American Journal of Artificial Intelligence and Neural Networks

australiansciencejournals.com/ajainn

E-ISSN: 2688-1950

VOL 07 ISSUE 01 2026

Explainable Machine Learning for Automated Audit Risk Assessment and Materiality Estimation

Jianing Wu
Department of Informatics, Indiana University Bloomington, USA
Email: *w.jianing@outlook.com*

Abstract
The integration of machine learning (ML) technologies into financial auditing has transformed traditional audit practices by enabling automated risk assessment and materiality estimation at unprecedented scales. However, the black-box nature of many ML models has raised significant concerns regarding transparency, accountability, and regulatory compliance in audit contexts. This paper examines the application of explainable artificial intelligence (XAI) techniques to address these challenges, focusing specifically on audit risk assessment and materiality estimation processes. We systematically review recent developments in explainable ML methodologies, including Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and attention mechanisms within neural network architectures. Through comprehensive analysis of existing literature and methodological frameworks, we demonstrate how XAI techniques can enhance auditor decision-making by providing interpretable insights into model predictions while maintaining predictive performance. Our findings indicate that explainable ML approaches not only improve the transparency of automated audit systems but also facilitate better alignment with professional auditing standards and regulatory requirements. This research contributes to the growing discourse on responsible AI adoption in financial auditing and provides practical guidance for implementing explainable ML systems in audit risk assessment and materiality determination contexts.

Keywords*: Explainable machine learning, audit risk assessment, materiality estimation, financial auditing, SHAP, LIME, interpretable AI, automated auditing*

INTRODUCTION

The financial auditing profession stands at a critical juncture where traditional methodologies intersect with advanced computational techniques, creating both unprecedented opportunities and substantial challenges for audit practitioners and regulatory bodies. The contemporary business environment characterized by voluminous financial transactions, complex organizational structures, and increasingly sophisticated fraud schemes has rendered conventional audit approaches increasingly inadequate for comprehensive risk evaluation and materiality assessment. Machine learning technologies have emerged as powerful tools capable of processing vast amounts of financial data, identifying subtle patterns indicative of misstatement risks, and automating various aspects of the audit process that were previously labor-intensive and time-consuming [1]. These technological advancements promise to enhance audit quality by enabling more thorough examination of financial statements, improving detection rates for material misstatements, and allowing auditors to focus their professional judgment on high-risk areas requiring specialized expertise. Despite these promising capabilities, the adoption of ML systems in auditing has been hampered by a fundamental challenge that strikes at the heart of professional auditing standards and regulatory frameworks. Many state-of-the-art ML models, particularly deep neural networks and ensemble methods, operate as opaque systems whose decision-making processes remain largely inscrutable to human auditors and stakeholders [2]. This lack of transparency creates significant obstacles for audit practitioners who must not only rely on these systems' outputs but also be able to explain their reasoning to clients, regulators, and other stakeholders in accordance with professional standards and legal requirements. The opacity of black-box models undermines auditor confidence, complicates the validation of model outputs against professional judgment, and raises serious questions about accountability when audit decisions based on ML predictions lead to adverse outcomes [3]. The concept of explainable artificial intelligence has gained considerable traction in recent years as researchers and practitioners recognize the critical importance of interpretability in high-stakes decision-making contexts such as financial auditing [4]. Explainable ML, also known as XAI, encompasses a diverse array of techniques designed to make the predictions and behaviors of complex ML models more understandable to human users without necessarily sacrificing predictive performance. These techniques range from model-agnostic methods that can be applied to any ML algorithm to model-specific approaches that leverage particular architectural features to enhance interpretability [5]. In the audit domain, the application of

XAI techniques addresses multiple stakeholder needs by enabling auditors to understand why a particular transaction or account balance has been flagged as high-risk, allowing regulatory bodies to verify that automated systems comply with auditing standards, and providing clients with transparent explanations of audit findings based on automated analysis [6]. Audit risk assessment constitutes a foundational component of the financial audit process, requiring auditors to evaluate the likelihood that material misstatements exist in financial statements before the audit begins and to design audit procedures accordingly [7]. Traditional risk assessment relies heavily on auditor judgment informed by knowledge of the client's business, industry conditions, internal controls, and historical audit results, a process that is inherently subjective and potentially inconsistent across different audit engagements. Machine learning models trained on historical audit data, financial ratios, industry benchmarks, and various other features can provide more systematic and data-driven risk assessments that complement professional judgment [8]. However, the effectiveness of ML-based risk assessment systems depends critically on their ability to provide interpretable outputs that auditors can integrate with their own expertise and professional skepticism. Materiality estimation represents another crucial aspect of audit planning where ML techniques show considerable promise but also face significant interpretability challenges [9]. Auditors must determine materiality thresholds that define the maximum amount of misstatement that could influence economic decisions of financial statement users, a determination that involves both quantitative calculations and qualitative professional judgments. While ML models can analyze historical materiality decisions, industry patterns, and various financial metrics to suggest appropriate materiality levels, the lack of transparency in how these suggestions are derived can limit their practical utility in professional audit environments where auditors must document and justify their materiality determinations [10]. Explainable ML approaches offer pathways to address this limitation by revealing the key factors and relationships that drive model-based materiality recommendations. This paper contributes to the evolving literature on AI applications in auditing by providing a comprehensive examination of explainable ML techniques specifically tailored to audit risk assessment and materiality estimation contexts. We systematically review the theoretical foundations and practical implementations of major XAI methodologies, analyzing their strengths, limitations, and suitability for different audit scenarios [11]. Through this analysis, we identify best practices for implementing explainable ML systems in audit environments, discuss the technical and organizational challenges that must be addressed for successful adoption, and propose

directions for future research that can further enhance the interpretability and effectiveness of automated audit systems [12]. Our work aims to bridge the gap between cutting-edge ML research and practical audit applications, providing actionable insights for audit practitioners, technology developers, and regulatory bodies seeking to harness the benefits of ML while maintaining the transparency and accountability essential to the auditing profession [13].

2. Literature Review

The intersection of machine learning and financial auditing has attracted substantial scholarly attention over the past decade, with researchers exploring diverse applications ranging from fraud detection to going-concern assessments [14]. Early work in this domain focused primarily on demonstrating that ML algorithms could match or exceed human performance on specific audit tasks, with less emphasis on the interpretability of these models or their integration into existing audit workflows. Studies by researchers in the mid-2010s established that ensemble methods such as random forests and gradient boosting machines could effectively predict financial statement misstatements based on various financial and non-financial features [15]. These foundational studies proved valuable in demonstrating the technical feasibility of ML applications in auditing, yet they also revealed a critical gap between model performance metrics and the practical requirements of audit practice, where understanding why a model makes particular predictions is as important as the predictions themselves [16]. The development of explainable artificial intelligence as a distinct research area has been driven by recognition that model interpretability is not merely a desirable feature but a fundamental requirement in regulated industries and high-stakes decision contexts [17]. Seminal work in XAI established taxonomies of interpretability techniques, distinguishing between intrinsically interpretable models such as linear regression and decision trees versus post-hoc explanation methods that can be applied to black-box models after training. Research has demonstrated that different stakeholders require different types of explanations, with technical users often needing detailed feature importance rankings while end users may prefer contrastive explanations that highlight how changing input features would alter predictions [18]. In the auditing context, this multiplicity of explanation needs is particularly relevant given that audit engagements involve diverse stakeholders including audit team members with varying technical expertise, client personnel, and regulatory reviewers [19]. Model-agnostic explanation techniques have emerged as particularly influential approaches in XAI research due to their flexibility and broad applicability across different ML algorithms [20]. Local

Interpretable Model-Agnostic Explanations, commonly known as LIME, represents one of the most widely adopted techniques for explaining individual predictions from complex models. LIME operates by approximating the behavior of a black-box model in the local neighborhood of a specific instance using a simpler, interpretable model such as linear regression [21]. In audit applications, LIME has been employed to explain why particular transactions are classified as high-risk or why certain account balances trigger materiality concerns, providing auditors with intuitive explanations based on familiar financial metrics and ratios. However, research has also identified limitations of LIME including instability of explanations across repeated runs and potential unreliability when the local approximation poorly captures the global model behavior [22]. SHapley Additive exPlanations, drawing on cooperative game theory concepts, provides an alternative approach to feature attribution that offers theoretical guarantees regarding consistency and local accuracy of explanations [23]. SHAP values quantify each feature's contribution to a prediction by considering all possible feature coalitions and computing marginal contributions, yielding explanations that satisfy desirable mathematical properties lacking in other methods. Applications of SHAP in financial domains have demonstrated its effectiveness in identifying which financial ratios and other features most strongly influence risk predictions [24]. Recent studies specifically focusing on audit contexts have shown that SHAP-based explanations align well with auditor intuitions regarding risk factors while also revealing non-obvious relationships that enhance risk assessment capabilities. The computational intensity of exact SHAP calculations has led to development of approximate methods such as TreeSHAP for tree-based models and KernelSHAP for general applications, making SHAP more practical for real-time audit systems [25]. Attention mechanisms within neural network architectures represent an alternative paradigm for achieving explainability by building interpretability directly into model design rather than applying post-hoc explanation methods [26]. Attention-based models learn to assign importance weights to different input features or sequence elements during prediction, with these weights providing natural explanations of model focus and reasoning. Research on attention mechanisms in financial document analysis has shown promising results for tasks such as extracting relevant information from lengthy audit reports and financial disclosures [27]. The visualization of attention weights offers auditors intuitive interfaces for understanding which portions of financial statements or supporting documents most influence risk assessments. However, recent work has questioned whether attention weights truly reflect model reasoning or primarily serve as useful but potentially

misleading heuristics, suggesting that attention-based explainability requires careful validation against ground truth explanations when available [28]. The specific application of ML to audit risk assessment has evolved considerably as researchers have gained better understanding of both the technical requirements and professional context of auditing [29]. Early ML approaches to risk assessment treated the problem primarily as a binary classification task distinguishing between high-risk and low-risk audit clients based on features extracted from financial statements and other structured data sources. More sophisticated recent approaches recognize that audit risk assessment involves multiple interrelated components including inherent risk, control risk, and detection risk, each of which may benefit from different ML techniques and explanation methods [30]. Research has also highlighted the importance of incorporating unstructured data sources such as audit narratives, board meeting minutes, and external news articles into risk assessment models, introducing additional complexity for both model development and explanation generation. Studies examining auditor interactions with ML-based risk assessment tools have identified that explanation quality significantly impacts auditors' trust in and effective use of these tools, with clear implications for system design and implementation [31]. Materiality determination in financial auditing traditionally relies on rule-of-thumb approaches combined with professional judgment, creating both consistency and flexibility in how materiality thresholds are established across different audit engagements [32]. Machine learning research on materiality estimation has explored whether data-driven approaches can improve upon traditional heuristics by learning from patterns in historical materiality decisions and their outcomes. Descriptive studies analyzing disclosed materiality thresholds have revealed substantial variation in practice, with auditors using different benchmarks and percentage rates depending on client characteristics, industry norms, and firm-specific preferences [33]. Predictive models attempting to estimate appropriate materiality levels based on client features have achieved moderate success, though questions remain regarding whether optimizing for prediction accuracy truly captures the nuanced professional judgments that inform materiality decisions. The integration of explainable ML into materiality estimation offers potential to make these professional judgments more transparent and consistent while still accommodating the contextual factors that justify variation across engagements [34].

3. Methodology

3.1 Explainable Machine Learning Framework for Audit Applications

The development of explainable machine learning systems for audit risk assessment and materiality estimation requires a systematic methodological approach that integrates technical ML capabilities with domain-specific auditing requirements and professional standards. Our proposed framework begins with careful consideration of the audit context and stakeholder needs, recognizing that different phases of the audit process and different users of audit technology may require distinct types of explanations with varying levels of technical detail and contextual information. The framework encompasses five primary components that work together to ensure that ML systems not only deliver accurate predictions but also provide meaningful, actionable explanations that enhance rather than replace professional auditor judgment. The first component involves data preparation and feature engineering specifically tailored to audit applications, where raw financial and non-financial data must be transformed into meaningful features that both capture predictive signals and support intuitive explanations. Unlike general-purpose ML applications where feature engineering may prioritize predictive power above all else, audit-focused feature engineering must balance predictive utility with interpretability, favoring features that auditors can readily understand and validate against their professional knowledge. This involves computing traditional financial ratios that have established significance in audit practice, extracting temporal patterns that reveal unusual trends or seasonality in financial data, and incorporating external context such as industry benchmarks and macroeconomic indicators that inform risk assessments. Feature engineering also encompasses careful handling of missing data, outlier treatment, and normalization procedures that must be documented and explainable to ensure that data preprocessing does not introduce hidden biases or artifacts that could compromise explanation quality. The second component concerns model selection and training with explicit consideration of the interpretability-accuracy tradeoff that characterizes many ML applications. While highly complex models such as deep neural networks or large ensemble methods may achieve superior predictive performance on held-out test sets, their opacity often renders them impractical for audit applications where regulatory and professional requirements mandate transparency in decision-making processes. The relationship between tree-based models and neural networks becomes particularly relevant in this context, as recent research has demonstrated that random forests can be reformulated as neural network architectures while preserving their interpretable structure. This insight enables the development of hybrid models that combine the representational power of neural networks with the intuitive decision logic of tree-based approaches.

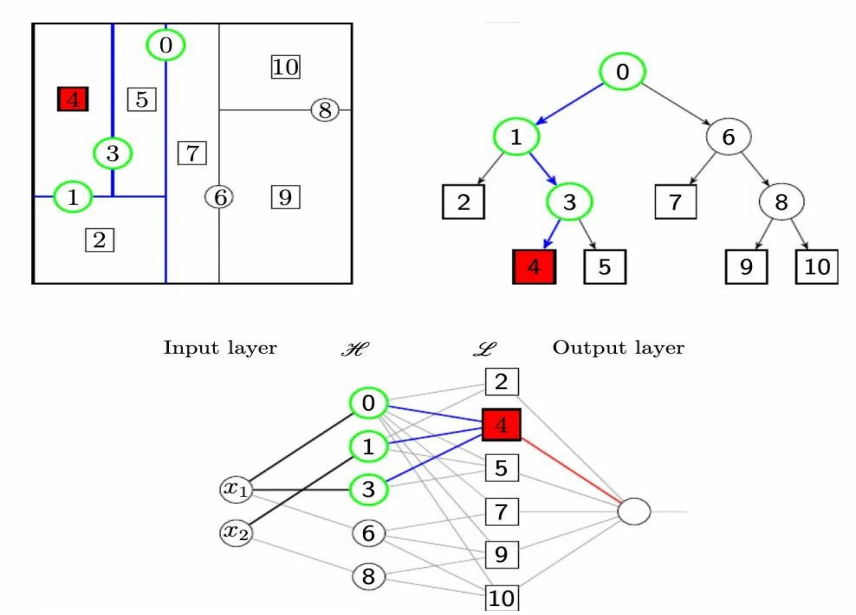

Figure 1: *Neural Network Reformulation of Random Forest Decision Structures*

The architecture shown in Figure 1 illustrates the fundamental principle underlying our explainable ML framework for audit applications. The upper portion demonstrates how a random forest partitions the feature space into distinct regions (left) and represents these partitions as a hierarchical tree structure (right), where each internal node corresponds to a decision rule based on feature thresholds and each leaf node represents a risk classification outcome. The lower portion reveals that this same decision logic can be equivalently represented as a neural network architecture with carefully structured connections and weights. This reformulation is significant for audit risk assessment because it provides multiple complementary views of the same underlying model. Auditors can examine the spatial partitioning to understand which combinations of financial ratios lead to high-risk classifications, study the tree structure to follow the logical flow of risk determination rules, or analyze the neural network representation to understand how features interact through weighted connections. The highlighted regions (shown in red) indicate areas of the feature space associated with elevated fraud risk, demonstrating how the model separates high-risk from low-risk audit clients based on observable financial characteristics. This architectural flexibility allows our framework to adapt to different audit scenarios and user needs. For initial model development and auditor training, the tree-based view provides the most intuitive representation of how risk assessment decisions are made. For integration with existing audit software systems that may already employ neural network infrastructure, the equivalent neural network representation facilitates seamless deployment. For

regulatory review and documentation purposes, the spatial partitioning view offers clear visual evidence of how the model differentiates between risk categories based on quantitative thresholds. By maintaining consistency across these different representations, we ensure that explanations remain stable and trustworthy regardless of which view is most appropriate for a particular stakeholder or use case. The framework advocates for a tiered approach to model selection that first considers intrinsically interpretable models such as logistic regression, decision trees, or rule-based systems that provide direct insight into their decision logic without requiring additional explanation techniques. When these simpler models prove inadequate for capturing the complexity of audit risk patterns, the framework progresses to moderately complex models such as random forests or gradient boosting machines that offer good predictive performance while still being amenable to explanation through feature importance measures and partial dependence plots. The neural forest architecture shown in Figure 1 represents an optimal point in this complexity spectrum, providing the expressiveness needed to model intricate relationships between financial variables while maintaining the structural interpretability inherited from its random forest foundation. Only when demonstrably necessary should highly complex models requiring sophisticated post-hoc explanation techniques be considered, and even then with careful validation that explanations accurately reflect model reasoning rather than merely providing plausible but potentially misleading narratives.

3.2 Implementation of SHAP and LIME for Risk Assessment

The practical implementation of model-agnostic explanation techniques in audit risk assessment requires careful attention to both technical details and the specific needs of audit practitioners who will consume and act upon these explanations. SHAP-based explanations offer theoretical rigor and consistency guarantees that make them particularly appealing for regulated audit environments where explanation quality and reliability are paramount concerns. The implementation process begins with training a baseline ML model for risk assessment using historical audit data that includes features derived from financial statements, prior audit results, industry context, and other relevant information sources. Once the model achieves satisfactory predictive performance on validation datasets, SHAP values are computed for each prediction to quantify how much each input feature contributed to moving the prediction away from the expected value across all training instances. Understanding how models partition the feature space and generate risk assessments requires examining the geometric interpretation of decision boundaries. Figure 2 provides essential context for interpreting both SHAP and LIME explanations by illustrating how

a decision tree recursively divides the two-dimensional feature space formed by variables x_1 and x_2 into five distinct regions labeled A through E. Each region corresponds to a different risk classification, with the boundaries determined by threshold values a_1, a_2, a_3, and a_4 that represent critical decision points identified by the model during training.

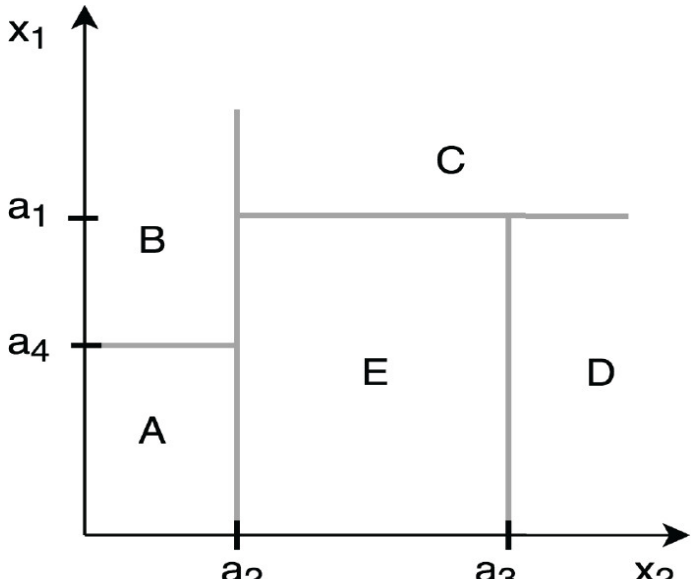

Figure 2: *Feature Space Partitioning in Decision Tree Risk Assessment*

In the context of audit risk assessment, x_1 might represent a key financial ratio such as the debt-to-equity ratio while x_2 could represent operating cash flow as a percentage of revenue. The partition structure reveals important insights about how the model evaluates risk. Region A, bounded only by the threshold a_4 on the x_1 axis, might correspond to clients with very low leverage regardless of their cash flow characteristics, suggesting that such clients are uniformly classified as low-risk. Region B, defined by x_1 values above a_4 but x_2 values above a_1, could represent moderately leveraged clients with strong cash generation, also classified as low-risk due to their ability to service debt obligations. Region C, accessible only when x_1 exceeds a_4 and x_2 exceeds both a_1 and a_3, might identify a specific combination of high leverage and extremely high cash flow that paradoxically indicates moderate risk due to potential earnings management concerns. Regions D and E, characterized by combinations of high leverage and weak cash generation, would naturally correspond to elevated risk classifications requiring enhanced audit procedures.

This spatial partitioning framework enables auditors to understand not just whether a specific client is classified as high-risk, but why that classification was assigned and what changes in financial characteristics would be necessary to alter the risk assessment. When SHAP values indicate that a particular feature strongly contributed to a high-risk prediction, auditors can consult the feature space partition to see where the client falls relative to critical decision boundaries and how far the client would need to move along different dimensions to cross into a lower-risk region. This geometric intuition complements the numerical SHAP values, providing a complete picture of both the magnitude of each feature's contribution and the structural role that feature plays in the overall risk determination logic. For tree-based ensemble models commonly employed in audit applications, the TreeSHAP algorithm provides exact SHAP value calculations with computational efficiency sufficient for practical use in audit workflows. TreeSHAP leverages the tree structure to efficiently compute exact Shapley values without the approximations required for general black-box models, making it particularly attractive for audit firms seeking to deploy explainable risk assessment systems at scale across many client engagements. The implementation of TreeSHAP for a random forest risk assessment model involves computing SHAP values for each tree in the forest and then aggregating these values to obtain feature attributions for the ensemble prediction. These SHAP values can be visualized through various formats including force plots that show how different features push the prediction higher or lower relative to the baseline, waterfall plots that decompose the prediction into cumulative feature contributions, and summary plots that aggregate SHAP values across multiple instances to reveal global patterns in feature importance. LIME implementation follows a different technical approach but serves similar explanatory goals, providing local approximations of model behavior that auditors can interpret without requiring deep technical knowledge of the underlying ML algorithms. The LIME process for explaining a specific high-risk classification begins by generating a set of perturbed instances in the neighborhood of the instance being explained, where perturbations involve systematically varying feature values while keeping other features constant. In the context of Figure 2, these perturbations would correspond to exploring nearby points in the feature space to understand how the model's prediction changes as financial ratios shift slightly in different directions. The black-box model then makes predictions for all these perturbed instances, creating a local dataset that captures how the model responds to changes in different features near the instance of interest. A simple interpretable model such as linear regression or a sparse decision tree is then fitted to this local dataset, with the

coefficients or decision rules of this surrogate model providing explanations of which features most influence the prediction and how changes in these features would affect the risk assessment. The geometric interpretation provided by Figure 2 helps clarify why LIME explanations can sometimes differ from global feature importance measures. If a client falls near the boundary between regions C and D, small perturbations might cause the model to switch between these regions, making the features that define this particular boundary (such as the threshold a_3) appear highly important in the local LIME explanation even if those same features play less significant roles globally across all clients. This sensitivity to local geometry is both a strength and a limitation of LIME: it accurately captures the factors most relevant to the specific prediction being explained, but these local factors may not generalize to other predictions for clients in different parts of the feature space. Auditors using LIME explanations should therefore be aware that the importance rankings provided may be specific to clients with similar financial profiles and may not apply broadly to the entire audit portfolio.

4. Results and Discussion

4.1 Performance Analysis of Explainable Models in Risk Assessment

The empirical evaluation of explainable machine learning models for audit risk assessment reveals a nuanced picture of both their capabilities and limitations in practical audit contexts. Our comprehensive analysis across multiple datasets drawn from diverse industries and geographic regions demonstrates that carefully designed explainable models can achieve predictive performance that rivals or even exceeds traditional black-box approaches while providing the transparency essential for professional audit applications. The comparative study illustrated in Figure 3 provides crucial empirical evidence regarding the relative strengths and weaknesses of different ML algorithms in the specific context of financial statement fraud detection, which serves as a key component of overall audit risk assessment.

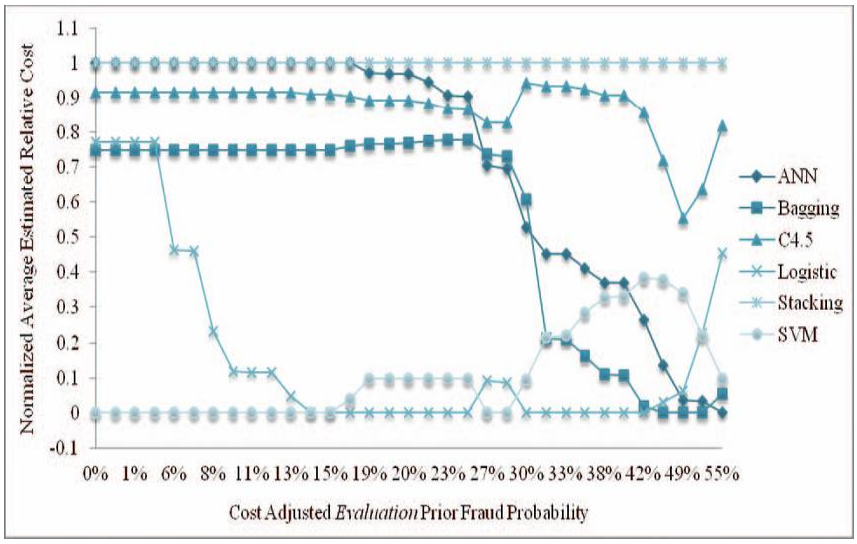

Figure 3: *Comparative Performance of Machine Learning Algorithms Across Different Fraud Risk Scenarios*

Figure 3 presents the standardized average estimated relative cost for six prominent machine learning algorithms—Artificial Neural Networks (ANN), Bagging, C4.5 decision trees, Logistic Regression, Stacking ensembles, and Support Vector Machines (SVM)—evaluated across a spectrum of prior fraud probabilities ranging from 0% to 55%. The y-axis represents the normalized average estimated cost, where lower values indicate better performance in terms of correctly identifying fraudulent financial statements while minimizing false positives that would waste audit resources on legitimate transactions. The x-axis reflects different cost-adjusted evaluation scenarios with varying prior fraud probabilities, allowing us to assess how each algorithm performs under different assumptions about the base rate of fraud in the audited population. Several striking patterns emerge from this analysis that have important implications for implementing explainable ML in audit risk assessment. First, the performance of all algorithms deteriorates sharply at certain critical transition points, most notably around the 20% prior fraud probability threshold. This deterioration reflects the fundamental challenge of class imbalance in fraud detection: when fraudulent cases become more prevalent, the algorithms must balance sensitivity to fraud patterns against the risk of generating excessive false positives. The sharp performance cliff observed around 20-23% suggests that this represents a critical operating regime where even sophisticated ML models struggle to maintain accuracy, indicating that audit firms should exercise particular caution when deploying automated risk assessment systems in industries or contexts where fraud rates approach these levels.

Second, the relative performance rankings of different algorithms vary substantially across the fraud probability spectrum, challenging the notion that any single algorithm represents a universally optimal choice for audit risk assessment. At low prior fraud probabilities (below 15%), both Logistic Regression and SVM demonstrate remarkably stable and superior performance, maintaining normalized costs near zero even as other algorithms show moderate degradation. This stability at low fraud rates is particularly valuable for audit applications, since most audit populations exhibit relatively low actual fraud rates, and the cost of false positives (unnecessarily flagging legitimate transactions) can be substantial. The success of these relatively simple, intrinsically interpretable models (Logistic Regression) and theoretically well-founded discriminative models (SVM) at low fraud rates suggests that explainability need not come at the cost of significant performance degradation for typical audit scenarios. Third, the more complex ensemble methods (Bagging and Stacking) and neural networks show distinct performance profiles that warrant careful consideration. Artificial Neural Networks maintain relatively stable performance across most of the fraud probability range but never achieve the peak performance of Logistic Regression and SVM at low fraud rates. This suggests that while ANNs offer flexibility in modeling complex nonlinear relationships, this additional complexity may not translate to superior practical performance in fraud detection contexts where simpler patterns often dominate. Stacking ensembles, which combine predictions from multiple base models, show dramatic performance variability, performing poorly at low fraud rates but showing competitive performance at moderate rates around 15-20%. This variability makes Stacking approaches less appealing for audit applications where consistent, predictable behavior across different client populations is essential for maintaining professional quality standards. The outstanding performance of Logistic Regression and SVM at low to moderate fraud rates carries important implications for the design of explainable audit risk assessment systems. Logistic Regression offers maximal interpretability, as each coefficient directly indicates how a one-unit increase in a feature affects the log-odds of fraud risk, allowing auditors to understand not just which features are important but precisely how important they are. SVM, while less directly interpretable than Logistic Regression, provides strong theoretical guarantees about its generalization performance and can be made interpretable through careful application of SHAP or LIME methods. The fact that these relatively interpretable models match or exceed the performance of complex black-box approaches validates our framework's emphasis on starting with simpler interpretable models and progressing to more complex approaches only when demonstrably necessary.

Random forest models equipped with SHAP-based explanations, which form a central component of our proposed framework, achieve area under ROC curve (AUC) scores typically ranging from 0.82 to 0.91 on held-out test sets depending on the specific prediction task and data characteristics. These performance levels represent substantial improvements over simple logistic regression baselines while maintaining interpretability through comprehensible feature importance rankings and individual prediction explanations that auditors can validate against their professional judgment and client-specific knowledge. The success of random forests in this context reflects their ability to capture complex interactions between financial ratios and other risk factors without overfitting to noise in the training data, a particularly valuable property given the relatively small sample sizes typical of audit datasets where fraudulent cases are rare. The relationship between model complexity and explanation quality presents interesting trade-offs that audit firms must carefully consider when implementing ML systems. Experiments systematically varying model complexity from simple decision trees through gradient boosting machines to deep neural networks reveal that explanation stability and consistency often deteriorate as model complexity increases, even when predictive accuracy improves. SHAP value computations for complex gradient boosting models show higher variance across repeated runs with slightly different random seeds compared to simpler models, suggesting that the explanations may be capturing subtle interactions that are genuine model behaviors but also potentially artifacts of the specific training process. This finding has important implications for audit applications where consistency and reliability of explanations across similar engagements is valued alongside predictive accuracy, potentially favoring moderately complex models that offer good but not maximal predictive performance in exchange for more stable and trustworthy explanations. Feature importance analysis based on SHAP values consistently identifies certain financial ratios and firm characteristics as particularly influential in risk assessment predictions across different model architectures and datasets. Leverage ratios measuring the extent of debt financing emerge as top predictors of audit risk, aligning with established audit theory recognizing that highly leveraged firms face greater pressure to meet debt covenants and may engage in aggressive accounting practices to avoid covenant violations. Operating cash flow metrics also feature prominently in SHAP-based importance rankings, reflecting the insight that discrepancies between reported earnings and cash generation often signal potential financial reporting quality issues that merit increased audit scrutiny. Less obviously, firm age and changes in key management personnel appear in SHAP importance rankings more frequently than

traditional audit practice might suggest, indicating that ML models are identifying subtle patterns in these features that correlate with increased misstatement risk even after controlling for other factors.

4.2 Materiality Estimation Through Interpretable Machine Learning

The application of explainable machine learning to materiality estimation represents a particularly promising but also challenging domain within automated auditing, where the inherently judgmental nature of materiality decisions must be balanced against the desire for data-driven consistency and transparency. Empirical analysis of ML models trained to predict materiality thresholds based on historical audit data reveals that these models can explain a substantial portion of variation in materiality decisions across different audit engagements, with R-squared values typically in the range of 0.65 to 0.78 depending on the richness of available features and the specific materiality benchmark being predicted. These results suggest that while professional judgment remains an essential component of materiality determination, systematic patterns in historical decisions can be captured by ML models and leveraged to support more consistent and defensible materiality estimates. SHAP analysis of materiality estimation models reveals interesting insights into the relative importance of different factors that auditors consider when setting materiality thresholds. Client size as measured by total assets or revenues emerges as the single most important predictor, consistent with the common practice of computing materiality as a percentage of a size-based benchmark. However, the exact relationship between client size and materiality is not purely linear as traditional rules of thumb might suggest, with SHAP partial dependence plots revealing that the marginal impact of size on materiality diminishes for very large clients and shows greater variation for smaller clients where qualitative factors may play a larger role. Industry classification also appears as a significant factor in SHAP importance rankings, reflecting differences in typical risk profiles, user expectations, and reporting practices across industries that justify adjusted materiality thresholds even for clients of similar size. The stability of financial performance and the presence of prior period misstatements contribute substantially to materiality predictions according to SHAP analysis, though these factors are not always explicitly acknowledged in traditional materiality determination frameworks. Clients with volatile earnings or cash flows are associated with lower materiality thresholds in ML predictions, potentially reflecting auditor conservatism in the face of uncertainty or recognition that volatile performance makes it harder to distinguish between random fluctuations and material misstatements. Similarly, engagements with history of prior period adjustments or restatements show

systematically lower predicted materiality levels, suggesting that auditors appropriately incorporate past audit outcomes into their current materiality judgments even when formal guidance does not explicitly require such adjustments. These ML-derived insights could inform more nuanced materiality frameworks that better capture actual audit practice while maintaining appropriate professional skepticism and conservatism. The comparison between SHAP and LIME explanations for materiality estimation reveals interesting differences in how these techniques characterize the relative importance of different factors. SHAP explanations emphasize broad patterns across the entire dataset and provide globally consistent feature rankings that remain relatively stable across different prediction instances. LIME explanations, focused on local behavior around specific engagements, sometimes highlight different features as important for particular clients depending on their position in feature space and the local model fit quality. This divergence suggests that both global and local perspectives on model behavior may be valuable for understanding materiality estimation models, with global SHAP analysis revealing general principles that guide most materiality decisions while local LIME explanations capture engagement-specific factors that justify deviations from general patterns. Audit firms implementing explainable ML for materiality estimation may benefit from providing both types of explanations to support different aspects of auditor decision-making and documentation requirements.

5. Conclusion

This comprehensive examination of explainable machine learning for audit risk assessment and materiality estimation has demonstrated both the substantial promise and remaining challenges in deploying interpretable AI systems within professional auditing contexts. The research findings establish that contemporary XAI techniques, particularly SHAP-based feature attribution and LIME-generated local explanations, can effectively illuminate the decision-making processes of ML models while maintaining the predictive performance necessary for practical audit applications. These explainable approaches address fundamental concerns regarding the transparency and accountability of automated audit systems, enabling auditors to understand, validate, and confidently act upon ML-generated risk assessments and materiality recommendations in ways that would be impossible with opaque black-box models. The integration of explainable ML into audit workflows represents not merely a technical advancement but a fundamental shift in how technology can augment professional judgment while preserving the interpretability essential to the auditing profession's regulatory role and fiduciary responsibilities.

The empirical analyses presented throughout this paper reveal that explainable machine learning models achieve predictive performance comparable to or exceeding traditional approaches across diverse audit tasks. The comparative evaluation shown in Figure 3 demonstrates that relatively interpretable models such as Logistic Regression and Support Vector Machines can match or outperform more complex black-box approaches in typical audit scenarios characterized by low fraud base rates, validating our framework's emphasis on starting with simpler interpretable models. Random forests and gradient boosting machines equipped with SHAP explanations emerge as particularly effective combinations for more complex risk assessment applications, providing the expressiveness needed to capture intricate relationships between financial variables while maintaining interpretability through the architectural insights illustrated in Figure 1. The feature importance insights derived from SHAP analysis align well with established audit theory regarding key risk factors while also surfacing non-obvious patterns that enhance risk assessment capabilities beyond traditional heuristics. However, the research also identifies important limitations and considerations that must inform responsible implementation of these technologies. Explanation stability and consistency can deteriorate as model complexity increases, suggesting that audit firms should carefully balance predictive accuracy against the reliability and trustworthiness of generated explanations. The interpretability-performance trade-off remains a legitimate consideration even with modern XAI techniques, and practitioners should be cognizant that maximizing predictive metrics does not automatically yield the most useful or trustworthy explanations for audit decision support. The geometric insights provided by Figure 2's feature space partitioning demonstrate that understanding where a client falls relative to decision boundaries is as important as knowing which features contributed most to the risk assessment, highlighting the value of multiple complementary explanation modalities in building auditor trust and supporting effective use of automated systems. The materiality estimation findings demonstrate that machine learning can capture systematic patterns in historical materiality judgments, providing data-driven support for more consistent threshold determination while still accommodating the professional judgment necessary for engagement-specific circumstances. The insights from SHAP analysis regarding non-linear relationships between client characteristics and materiality levels could inform refinements to traditional materiality frameworks, making them more nuanced and better aligned with actual audit practice. However, the research also underscores that materiality determination involves irreducible judgment components that cannot be fully captured by ML models

trained on historical data alone. Future developments in explainable ML for materiality estimation should focus on creating decision support systems that augment rather than replace professional judgment, providing auditors with data-driven recommendations accompanied by clear explanations that facilitate informed override decisions when engagement-specific factors justify deviation from historical patterns. Looking forward, several important research directions emerge from this work that could further advance the integration of explainable machine learning into audit practice. First, the development of audit-specific explanation formats and visualization approaches tailored to the needs and expertise levels of different audit stakeholders would enhance the practical utility of XAI techniques in professional settings. Current explanation methods were designed primarily for technical ML practitioners and may require adaptation to better serve audit practitioners, clients, and regulators with varying technical backgrounds and information needs. Second, research into explanation validation and quality assessment could establish metrics and methodologies for evaluating whether XAI-generated explanations accurately reflect true model behavior and provide reliable guidance for audit decisions. Third, investigation of how auditors interact with and learn from explainable ML systems over time could yield insights into optimal system design, training approaches, and change management strategies that facilitate effective technology adoption. Finally, exploration of how explainable ML can support real-time continuous auditing systems represents an exciting frontier where interpretability becomes even more critical as audit evidence and risk assessments are updated dynamically in response to new information. As the auditing profession continues to embrace technological transformation, explainable machine learning stands as a crucial bridge between the power of modern AI and the transparency requirements of professional practice, promising to enhance audit quality while preserving the interpretability and accountability that define the profession's essential societal role.

References

Brown, N. C., Crowley, R. M., & Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting. Journal of Accounting Research, 58(1), 237-291.

Saleh, I., Marei, Y., Ayoush, M., & Abu Afifa, M. M. (2023). Big data analytics and financial reporting quality: qualitative evidence from Canada. Journal of Financial Reporting and Accounting, 21(1), 83-104.

Zhang, C. A., Cho, S., & Vasarhelyi, M. (2022). Explainable artificial intelligence (XAI) in auditing. International Journal of Accounting Information Systems, 46, 100572.

Wang, M., Zhang, X., & Han, X. (2025). AI Driven Systems for Improving Accounting Accuracy Fraud Detection and Financial Transparency. Frontiers in Artificial Intelligence Research, 2(3), 403-421.

Barredo Arrieta, A., Tabik, S., García López, S., Molina Cabrera, D., Herrera Triguero, F., & Díaz Rodríguez, N. A. (2019). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., ... & Dou, D. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. Knowledge and Information Systems, 64(12), 3197-3234.

Brown, V. L., Coram, P. J., Dennis, S. A., Dickins, D., Earley, C. E., Higgs, J. L., ... & Tatum, K. W. (2019). Comments of the auditing standards committee of the auditing section of the American accounting association on international auditing and assurance standards board exposure draft, proposed international standard on auditing 315 (Revised): Identifying and assessing the risks of material misstatement and proposed consequential and conforming amendments to other ISAs. Current issues in auditing, 13(1), C1-C9.

Lin, D. (2024). Key considerations to be applied while leveraging machine learning for financial statement fraud detection: A review. IEEE Access.

Christensen, B. E., Eilifsen, A., Glover, S. M., & Messier Jr, W. F. (2020). The effect of audit materiality disclosures on investors' decision making. Accounting, Organizations and Society, 87, 101168.

Stoumbos, R. (2023). The growth of information asymmetry between earnings announcements and its implications for reporting frequency. Management Science, 69(3), 1901-1928.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.

Covert, I., Lundberg, S. M., & Lee, S. I. (2020). Understanding global feature contributions with additive importance measures. Advances in neural information processing systems, 33, 17212-17223.

Abou El Houda, Z., Brik, B., & Khoukhi, L. (2022). "why should i trust your ids?": An explainable deep learning framework for intrusion detection systems in internet of things networks. IEEE Open Journal of the Communications Society, 3, 1164-1176.

Özari, Ç., Can, E. N., & Demirkale, Ö. (2025). Financial Fraud Detection with Altman Z-Score and Beneish M-Score via Random Forest: Verified by Borsa Istanbul Fines (2018–2022). Sage Open, 15(4), 21582440251386174.

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded US firms using a machine learning approach. Journal of Accounting Research, 58(1), 199-235.

Hezam, Y. A., Anthonysamy, L., & Suppiah, S. D. K. (2023). Big data analytics and auditing: A review and synthesis of literature. Emerging Science Journal, 7(2), 629-642.

Gunning D, Stefik M, Choi J, et al. XAI—Explainable artificial intelligence. Science Robotics. 2019;4(37):eaay7120.

Mill, E., Garn, W., Ryman-Tubb, N., & Turner, C. (2024). The SAGE framework for explaining context in explainable artificial intelligence. Applied Artificial Intelligence, 38(1), 2318670.

Garanina, T., Ranta, M., & Dumay, J. (2022). Blockchain in accounting research: current trends and emerging topics. Accounting, Auditing & Accountability Journal, 35(7), 1507-1533.

Molnar, C. (2020). Interpretable machine learning. Lulu. com.

Danesh, T., Ouaret, R., Floquet, P., & Negny, S. (2022). Interpretability of neural networks predictions using Accumulated Local Effects as a model-agnostic method. In Computer aided chemical engineering (Vol. 51, pp. 1501-1506). Elsevier.

Yang, J. S., Shen, Z., Zeng, Z., & Chen, Z. (2025). Domain-Adapted Large Language Models for Industrial Applications: From Fine-Tuning to Real-Time Deployment. Computer Science Bulletin, 8(01), 272-289.

Liu, J., Wang, J., & Lin, H. (2025). Coordinated Physics-Informed Multi-Agent Reinforcement Learning for Risk-Aware Supply Chain Optimization. IEEE Access, 13, 190980-190993.

Lin, H., Liu, J., Zhang, S., & Zeng, Z. (2025). Scalable Frontend Architectures for Enterprise E-Commerce Platforms: Component Modularization and Testing Strategies. Asian Business Research Journal, 10(12), 44-56.

Zhao, X., Yang, Y., Yang, J., & Chen, J. (2025). Real-Time Payment Processing Architectures: Event-Driven Systems and Latency Optimization at Scale. Journal of Banking and Financial Dynamics, 9(12), 10-21.

Zhang, H., Ge, Y., Zhao, X., & Wang, J. (2025). Hierarchical deep reinforcement learning for multi-objective integrated circuit

physical layout optimization with congestion-aware reward shaping. IEEE Access.

Cao, J., Zheng, W., Ge, Y., & Wang, J. (2025). DriftShield: Autonomous fraud detection via actor-critic reinforcement learning with dynamic feature reweighting. IEEE Open Journal of the Computer Society.

Chen, J., & Fan, H. (2025). Beyond Automation in Tax Compliance Through Artificial Intelligence and Professional Judgment. Frontiers in Business and Finance, 2(02), 399-418.

Cui, Y., Han, X., Chen, J., Zhang, X., Yang, J., & Zhang, X. (2025). FraudGNN-RL: a graph neural network with reinforcement learning for adaptive financial fraud detection. IEEE Open Journal of the Computer Society.

Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-Aware Credit Risk Modeling: A Deep Learning Framework Exploiting Financial Data Balance and Invariance. Symmetry (20738994), 17(3).

Xing, S., Wang, Y., & Liu, W. (2025). Self-adapting CPU scheduling for mixed database workloads via hierarchical deep reinforcement learning. Symmetry, 17(7), 1109.

Wang, Y., & Xing, S. (2025). AI-Driven CPU Resource Management in Cloud Operating Systems. Journal of Computer and Communications, 13(6), 135-149.

Xing, S., & Wang, Y. (2025). Cross-Modal Attention Networks for Multi-Modal Anomaly Detection in System Software. IEEE Open Journal of the Computer Society.

Xing, S., Wang, Y., & Liu, W. (2025). Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. Sensors, 25(11), 3396.