Technema: Journal of Intelligent Engineering and Computing
Vol 1 No 2 June 2026, Hal 261-276
ISSN: 3163-6020 (Print) ISSN: 3163-5997 (Electronic)
Open Access: https://sovereignresearch.org/technema

Implementation of SHAP in Ensemble Learning for Interpreting Audit Risk Classification: A Preliminary Study Using a Public Audit Dataset

Yanuangga Galahartlambang[1*], Titik Khotiah[2], Ilham Basri K[3], Masrur Anwar[4], Mohamad Akhsan Rofiqi[5]
[1-5] Institut Teknologi dan Bisnis Ahmad Dahlan Lamongan, Indonesia
email: yanuangga@ahmaddahlan.ac.id[1]

Article Info :
Received:
12-06-2026
Revised:
25-06-2026
Accepted:
28-06-2026

Abstract

Ensemble learning models can achieve strong predictive performance on structured audit-risk data, yet their complex decision logic can limit transparency and technical accountability. This study develops a proof-of-concept Explainable Artificial Intelligence pipeline using SHapley Additive exPlanations (SHAP) to interpret Random Forest and XGBoost classifiers on the public UCI Audit Data. The experiment processed 776 observations and 27 columns, with one missing value in Money_Value imputed using the median of 0.09 and the non-numeric LOCATION_ID removed, resulting in 25 predictors and one binary target. A stratified 80:20 split with random_state = 42 produced 620 training observations and 156 test observations, with Random Forest achieving 1.0000 across accuracy, precision, recall, F1-score, and ROC-AUC, while XGBoost achieved 0.9936 accuracy, 1.0000 precision, 0.9836 recall, 0.9917 F1-score, and 1.0000 ROC-AUC. SHAP analysis identified Audit_Risk as the dominant global predictor, while its contribution to a representative at-risk prediction reached +5.82, shifting the raw model margin from −0.506 to 5.107 and yielding an estimated probability of approximately 0.994. However, because the dataset's target construction is closely associated with the Audit Risk Score, retaining Audit_Risk as a predictor introduces potential structural target leakage. The findings position the near-perfect performance as evidence of successful pipeline implementation and model interpretability rather than external generalization, supporting future leakage-controlled ablation, cross-validation, and external-data validation.

Keywords: *Audit Risk, Ensemble Learning, Explainable AI, SHAP, XAI.*

INTRODUCTION

The use of machine learning on structured and tabular data has expanded organizations' capacity to identify complex risk patterns that are difficult to capture through conventional statistical models and static rule-based procedures. In financial and auditing contexts, this development has created opportunities to process large volumes of organizational and transaction-level information while identifying nonlinear relationships, feature interactions, and heterogeneous risk profiles that may not be readily observable through conventional analytical procedures. Tree-based and ensemble learning methods have become particularly relevant because their flexible decision structures allow predictive systems to accommodate complex relationships within financial and audit-related datasets, while XGBoost has demonstrated strong scalability and predictive capability for structured data (Chen & Guestrin, 2016). Recent applications have extended machine learning toward audit quality assessment, financial risk prediction, fraud detection, and audit detection-risk reduction, indicating a broader movement from conventional rule-based assessment toward data-driven risk analytics (Resul & Ganji, 2025; Ulambayar et al., 2026). This development is accompanied by a growing demand for models whose outputs can be examined and communicated to auditors, financial analysts, regulators, and other stakeholders rather than being treated as opaque computational results. The central challenge is that increasing predictive capacity can simultaneously reduce the traceability of the model's reasoning, particularly when ensemble models combine numerous nonlinear decision structures that are difficult to inspect directly. This tension has positioned interpretability and explainability as increasingly important components of intelligent engineering and computing systems applied to high-impact decision environments.

The explainability problem becomes particularly important when machine learning is applied to audit risk classification because a prediction cannot be adequately represented by a class label alone. An auditor or decision maker may also require information about which variables contributed most strongly to a classification, whether their contributions increased or decreased the predicted risk, and why an individual observation received a particular risk designation. Research on supervised machine learning explainability shows that model interpretation encompasses different methodological perspectives, including intrinsic interpretability, post-hoc explanation, global understanding, and local prediction analysis (Burkart & Huber, 2021). XAI has consequently developed from a visualization-oriented concept into a broader framework concerned with transparency, accountability, human understanding, and responsible use of intelligent systems (Arrieta et al., 2020; Minh et al., 2022). In high-stakes applications, however, explaining a black-box model after prediction does not necessarily resolve the underlying problem of whether the model itself is appropriate for the decision context, particularly when the explanation merely describes an already opaque predictive mechanism (Rudin, 2019). Global explanations can reveal general patterns of feature contribution across observations, whereas local explanations can clarify the factors associated with an individual prediction, yet both forms remain dependent on the data, target definition, model architecture, and assumptions underlying the explanation method. The interpretability of an audit-risk classifier therefore needs to be considered as part of the analytical validity of the prediction rather than as an additional visualization layer applied after model development.

Among post-hoc explanation approaches, SHapley Additive exPlanations (SHAP) has become prominent because it provides an additive attribution framework based on Shapley values from cooperative game theory, enabling the contribution of individual features to a model prediction to be quantified relative to a baseline output (Lundberg & Lee, 2017). For tree-based models, TreeExplainer further enables computationally efficient attribution and supports the transition from individual prediction explanations toward aggregated global understanding of model behavior (Lundberg et al., 2020). This methodological capability has stimulated applications of SHAP across financial fraud detection, credit-card fraud, bankruptcy prediction, and other financial-risk settings, where feature attribution can make predictive outputs more traceable to domain users (Nguyen et al., 2025; Selvam & Sughasiny, 2025; Yanuangga, 2025). Recent studies have also demonstrated that SHAP can be integrated with ensemble learning to investigate model contribution dynamics and improve the interpretability of fraud-oriented predictive systems (Thanathamathee et al., 2024; Erasmus & Paepae, 2026). In audit-related research, SHAP has similarly been employed to interpret machine-learning predictions of audit opinions, suggesting that explainability can connect predictive performance with information that is potentially meaningful for audit decision-making (Ashtari et al., 2023). Group-based SHAP analysis further illustrates how feature contributions can be examined beyond isolated observations, providing a more structured perspective on the behavior of financial fraud detection models (Lin & Gao, 2022). These developments indicate that SHAP is not merely a mechanism for producing explanatory graphics, but can serve as an analytical bridge between ensemble prediction, feature contribution, and domain-level interpretation.

Despite this progress, the presence of high predictive performance and apparently coherent SHAP visualizations does not automatically establish the scientific validity of an explainable classification experiment. SHAP explanations describe how a trained model uses the supplied predictors, meaning that an apparently dominant feature may simply reflect information that is already embedded in the target construction rather than an independently meaningful risk determinant. This concern is especially relevant to the UCI Audit Data, in which the original study classified firms according to fraud and non-fraud categories associated with the Audit Risk Score, creating a potential structural relationship between the predictor space and the target definition (Hooda et al., 2018). The dataset documentation likewise provides the original Audit Data as a public machine-learning dataset derived from audit-related observations, making it useful for reproducible experimentation while simultaneously requiring careful examination of how its variables and labels were constructed (Hooda, 2018). A model may therefore achieve exceptionally strong performance because it has learned a relationship that is intrinsically close to the labeling mechanism, while SHAP may subsequently identify the same variable as the most influential feature, producing an explanation that appears persuasive but does not necessarily demonstrate independent predictive discovery. Related applications of explainable machine learning in fraud and financial analytics have shown the value of SHAP for interpreting

complex models, but they do not eliminate the need to evaluate whether the explanatory features are temporally, conceptually, and causally appropriate for the prediction target (Mupenzi, 2026; Qazi & Jadoon, 2026). This distinction is critical because target leakage can transform an apparently successful predictive experiment into a demonstration that the model has reproduced information already encoded in the outcome variable. Experimental quality therefore depends not only on predictive metrics and explanation scores but also on an explicit audit of feature availability, target construction, and the structural relationships between predictors and labels.

The unresolved issue is consequently not whether ensemble models can classify audit-related risk or whether SHAP can generate interpretable feature attributions, but whether the resulting explanations remain analytically meaningful when predictive variables are closely connected to the definition of the target. Existing research has increasingly combined ensemble learning and SHAP to improve interpretability, including applications in financial fraud, ERP anomaly detection, credit-card fraud, and job-placement prediction, yet these studies generally emphasize model performance and explanatory patterns more strongly than the possibility that a dominant feature may be structurally responsible for the observed classification outcome (Sibagariang, 2025; Qazi & Jadoon, 2026). The same issue has implications for audit analytics because an explanation that identifies a score-derived variable as the primary determinant may provide little additional knowledge if that score is already embedded in the procedure used to construct the target class. Recent work on explainable financial and audit analytics reinforces the practical importance of interpretable models, but the coexistence of strong predictive performance and potential target leakage creates a methodological problem that cannot be resolved through SHAP visualization alone (Ashtari et al., 2023; Salih et al., 2025). Addressing this problem is scientifically important because it separates genuine evidence of predictive learning from the mechanical recovery of information contained in the labeling process. It is also practically relevant because auditors and organizational decision makers could otherwise interpret a highly accurate model as evidence of robust risk discrimination when its performance may depend on information that would not constitute an independent predictor in a real deployment setting. A rigorous preliminary investigation must therefore evaluate predictive performance and model explanations jointly with the provenance and construction of the target variable. This perspective creates a need for an explainability-oriented audit-risk experiment that treats model interpretation and experimental validity as interconnected rather than independent analytical stages.

Against this background, the present study positions itself as a preliminary and reproducible investigation of ensemble learning and SHAP-based interpretation for audit risk classification using a public audit dataset. The study compares Random Forest and XGBoost as representative ensemble approaches for classifying the Risk target and examines their predictive behavior through both global and local SHAP explanations using TreeExplainer. The analysis specifically investigates whether the apparent importance of Audit_Risk provides substantive evidence about audit-risk determinants or instead reflects a structural relationship between the predictor and the target construction. Particular attention is given to the relationship between near-perfect predictive performance, feature dominance, and the possibility of target leakage, allowing model accuracy to be interpreted in relation to experimental validity rather than as an isolated performance indicator. The theoretical contribution lies in framing explainability as an issue of model-and-data validity, where feature attribution must be interpreted in conjunction with the provenance and construction of the classification target. The methodological contribution lies in establishing a transparent proof-of-concept that integrates ensemble comparison, global and local SHAP analysis, and critical examination of target leakage within a reproducible audit-risk classification setting. The study ultimately provides an empirical foundation for refining feature-selection, target-definition, and experimental-design strategies in subsequent research on intelligent and explainable audit-risk analytics.

RESEARCH METHODS

This study employed an empirical experimental design using the UCI Audit Data, a public audit dataset originating from the Auditor Office of India for classifying suspicious firms (Hooda et al., 2018). The experimental architecture consisted of dataset acquisition, data-quality inspection, preprocessing, ensemble model development, predictive evaluation, and SHAP-based interpretation, with Python implemented through a Jupyter Notebook as the computational environment. The dataset object processed in the experiment contained 776 observations and 27 columns, although the UCI repository

metadata reports 777 instances, and the analysis therefore followed the audit_risk.csv file actually processed in the notebook. Random Forest and XGBoost were implemented as the ensemble classification models, followed by SHAP TreeExplainer for interpreting the XGBoost predictions.

Data Pre-processing and Model Implementation

Data-quality inspection identified one missing value in Money_Value, which was imputed using the median value of 0.09, while LOCATION_ID was removed because it represented a non-numeric identifier rather than an analytical predictor. After preprocessing, the dataset contained 25 numerical predictors and one binary target, Risk, consisting of 471 observations classified as Risk = 0 and 305 observations classified as Risk = 1. The data were divided using an 80:20 stratified split with test_size = 0.2, stratify = y, and random_state = 42, producing 620 training observations and 156 test observations. Random Forest was configured with n_estimators = 200, max_depth = 6, class_weight = balanced, and random_state = 42, whereas XGBoost used n_estimators = 200, max_depth = 4, learning_rate = 0.1, eval_metric = logloss, and random_state = 42 (Chen & Guestrin, 2016).

Table 1. Summary of the dataset and data preprocessing

Component	Experimental Result
Raw data size	776 observations × 27 columns
Missing values	1 value in Money_Value
Imputation strategy	Median = 0.09
Removed column	LOCATION_ID (non-numeric/identifier)
Final data	776 observations; 25 predictors + 1 target
Risk target distribution	Risk = 0: 471; Risk = 1: 305
Data split	Stratified 80:20; random_state = 42
Training / test data	620 / 156 observations

Table 2. Model configurations used in the study

Model	Main Configuration
Random Forest	n_estimators = 200; max_depth = 6; class_weight = balanced; random_state = 42
XGBoost	n_estimators = 200; max_depth = 4; learning_rate = 0.1; eval_metric = logloss; random_state = 42

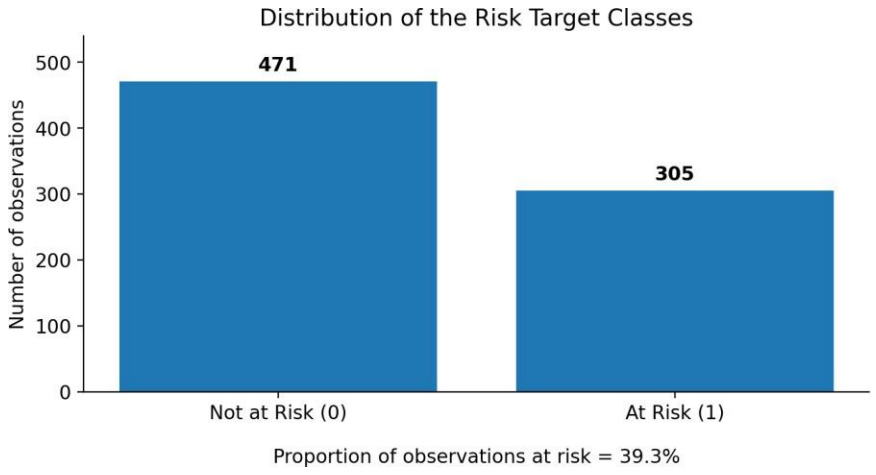

Figure 1. Distribution of the Risk target classes in the experimental dataset

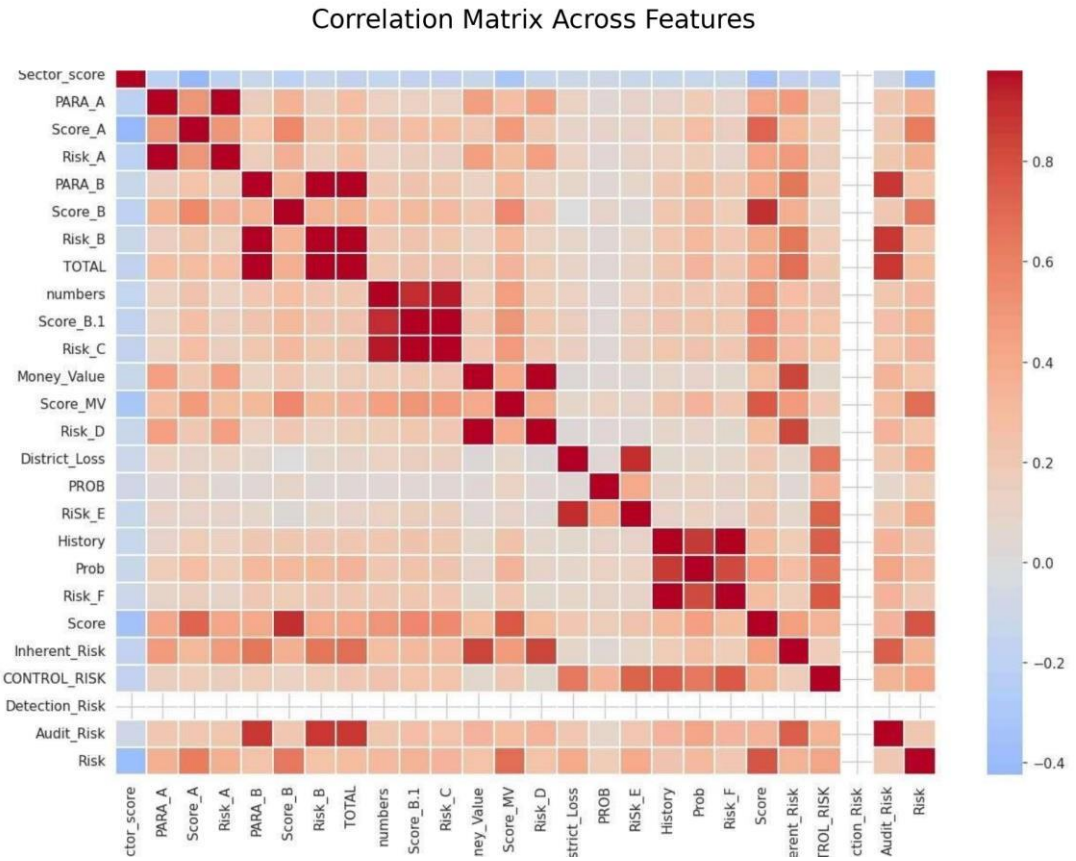

Figure 2. Correlation matrix of the numerical features in the UCI Audit Data

Experimental Testing, Validation, and SHAP Evaluation

Model performance was evaluated on the held-out test set of 156 observations using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices to assess overall classification performance and the detection of the at-risk class. Since no hyperparameter search or cross-validation was performed, validation relied on the fixed stratified hold-out procedure with random_state = 42 to maintain reproducibility. SHAP analysis was conducted for the XGBoost model using shap.TreeExplainer, generating SHAP values for all 156 test observations and 25 predictors, with global interpretation based on the summary plot and mean absolute SHAP values and local interpretation based on a waterfall plot for one at-risk observation (Lundberg et al., 2020). Under the implemented configuration, TreeExplainer produced raw-margin or log-odds outputs, while the corresponding probability was obtained through the logistic transformation, and the resulting feature contributions were examined alongside model performance to assess whether dominant predictors represented meaningful audit-risk information or potentially reflected the construction of the target variable (Lundberg & Lee, 2017).

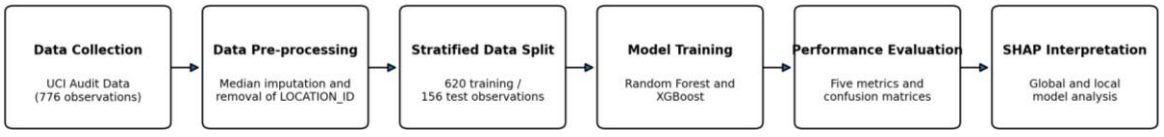

Figure 3. Experimental methodology and SHAP explanation-extraction workflow

RESULTS AND DISCUSSION

Model Performance and Classification Results

The experimental results demonstrate that both ensemble models achieved exceptionally high classification performance on the held-out test data, with Random Forest correctly classifying all 156 observations and XGBoost misclassifying only one observation. Random Forest obtained an accuracy,

precision, recall, F1-score, and ROC-AUC of 1.0000, indicating that no false-positive or false-negative classification occurred within the evaluated test subset. XGBoost achieved an accuracy of 0.9936, precision of 1.0000, recall of 0.9836, F1-score of 0.9917, and ROC-AUC of 1.0000. The small difference between the two models indicates that both algorithms were able to identify an extremely clear separation between the two Risk classes under the experimental data configuration. Such performance is consistent with the ability of tree-based ensemble methods to capture nonlinear relationships and interactions among structured predictors without requiring explicit specification of functional forms (Chen & Guestrin, 2016). From an intelligent-computing perspective, the results confirm that the implemented ensemble pipeline was capable of learning the classification structure present in the processed audit dataset. The magnitude of the performance, however, requires further examination because exceptionally high predictive scores may reflect properties of the dataset and target construction rather than generalizable discrimination capability.

The detailed evaluation results in Table 3 show that Random Forest maintained perfect values across all five evaluation metrics, whereas the only reduction for XGBoost occurred in recall and consequently in F1-score. Precision remained 1.0000 for XGBoost, indicating that every observation predicted as at risk belonged to the positive class in the test set. The recall value of 0.9836 indicates that one actual at-risk observation was not detected by XGBoost, producing the single false-negative error observed in the experiment. The difference between the models is therefore concentrated on sensitivity toward the positive class rather than on false-positive classification. This distinction is particularly relevant in audit-risk classification because a false negative represents an at-risk entity that is classified as not at risk and could consequently receive less analytical attention than warranted. The results illustrate why multiple evaluation metrics are necessary when assessing classification systems, since an accuracy value of 0.9936 alone would not reveal the specific weakness represented by the missed positive observation. The use of precision, recall, F1-score, and ROC-AUC provides a more informative assessment of the classification behavior than accuracy alone, particularly when the objective involves identifying potentially risky observations.

Table 3. Comparison of model-evaluation metrics on the test set

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	1.0000	1.0000	1.0000	1.0000	1.0000
XGBoost	0.9936	1.0000	0.9836	0.9917	1.0000

The confusion matrices in Figure 4 provide a more direct representation of the classification behavior underlying the metrics in Table 3. Random Forest produced 471 true-negative and 305 true-positive classifications, with zero false positives and zero false negatives. XGBoost produced the same number of correctly classified observations except for one false negative among the 305 at-risk observations, while no false-positive classification was recorded. This pattern explains why the precision of XGBoost remained perfect even though its recall declined, since the model did not incorrectly assign the at-risk class to any observation belonging to the not-at-risk class. The absence of false positives suggests a highly distinct decision boundary within the observed test sample, while the isolated false negative indicates that XGBoost encountered one observation whose predictor configuration did not fully conform to the learned positive-class pattern. Similar applications of machine learning to audit and financial-risk prediction have demonstrated the usefulness of ensemble approaches for identifying complex risk patterns, although predictive performance needs to be interpreted in relation to the characteristics of the underlying data (Resul & Ganji, 2025).

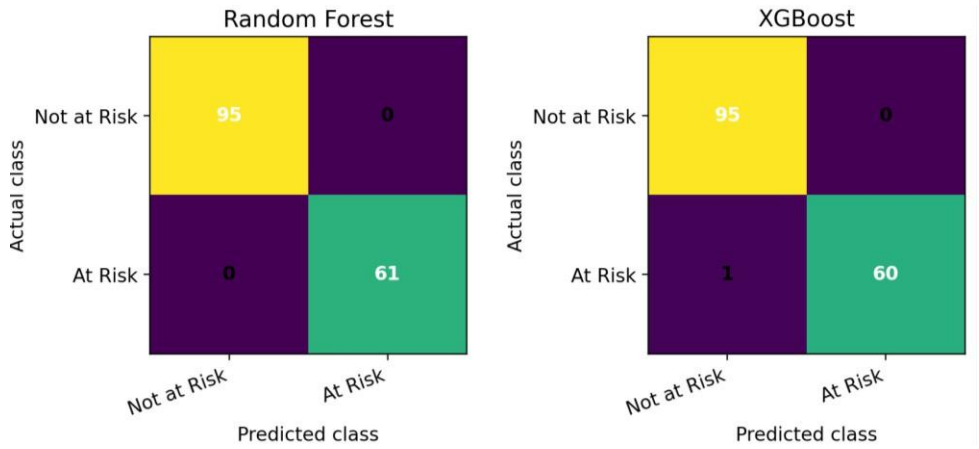

Figure 4. Confusion matrices for Random Forest and XGBoost on the test set

The comparative performance also reflects differences in the structural learning mechanisms of the two ensemble algorithms. Random Forest aggregates predictions from multiple decision trees constructed through randomized sampling and feature selection, which can produce stable classification boundaries when the available predictors provide highly discriminative signals. XGBoost instead builds trees sequentially by emphasizing observations associated with previous prediction errors, allowing the model to refine its decision function through gradient boosting. In this experiment, both mechanisms converged toward almost identical classification outcomes despite their different ensemble structures, suggesting that the dominant information required for separating the Risk classes was readily accessible to both algorithms. The result is consistent with prior research demonstrating the applicability of XGBoost to interpretable financial and audit-oriented prediction tasks, where nonlinear relationships among financial variables can be captured effectively by boosted-tree architectures (Ashtari et al., 2023). The marginal advantage of Random Forest in the present test split should therefore not be interpreted as evidence that Random Forest is intrinsically superior to XGBoost for audit-risk classification. The comparison instead establishes that the classification signal embedded in this dataset was sufficiently strong for two different ensemble mechanisms to reproduce nearly the same decision structure.

The perfect performance of Random Forest and near-perfect performance of XGBoost should also be interpreted cautiously because the experiment used only a single 80:20 hold-out split without cross-validation or external validation. A single test partition provides evidence about model behavior on the selected observations but does not quantify how performance would vary across alternative samples drawn from the same population. The relatively small dataset, consisting of only 776 processed observations, further limits the extent to which the reported metrics can be generalized to broader audit populations. More importantly, the predictor set contains derived risk-related variables, including Audit_Risk, that are structurally close to the definition of the Risk outcome. The original audit dataset was developed for fraudulent-firm classification using audit-related risk measures, creating a possibility that some predictors contain information that is directly or indirectly associated with the labeling mechanism (Hooda et al., 2018). Under such circumstances, a model can obtain extremely high predictive performance by reproducing an existing scoring relationship rather than discovering an independently generalizable pattern of audit risk. This issue is particularly important in explainable machine learning because feature-attribution methods can faithfully reveal the model's reliance on such variables without establishing that the variables represent causally meaningful determinants of the target (Arrieta et al., 2020).

The classification results are therefore best interpreted as evidence that the experimental pipeline successfully learned the separation encoded in the processed UCI Audit Data rather than as definitive evidence of exceptional real-world audit-risk prediction. The perfect Random Forest result and the single-error XGBoost result establish a strong baseline for subsequent interpretation, but their scientific value depends on identifying what information the models used to achieve this separation. In particular, the convergence of two different ensemble methods toward near-perfect discrimination raises the question of whether a small number of highly informative variables dominate the prediction process. This question is central to the present study because model performance alone cannot distinguish

genuine predictive structure from information that is structurally embedded in the target definition. Explainability methods are valuable in this setting when they expose the features responsible for individual and aggregate predictions, allowing unusually strong predictive behavior to be examined rather than accepted at face value (Burkart & Huber, 2021). The subsequent SHAP analysis therefore serves not merely to visualize feature importance but to investigate the computational basis of the observed classification performance. Particular attention is required for Audit_Risk, whose relationship with the target construction provides a potential explanation for the exceptionally high scores reported in this experiment.

Global Interpretability of the XGBoost Model

The global SHAP analysis provides a more detailed explanation of the predictive behavior that produced the near-perfect classification results reported in the previous section. Figure 5 places Audit_Risk as the most influential predictor, with a substantially wider SHAP contribution range than the remaining variables across the 156 test observations. The distribution of SHAP values indicates that low values of Audit_Risk generally shifted the model output toward the not-at-risk class, whereas higher values shifted the output toward the at-risk class. This directional consistency suggests that the XGBoost model did not distribute its predictive information evenly across the feature space but instead relied heavily on a small number of variables with strong discriminatory capacity. Such global attribution is consistent with the role of SHAP in transforming complex tree-based predictions into feature-level contributions that can be examined across individual observations and aggregated at the dataset level (Lundberg et al., 2020). The finding is particularly relevant because the primary objective of global interpretation is not merely to identify the largest feature but to determine whether the learned decision structure is technically and substantively reasonable. In this case, the pronounced dominance of Audit_Risk provides the first indication that the exceptional predictive performance may be closely associated with the construction of the audit-risk variables themselves.

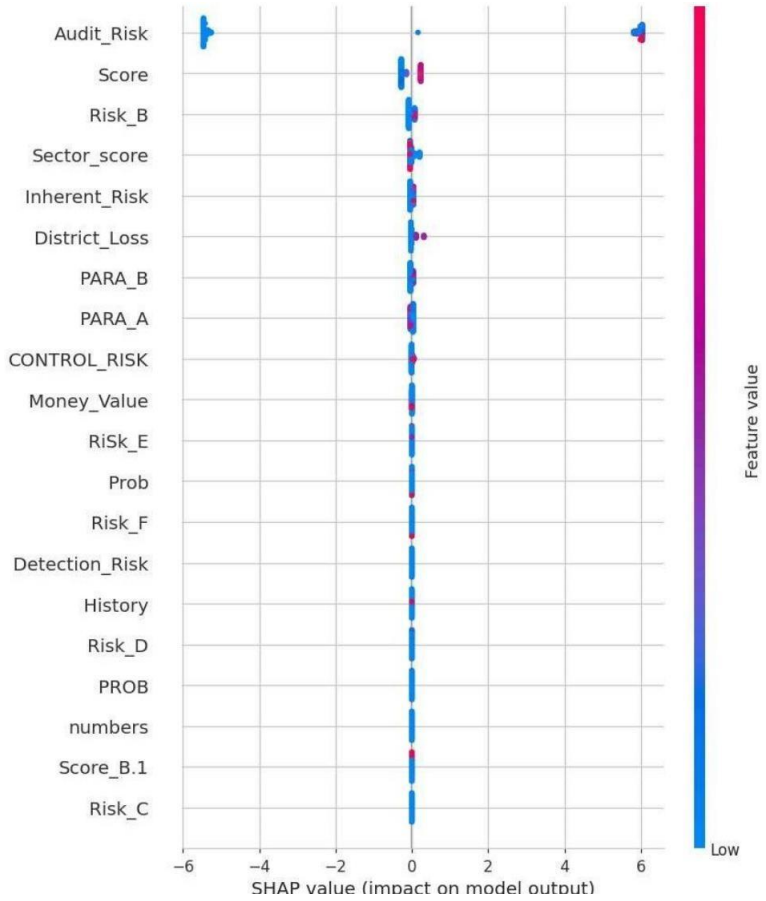

Figure 5. SHAP summary plot for the XGBoost model across 156 test observations

The remaining features in Figure 5 form a substantially lower contribution group consisting of Score, Risk_B, Sector_score, Inherent_Risk, District_Loss, PARA_B, PARA_A, and CONTROL_RISK. Score occupies the second position, followed by Risk_B and Sector_score, while the remaining variables contribute within comparatively narrower SHAP ranges. The relative ordering indicates that the model retained information from several audit-related variables, but their aggregate influence was considerably smaller than that of Audit_Risk. The presence of Inherent_Risk among the higher-ranked predictors is technically plausible because inherent risk represents an important component of audit-risk assessment, while the contribution of sector- and district-related variables suggests that contextual characteristics also entered the learned decision structure. Similar feature-attribution patterns have been reported in financial fraud applications, where SHAP has been used to distinguish the relative contribution of multiple financial indicators within complex predictive models (Lin & Gao, 2022). The ranking should nevertheless be interpreted as model-specific importance rather than evidence that these variables possess equivalent causal or substantive importance in audit practice. The distinction matters because SHAP quantifies the contribution of a predictor to the trained model's output under the observed feature configuration, rather than establishing that the predictor independently determines the underlying audit-risk condition.

The mean absolute SHAP ranking shown in Figure 6 reinforces the pattern observed in the summary plot by placing Audit_Risk substantially above the other predictors. Mean absolute SHAP values summarize the magnitude of a feature's contribution without considering whether the contribution moves the prediction toward Risk = 0 or Risk = 1, making the measure useful for comparing the overall influence of predictors across the test observations. The consistency between the summary plot and the mean |SHAP| ranking indicates that the dominance of Audit_Risk is not attributable to a small number of isolated observations but represents a recurring component of the XGBoost decision function. This type of global aggregation is aligned with the broader XAI objective of moving from individual explanations toward an understanding of model behavior across a population of observations (Arrieta et al., 2020). The result also demonstrates the practical value of SHAP in ensemble learning because conventional tree-based feature importance alone may not clearly communicate the direction and magnitude of individual contributions. In the present experiment, the combination of ranking and directional information makes it possible to distinguish the existence of feature dominance from the direction in which that feature affects the classification output.

Figure 6. Feature-importance ranking based on mean |SHAP value|

The strong contribution of Audit_Risk requires particular attention because the variable is not an independent raw characteristic of an audited firm but a derived risk measure within the dataset. The original UCI Audit Data documentation describes the Audit Risk Score as being calculated from components associated with inherent risk, control risk, and detection risk, and the original fraudulent-firm classification procedure is closely connected to the resulting audit-risk assessment (Hooda et al., 2018). The global SHAP pattern therefore indicates more than simple feature importance because it reveals that the XGBoost model has identified a variable closely associated with the mechanism underlying the target classification. From a computational perspective, this explains why two structurally different ensemble algorithms could achieve almost identical performance on the same test set. From a methodological perspective, however, the result raises concern that the model may be learning a representation of the target-generation procedure rather than discovering an independent pattern of audit risk. Comparable applications of SHAP in financial and fraud analytics demonstrate that explainability can expose dominant model features, but interpretation still depends on whether those features are legitimately available and conceptually separated from the prediction target (Thanathamathee et al., 2024). The observed feature dominance should therefore be treated as diagnostic evidence that motivates a closer examination of target construction rather than as direct evidence that Audit_Risk is the most important substantive determinant of fraud risk.

The distinction between model explanation and substantive interpretation is central to evaluating the present result. SHAP can accurately identify which variables the XGBoost model uses to produce a prediction, but it cannot determine whether those variables should have been included in the predictive experiment in the first place. This limitation is especially important when a dataset contains composite scores, derived indicators, or variables generated from procedures closely related to the target label. Previous research on explainable machine learning emphasizes that explanations should be interpreted within the context of model assumptions, data-generating processes, and the intended decision environment rather than treated as independent evidence of causal relationships (Minh et al., 2022). In audit-risk classification, the distinction is particularly consequential because a highly interpretable model can still produce a methodologically problematic explanation if its dominant predictor contains information that would not be legitimately available at the point of prediction. The present global SHAP results therefore support the interpretation of XGBoost behavior while simultaneously exposing a potential weakness in the experimental feature design. This dual role makes SHAP particularly valuable for the current preliminary study because the method identifies not only influential variables but also an empirical signal that requires further investigation through leakage-controlled experiments.

The global interpretation consequently establishes a coherent relationship between predictive performance and feature attribution in the experimental dataset. The near-perfect classification results are accompanied by a highly concentrated attribution structure in which Audit_Risk accounts for a markedly larger portion of the model's decision behavior than the remaining predictors. This pattern differs from a situation in which predictive performance is distributed across numerous relatively independent features, because the observed concentration suggests that the classification boundary may be governed primarily by a derived risk signal. Research applying SHAP to bankruptcy prediction and other financial-risk problems similarly demonstrates that feature attribution can transform ensemble models into more traceable analytical instruments, although the substantive meaning of the dominant variables remains dependent on the underlying data structure (Nguyen et al., 2025). The present findings therefore support the technical effectiveness of SHAP for global interpretation while placing an important qualification on the interpretation of feature importance. Audit_Risk should not yet be described as an independently validated determinant of the Risk target because its construction is closely connected to the audit-risk framework underlying the dataset. The global explanation instead provides a basis for investigating whether the observed dominance persists after potentially leakage-prone variables are removed. Such an investigation becomes particularly important before the model can be considered suitable for broader audit-risk prediction beyond the experimental dataset.

Local Interpretability of the XGBoost Model

The local SHAP analysis provides a more granular view of how the XGBoost model generated an individual at-risk prediction, complementing the global feature-ranking pattern presented in Section

4.2. For the second observation in the test set, the baseline value was , while the final model output reached , indicating a substantial movement of the prediction from the baseline toward the at-risk class. Audit_Risk was the dominant contributor, with a feature value of 1.444 producing a SHAP contribution of approximately +5.82. This contribution was substantially larger than the effects of all other predictors, demonstrating that the individual prediction was driven primarily by a single feature rather than by an evenly distributed combination of audit-related variables. The magnitude and direction of SHAP contributions provide an additive representation of the model decision, enabling the contribution of each predictor to be examined relative to the model's baseline output (Lundberg & Lee, 2017). Such local attribution is particularly relevant in audit-oriented applications because an individual classification can be examined in terms of the evidence used by the computational model rather than being presented solely as an unexplained class label.

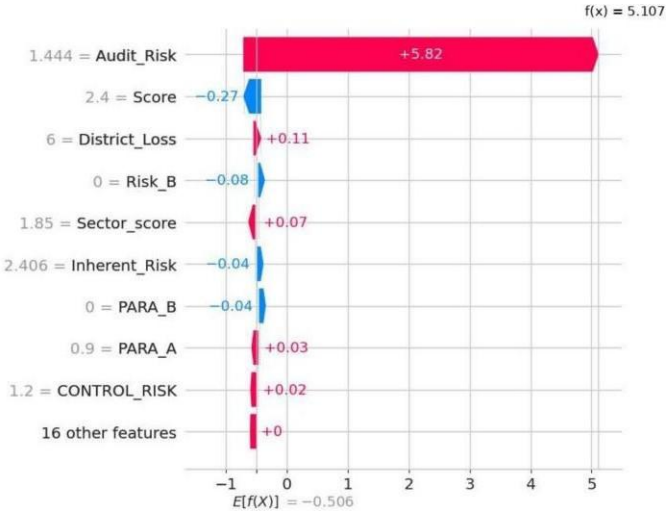

Figure 7. Waterfall plot providing a local explanation for the second test observation

The waterfall plot in Figure 7 shows that Audit_Risk shifted the prediction strongly toward the at-risk class, whereas several other predictors exerted comparatively small positive or negative effects. Score contributed −0.27, District_Loss contributed +0.11, and Sector_score contributed +0.07, while Risk_B and Inherent_Risk contributed −0.08 and −0.04, respectively. The remaining variables, including PARA_B, PARA_A, and CONTROL_RISK, produced contributions close to zero relative to the dominant effect of Audit_Risk. This pattern indicates that the local decision was not produced through a balanced accumulation of weak signals but through a highly concentrated contribution structure. Research applying SHAP to financial fraud detection has similarly demonstrated that local explanations can identify the specific variables responsible for individual predictions and provide greater transparency than aggregate feature-importance measures alone (Selvam & Sughasiny, 2025). The result is important for interpreting individual audit-risk predictions because a practitioner examining this observation would receive a clear computational explanation of why the model assigned it to the at-risk category. At the same time, the concentration of attribution on Audit_Risk requires scrutiny because transparency about a model's reasoning does not establish that the dominant information represents an independently valid basis for classification.

Table 4. Local feature contributions for the second test observation

Feature	Feature value	SHAP contribution
Audit_Risk	1.444	+5.82
Score	2.4	-0.27
District_Loss	6	+0.11
Risk_B	0	-0.08
Sector_score	1.85	+0.07
Inherent_Risk	2.406	-0.04

PARA_B	0	-0.04
PARA_A	0.9	+0.03
CONTROL_RISK	1.2	+0.02
16 other features	-	approximately 0.00

The numerical details in Table 4 clarify the extent to which the individual prediction was dominated by Audit_Risk. Its contribution of +5.82 is several orders of magnitude larger than the contributions of the other listed variables, while the combined effects of the remaining predictors make only a limited adjustment to the final model output. Score was the most influential opposing feature, but its negative contribution of −0.27 was insufficient to offset the positive contribution associated with Audit_Risk. The positive effects of District_Loss, Sector_score, and PARA_A further strengthened the at-risk direction, although their magnitudes were comparatively modest. This configuration illustrates the additive nature of SHAP explanations, where the final prediction can be decomposed into a baseline component and feature-specific contributions rather than interpreted solely through the final class label (Salih et al., 2025). The local result therefore confirms the global observation that Audit_Risk occupies a structurally dominant position within the learned decision function. The consistency between the individual-level contribution and the global ranking strengthens the interpretation that the feature's importance is systematic rather than an artifact of the aggregation procedure.

The numerical output of the waterfall plot also requires careful interpretation because the value represents the model's raw margin rather than a direct probability under the implemented TreeExplainer configuration. Applying the logistic transformation converts this value into an estimated probability of approximately 0.994 for the at-risk class. The baseline value of −0.506 corresponds to a probability of approximately 0.376, indicating that the local feature contributions substantially shifted the model from a relatively lower baseline tendency toward a highly confident at-risk prediction. The transition is therefore not simply a change in class label but a substantial movement in the model's decision score caused primarily by the dominant positive contribution of Audit_Risk. TreeExplainer is designed to provide additive explanations for tree-based models and can connect local prediction behavior with broader patterns in model output (Lundberg et al., 2020). Accurate interpretation of the displayed scale is important because treating the raw margin as a probability would overstate the semantic meaning of the numerical value represented in the waterfall plot. The distinction also demonstrates why technical knowledge of the explanation mechanism is necessary when SHAP visualizations are used as evidence in empirical machine-learning research.

The local explanation provides an additional indication of potential structural dependence between the model and the target-generation mechanism. A feature that contributes +5.82 to an individual prediction while most other variables contribute only marginally suggests that the classifier can determine the observation's class largely from information encoded in Audit_Risk. The original audit dataset was developed around audit-risk measures used for fraudulent-firm classification, meaning that the relationship between derived risk scores and the target cannot be considered independent of the labeling process (Hooda et al., 2018). In this context, the local explanation is valuable precisely because it makes the potential problem visible at the observation level rather than leaving it hidden behind the aggregate accuracy of 0.9936. Explainability research emphasizes that an explanation describes the behavior of the trained model, while the validity of the underlying predictive task depends on the quality and appropriateness of the data and target definition (Arrieta et al., 2020). The local result therefore should not be interpreted as evidence that Audit_Risk independently causes an entity to become risky. It instead indicates that the model has learned to associate the observed value of this derived variable very strongly with the target classification.

The agreement between global and local explanations provides a useful internal consistency check for the experimental interpretation. Globally, Audit_Risk was ranked substantially above the other predictors according to mean absolute SHAP values, while locally it generated the overwhelming positive contribution for the selected at-risk observation. The same feature consequently dominates both the aggregate explanation and the individual prediction, suggesting that the observed attribution pattern is embedded in the model rather than limited to one analytical perspective. This consistency is aligned with the broader objective of explainable ensemble learning, in which local explanations can be aggregated to develop an understanding of how a complex model behaves across observations (Qazi & Jadoon, 2026). From an audit-risk perspective, the result demonstrates the practical potential of SHAP

to trace a classification decision back to its principal computational drivers. From a methodological perspective, however, the same consistency reinforces the need to distinguish interpretability from predictive validity because a model can be highly transparent about a potentially problematic feature dependency. The local explanation therefore strengthens the diagnostic value of the study by showing concretely how the dominant global feature operates within an individual prediction. Further analysis of the dependence relationship between Audit_Risk and its SHAP contribution is required to determine whether the observed dominance represents a continuous risk relationship or a near-deterministic separation encoded in the dataset.

Audit_Risk Dependence Pattern and Structural Target Dependence

The dependence analysis further clarifies the relationship between Audit_Risk and the XGBoost prediction by examining how changes in the feature correspond to changes in its SHAP contribution across the test observations. Figure 8 shows an almost discrete pattern in which observations with very low Audit_Risk values are associated with SHAP values around −5.5, whereas observations outside this low-value region generally exhibit SHAP values close to +6. The transition occurs over a relatively narrow range, indicating that the model does not use Audit_Risk as a smoothly varying predictor in which incremental increases produce proportional changes in the prediction. Instead, the feature appears to function primarily as a strong separator between two regions of the model's decision space. This behavior is consistent with the ability of tree-based ensemble models to construct nonlinear threshold-based decision rules from structured data (Chen & Guestrin, 2016). The magnitude of the transition is particularly important because it provides a direct explanation for the near-perfect classification performance observed in the test set. The dependence pattern therefore shifts the interpretation from simple feature importance toward an examination of how the model operationalizes the information contained in Audit_Risk.

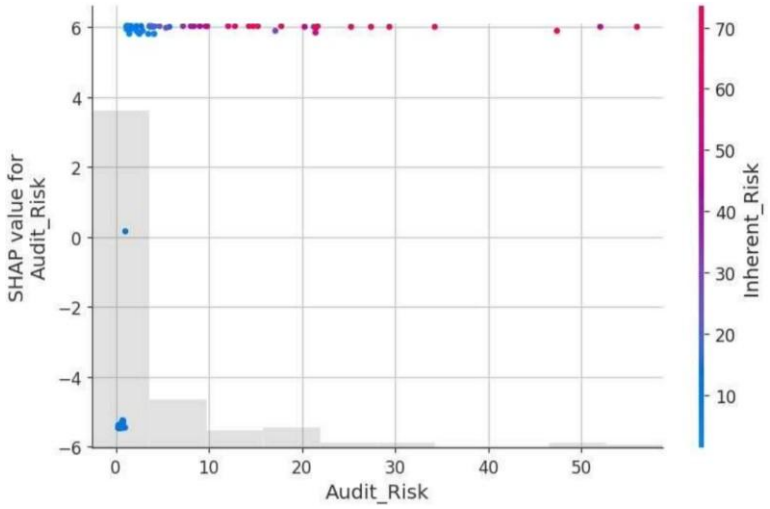

Figure 8. SHAP dependence plot for Audit_Risk, colored by the interaction with Inherent_Risk

The concentration of SHAP values around two opposing levels indicates that the XGBoost model has learned a highly concentrated decision boundary associated with Audit_Risk. Very low values move the model strongly toward Risk = 0, while values beyond the apparent transition region move the prediction strongly toward Risk = 1. Once the positive contribution reaches the upper range, additional increases in Audit_Risk produce relatively little additional change in SHAP magnitude, indicating a saturation effect rather than a continuous increase in predictive influence. This pattern is technically consistent with tree-based partitioning, where repeated splits can transform a continuous numerical variable into discrete decision regions. The finding also explains why the local observation examined in Section 4.3 generated a raw model output of 5.107, since its Audit_Risk value of 1.444 falls within the region associated with a strong positive contribution. SHAP is particularly useful in identifying this type of nonlinear behavior because the dependence representation connects feature values directly with their contribution to model output rather than reducing the relationship to a single

importance score (Lundberg et al., 2020). The observed threshold-like structure therefore provides stronger evidence of how the classifier makes decisions than the global ranking alone.

The interaction coloring by Inherent_Risk in Figure 8 provides additional information about the relationship between the principal predictor and another component of the audit-risk structure. Although the color variation indicates differences in Inherent_Risk across observations, the dominant separation in SHAP contribution remains organized around Audit_Risk rather than being primarily determined by the interaction variable. This suggests that Inherent_Risk may modify the model's decision context without replacing Audit_Risk as the principal discriminative signal. The finding is relevant because audit-risk variables are not necessarily independent constructs, and their relationships can influence how ensemble models distribute predictive attribution among correlated predictors. SHAP-based analyses of financial fraud models similarly demonstrate that interactions and correlated predictors can shape the apparent importance assigned to individual variables, requiring interpretation beyond a simple feature ranking (Lin & Gao, 2022). The observed interaction structure should therefore be viewed as evidence of a multivariable risk representation in which Audit_Risk remains dominant despite the presence of related risk components. Such a pattern is analytically plausible for the dataset but also increases the need to examine whether the variables collectively reproduce information used in the target-generation procedure.

The consistency between the dependence plot, global SHAP summary, and local waterfall explanation provides an important internal validation of the interpretability results. The global analysis identified Audit_Risk as the dominant feature, the local analysis showed that the same feature contributed +5.82 to an individual at-risk prediction, and the dependence analysis demonstrates that its contribution changes sharply according to the feature's value. These three levels of evidence describe the same model behavior from complementary perspectives rather than producing contradictory explanations. This coherence is consistent with the intended function of SHAP as a framework capable of connecting local feature attributions with broader model behavior for tree-based learning systems (Lundberg & Lee, 2017). The agreement is also important because an isolated feature-importance ranking could otherwise be interpreted as a statistical artifact or as a consequence of the aggregation procedure. In the present experiment, the repeated dominance of Audit_Risk across analytical representations indicates that the model's dependence on this variable is structurally embedded in its decision function. The interpretability results therefore provide a stronger basis for investigating the source of this dependence.

The methodological significance becomes more apparent when the dependence pattern is considered alongside the construction of the UCI Audit Data. The dataset documentation identifies Audit_Risk as a derived measure associated with the audit-risk calculation, while the classification task itself is based on fraudulent and non-fraudulent firm distinctions related to audit-risk assessment (Hooda et al., 2018). A feature that is mathematically or procedurally close to the target can provide information that is highly predictive without representing an independently available predictor at the intended prediction point. This condition can generate exceptionally high performance because the learning algorithm effectively reconstructs an existing decision rule rather than estimating a generalizable relationship from independent evidence. The SHAP dependence pattern strengthens this concern because the feature does not merely rank highly; it produces an almost binary contribution structure that closely mirrors the binary target separation. Similar concerns about interpretability arise when XAI methods are applied to high-stakes predictive systems without sufficient consideration of the relationship between feature construction and target definition (Rudin, 2019). The result indicates that explainability should function as part of experimental quality assessment rather than as a post hoc visualization procedure applied after predictive performance has been established.

The dependence pattern also has implications for the interpretation of the study's predictive results and its positioning as a preliminary investigation. The almost deterministic relationship between Audit_Risk and model contribution explains why Random Forest and XGBoost achieved accuracy values of 1.0000 and 0.9936, respectively, despite their different ensemble mechanisms. It also suggests that the predictive task may be considerably easier than a conventional audit-risk forecasting problem in which the target is defined independently from the predictors available at the prediction stage. Explainable machine learning can expose this type of structural characteristic because the attribution pattern reveals not only which variables are influential but also how their values affect the model output across observations (Minh et al., 2022). The present findings therefore support the technical reliability

of the implemented SHAP workflow while simultaneously limiting the substantive interpretation of the reported predictive performance. A leakage-controlled experiment that excludes Audit_Risk and other derived variables would provide a more rigorous test of whether the remaining audit attributes contain independent predictive information. Such feature ablation would also allow future research to distinguish genuine audit-risk classification capability from reconstruction of the existing scoring mechanism. The dependence analysis consequently serves as a bridge between model interpretability and experimental validity, providing an empirical basis for the limitations and research roadmap discussed in the subsequent section.

CONCLUSION

This study developed a proof-of-concept ensemble learning pipeline integrating Random Forest, XGBoost, and SHAP for interpreting audit-risk classification using the UCI Audit Data. The experiment processed 776 observations, with 620 observations allocated to training and 156 observations retained for testing through a stratified 80:20 split. Random Forest achieved 1.0000 for accuracy, precision, recall, F1-score, and ROC-AUC, whereas XGBoost achieved an accuracy of 0.9936, precision of 1.0000, recall of 0.9836, F1-score of 0.9917, and ROC-AUC of 1.0000. SHAP generated consistent global and local explanations, with Audit_Risk identified as the dominant predictor across the summary plot, mean absolute SHAP ranking, waterfall plot, and dependence analysis. For the selected at-risk observation, Audit_Risk produced a contribution of +5.82 and shifted the model's raw margin from −0.506 to 5.107. The dependence analysis further indicated that Audit_Risk operated as a strong threshold-like separator between the two risk classes. These findings confirm that the implemented ensemble and SHAP pipeline was technically capable of producing highly accurate audit-risk classifications while providing interpretable evidence regarding the features underlying the model predictions.

The principal finding of this study is not merely the exceptionally high predictive performance, but the ability of SHAP to expose the model's dependence on a feature closely related to target construction. The dominant contribution of Audit_Risk indicates potential structural target leakage because the source dataset uses the Audit Risk Score as part of the basis for distinguishing risk classifications. Consequently, the near-perfect performance should be interpreted primarily as evidence that the experimental pipeline successfully reproduced the classification structure contained in the processed dataset rather than as evidence of strong external generalization. The SHAP results demonstrate that explainable artificial intelligence can function as a diagnostic component of audit-risk modelling by revealing methodological characteristics that conventional performance metrics cannot identify independently. Future research should prioritize leakage-controlled feature ablation by removing Audit_Risk and other derived variables that may contain information closely connected to the target-generation mechanism. Subsequent experiments should also incorporate stratified cross-validation, probability calibration, SHAP stability assessment, and validation using external or domain-specific primary data to evaluate the robustness of both predictive performance and explanations. These developments can extend the present technical demonstration into a more valid, auditable, and methodologically defensible intelligent system for audit-risk classification and interpretation.

REFERENCES

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Ashtari, S., et al. (2023). Improving audit opinion prediction accuracy using metaheuristics-tuned XGBoost algorithm with interpretable results through SHAP value analysis. *Applied Soft Computing*, 149, 110955. https://doi.org/10.1016/j.asoc.2023.110955

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245-317. https://doi.org/10.1613/jair.1.12228

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Erasmus, N. C., & Paepae, T. (2026). An Explainability-Driven SHAP-Weighted Ensemble Framework for Fraud Detection: Insights into Model Contribution Dynamics. *Information, 17*(6), 607. https://doi.org/10.3390/info17060607

Hooda, N., Bawa, S., & Rana, P. S. (2018). Fraudulent Firm Classification: A Case Study of an External Audit. *Applied Artificial Intelligence, 32*(1), 48–64. https://doi.org/10.1080/08839514.2018.1451032

Hooda, N. (2018). Audit Data [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5930Q.

Lin, K., & Gao, Y. (2022). Model interpretability of financial fraud detection by group SHAP. *Expert Systems with Applications*, 210, 118354. https://doi.org/10.1016/j.eswa.2022.118354

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. https://dl.acm.org/doi/10.5555/3295222.3295230

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence, 2*(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review, 55*(5), 3503-3568. https://doi.org/10.1007/s10462-021-10088-y

Mupenzi, J. (2026). An Explainable Hybrid Machine Learning Framework for Financial and Tax Fraud Analytics in Emerging Economies. *Ultimatics: Jurnal Teknik Informatika, 17*(2), 194–202. https://doi.org/10.31937/ti.v17i2.4483

Nguyen, H. H., Viviani, J. L., & Ben Jabeur, S. (2025). Bankruptcy prediction using machine learning and Shapley additive explanations: H.-H. Nguyen et al. *Review of Quantitative Finance and Accounting, 65*(1), 107-148. https://doi.org/10.1007/s11156-023-01192-x

Qazi, A., & Jadoon, A. A. K. (2026). Explainable ensemble learning using SHAP for ERP anomaly detection. *Scientific Reports*. https://doi.org/10.1038/s41598-026-57913-4

Resul, A. P. A. K., & GANJI, F. (2025). Using decision tree algorithms and artificial intelligence to increase audit quality: A data-based approach to predicting financial risks. *International Journal of Business Management and Entrepreneurship, 4*(1), 87-99. https://mbajournal.ir/index.php/IJBME/article/view/65

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

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems, 7*(1), 2400304. https://doi.org/10.1002/aisy.202400304

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*. http://arxiv.org/abs/1708.08296

Selvam, S., & Sughasiny, M. (2025). Smart and Explainable Credit Card Fraud Detection Using XGBoost and SHAP. Journal of ISMAC, 7(2), 155–169. https://doi.org/10.36548/jismac.2025.2.004

Sibagariang, S. (2025). Interpretable Machine Learning for Job Placement Prediction: A SHAP-Based Feature Analysis. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi, 14*(3), 190-198. https://doi.org/10.22146/jnteti.v14i3.20516

Thanathamathee, P., Sawangarreerak, S., Chantamunee, S., & Mohd Nizam, D. N. (2024). SHAP-Instance Weighted and Anchor Explainable AI: Enhancing XGBoost for Financial Fraud Detection. *Emerging Science Journal, 8*(6), 2404–2430. https://doi.org/10.28991/ESJ-2024-08-06-016

Ulambayar, T., Pagjii, O., Luvsandash, O., Garamdorj, G., & Sambuu, U. (2026). Audit Detection Risk Reduction Using Machine Learning: Evidence from 3.3 million Transactions. Research Square https://doi.org/10.21203/rs.3.rs-9968348/v1

Yanuangga G. (2025). Explainable AI (XAI) Analysis Using SHAP for Credit Card Fraud. Scripta-Technica, 1(2). https://doi.org/10.65310/scxk4755