to satisfy the needs of operational, regulatory, and ethical considerations of high-stakes fraud detection scenarios in which various stakeholders need explanations with different purposes. Actionable understanding that help to justify decisions should be supported in order to allow analysts to make effective decisions, compliance teams to confirm compliance with regulations, and policyholders to know decisions about individual interests. Patterns of architectural designs and strategies of workflow integration offer practical advice on how to deliver explainable capabilities of fraud detection at scale without compromising on performance and reliability constraints that are vital in making it operationally viable. Human-in-the-loop features, such as trust calibration, escalation reasoning, and explanation fatigue reduction, focus on successful human-AI cooperation in the investigative processes when the entirely automated and the entirely manual modes of investigation do not deliver the best results. The ethical consequences related to fairness and the transparency requirements and risks of misleading explanations are sustained as the fraud detection systems are becoming more advanced. The future directions will entail adaptive explanations and multimodal explainability and robust validation metrics so that insurance organizations can balance the goal of fraud prevention with the obligation of transparency and fairness.

References

[1] Amina Adadi and Mohammed Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, 2018. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8466590

[2] Riccardo Guidotti et al., "A Survey Of Methods For Explaining Black Box Models," arXiv:1802.01933, 2018. [Online]. Available: https://arxiv.org/abs/1802.01933

10.48047/jocaaa.2026.35.01.59

[3] Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," arXiv:1811.10154, 2019. [Online]. Available: https://arxiv.org/abs/1811.10154

[4] Finale Doshi-Velez and Been Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017. [Online]. Available: https://arxiv.org/abs/1702.08608

[5] Scott Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," arXiv:1705.07874, 2017. [Online]. Available: https://arxiv.org/abs/1705.07874

[6] Marco Tulio Ribeiro et al., "Why Should I Trust You?" Explaining the Predictions of Any Classifier," arXiv:1602.04938, 2016. [Online]. Available: https://arxiv.org/abs/1602.04938

[7] Andreas Holzinger et al., "What do we need to build explainable AI systems for the medical domain?" arXiv:1712.09923, 2017. [Online]. Available: https://arxiv.org/abs/1712.09923

[8] Zachary C. Lipton, "The Mythos of Model Interpretability," arXiv:1606.03490, 2017. [Online]. Available: https://arxiv.org/abs/1606.03490

[9] Alejandro Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," arXiv:1910.10045, 2019. [Online]. Available: https://arxiv.org/abs/1910.10045

[10] Robert R. Hoffman et al., "Metrics for Explainable AI: Challenges and Prospects," arXiv:1812.04608, 2019. [Online]. Available: https://arxiv.org/abs/1812.04608