Communication in Physical Sciences, 2022, 8(4):830-848

The Use of Supervised and Unsupervised Learning Methods for Detecting Auditing Anomalies

Olatunde Ayeomoni

Received: 12 November 2022/Accepted: 22 December 2022/Published: 30 December 2022

Abstract: *The rapid growth in transaction volumes and increasing sophistication of fraudulent schemes have reduced the effectiveness of traditional audit sampling in modern financial systems. This study evaluates supervised and unsupervised machine learning techniques for audit anomaly detection and compares their effectiveness across anomaly types and organizational contexts. Using a real-world dataset of 247,683 financial transactions, we implemented five supervised algorithms—Logistic Regression, Random Forest, Support Vector Machines, XGBoost, and Artificial Neural Networks—and five unsupervised methods—K-means, DBSCAN, Isolation Forest, Local Outlier Factor, and Autoencoders. Supervised models were evaluated using precision, recall, F1-score, and AUC-ROC, while unsupervised models were assessed using detection rate, false positive rate, anomaly scores, and silhouette coefficients. Results show that ensemble supervised techniques perform best when labeled data are available. XGBoost achieved the highest performance with an F1-score of 0.89 and AUC of 0.94, while maintaining a false positive rate below 1%. Among unsupervised approaches, Isolation Forest achieved an 87.3% detection rate with a 4.2% false positive rate, and autoencoders demonstrated competitive performance in high-dimensional anomaly detection. The findings indicate that supervised models are more accurate for detecting known fraud patterns, whereas unsupervised methods are essential for identifying emerging or previously unseen anomalies, particularly in environments with limited labeled data. The study supports the adoption of hybrid audit analytics frameworks that combine both paradigms to improve detection coverage while balancing interpretability, computational efficiency, and operational feasibility. These results provide practical guidance for auditors implementing machine learning–based anomaly detection systems.*

Keywords*: Audit anomaly detection; Machine learning; Supervised learning; Unsupervised learning; Fraud detection*

Olatunde Ayeomoni.
University of Cincinnati, School of Information Technology, Cincinnati, Ohio , USA
Email: olatundeayeomoni@gmail.com

1.0 Introduction

Financial auditing has undergone a significant transformation since the collapse of Enron and the enactment of the Sarbanes–Oxley Act in 2002, which strengthened requirements for internal controls and audit processes. Rezaee & Riley, 2010). However, even with these regulatory developments, organizations all over the world still have to struggle with corporate fraud, and the Association of Certified Fraud Examiners approximates that organizations miss a significant portion of their annual revenue to fraud (ACFE, 2020). Conventional audit practices of random sampling 1-5% of the transactions are ineffective in a time when businesses conduct millions of transactions per day on multi-faceted, distributed systems. The scale, velocity, and complexity of modern financial data exceed human analytical capacity, making the detection of rare fraudulent transactions increasingly difficult.

Machine learning offers a scalable solution to this challenge by enabling continuous analysis of entire transaction populations rather than limited audit samples. These algorithms learn patterns of normal and abnormal financial behavior at a scale far beyond human capability, enabling earlier detection of suspicious activities. (Jans *et al*., 2014). Nevertheless, applying machine learning to auditing situations has subtle complications that go beyond applying algorithms to the data. The auditors are facing drastically unbalanced datasets with less than 1 percent fraudulent transactions out of all the volumes, and balancing the conflicting goals of fraud detection and operational interference caused by false positives (Perols, 2011).

The machine learning environment in anomaly detection faces two basic paradigms of different strengths and weaknesses. Supervised methods of learning need historical records with labeled anomalies, which are found and confirmed so that the algorithms can acquire discriminating patterns that can distinguish between legitimate and suspicious transactions (Ngai *et al*., 2011). Such techniques have shown amazing results in a controlled environment, and research studies have recorded detection rates of more than 90 percent with regard to some types of fraud. However, the supervised methods are severely restricted within the auditing settings. The dependence on historical labels opens an opportunity to new fraudulent schemes which are not the same as those observed in the past, and the scarcity of labeled fraudulent cases, which may represent only a tiny fraction of millions of legitimate transactions, Furthermore, the adversarial nature of fraud means that fraudsters continuously adapt their techniques, causing models trained on historical data to degrade over time (Phua *et al*., 2010).

Another paradigm proposed is unsupervised learning methods that detect anomalies by use of statistical deviation as compared to normal patterns without pre-labeled data (Chandola *et al*., 2009). These techniques are particularly valuable for uncovering previously unknown fraud schemes and adapting to emerging attack patterns. This freedom of unlabeled data also deals with a practical fact in most organizations: it is often not clear or even discoverable what historical transactions were actually fraudulent. Nonetheless, unsupervised methods present challenges of their own, especially in interpreting and validating identified anomalies. In the absence of labelled benchmarks, auditors have difficulty in separating actual fraud and legitimate but unusual transactions, a challenge that is further amplified in organizations with inherently diverse transaction patterns.

However, existing studies often emphasize algorithmic performance in isolation rather than conducting systematic cross-paradigm comparisons under realistic audit conditions. The available literature is fragmented, though it is quite abundant. Research usually analyzes single algorithms separately instead of making comparisons, and usually uses in-house datasets that are difficult to recreate or uses synthetic data, which is not always reflective of reality (West and Bhattacharya, 2016; Akinsanya *et al.,* 2022). Furthermore, a lot of work is based on computer science thinking that is focused on technical complexity rather than on audit practice, including interpretability, regulatory compliance, and integration with existing audit workflows.

This study addresses these gaps through a comprehensive empirical evaluation of supervised and unsupervised machine learning methods for audit anomaly detection. Our study has three important objectives. In the first step, we evaluate the performance of widely used and high-performing algorithms in the two paradigms based on a large real-life dataset that includes various types of transactions and organizational setup. Second, we perform comparative analysis to establish the circumstances under which each of the approaches proves to have benefits, and offer evidence-based suggestions to the practitioners. Third, we study the practical implementation factors such as computational efficiency, interpretability and false positive rates, which are critical to the practical viability of deployment but are not adequately covered in scholarly articles.

By providing a rigorous comparison grounded in realistic audit data and operational constraints, this study bridges the gap between machine learning research and practical audit implementation. The findings support evidence-based adoption of intelligent audit analytics and highlight how supervised and unsupervised approaches can be combined within comprehensive fraud detection frameworks.

The remainder of this paper is organized as follows. Part 2 provides the theoretical framework, de reviewing the concepts of anomaly detection, machine learning paradigms, and other work in audit analytics. Section 3 presents our methodology, such as a description of the datasets, specifications of the algorithms, and metrics of evaluation. Section 4 shows the comparison of the results of the methods of supervision and unsupervision as well as practical implications. Section 5 of the paper will finish with an audit practice recommendation and future research directions.

2.0 Theoretical Framework

2.1 Audit Anomaly Detection: Concept and Definition

Audit anomalies refer to deviations from expected patterns in financial transactions, internal controls, or accounting records. Although fraud represents the most severe category, auditors must also detect anomalies arising from system malfunctions, human error, or policy violations, which, despite being unintentional, can still compromise financial integrity. (Arens *et al*., 2016). Audit anomalies are commonly classified into three major categories. Fraudulent activities are those acts of intentional manipulation to obtain a personal benefit, such as expense reimbursement fraud, procurement kickback schemes, or more complex practices like revenue recognition manipulation and asset misappropriation. Errors refer to the second category and include unintentional mistakes in entering data, performing a calculation, or classifying the data which nonetheless result in misstatements in financial reporting. The third category is irregularities and comprises transactions that break organizational policies or regulatory conditions without necessarily involving fraud, such as purchases over authorization limits or vendor payments that are not adequately documented.

Traditional audit sampling is rooted in statistical quality control principles originally developed for manufacturing and later adapted to financial auditing in the mid-twentieth century. (Cushing and Loebbecke, 1986). The approaches are generally random or stratified sampling to pick a sample of transactions that are then analyzed in greater detail, based on the implicit assumption that sample results are representative of the broader population. This assumption is, however, problematic for detecting anomalies. By definition, fraudulent transactions are rare, and often occur at rates well below 1%, i.e. even random samples of 5 percent of transactions can completely fail to detect fraudulent activities. In addition, fraudsters carefully design illegal activities in such a way that they can blend with legitimate transaction patterns to further lower the chances of detection due to the sampling (Singleton and Singleton, 2010).

Continuous auditing represents another paradigm that leverages automated systems (Alles *et al*., 2006). Naturally, this strategy is in line with machine learning applications, which can process large volumes of data to locate possible anomalies to be investigated by a human. The change in the sampling into the wholesome analysis makes the auditor's role shift from primary examination to exception-based investigation compared to a primary examiner, since instead of spreading the human resource on different transactions that are random the workforce is centered on the most suspicious transactions highlighted by algorithms.

2.2 Machine Learning Paradigms for Anomaly Detection

Supervised learning is conceptually based on learning from labeled examples (Hastie *et al*., 2009; Aboagye *et al.,* 2022). Algorithms identify patterns of differentiation between the classes given a training dataset in which each transaction is labeled with its status, legitimate or anomalous. These patterns are encoded into a trained model, which classifies new, unclassified transactions. This paradigm is superior when the abnormalities in history have regular and recognizable traits that distinguish them from typical transactions. As an illustration, expense fraud could be conducted in a systematic manner, where the amount is less than the approval limits, at a particular time period or in a particular department. These discriminating patterns can be learnt in supervised algorithms and generalized by the algorithm to identify similar future instances.

Nevertheless, supervised learning has a number of limitations in its application in audit settings. The need to have labeled training data poses an immediate practical problem because there might be no extensive fraud labels available to organizations or the labels of doubtful quality are available. Algorithms are prone to predicting the majority class, which is the case even in the presence of labels where legitimate transactions are by far more frequent than anomalies, so the algorithm can learn to label all the data as legitimate (He and Garcia, 2009). Moreover, supervised models have limited ability to identify new types of frauds, which are not similar to past trends. This weakness is especially alarming considering that fraud is an adaptive concept, meaning that fraudsters adjust their strategies based on detection procedures.

In unsupervised learning techniques, the labelling condition is assumed away by detecting aberrant data according to their relative statistical characteristics to the larger population (Goldstein and Uchida, 2016). The assumption behind this is that the legitimate transactions tend to follow patterns or cluster in the feature space, whereas the anomalies do not follow the pattern. Various approaches to unsupervised methods formulate this assumption in different ways. Distance-based approaches detect anomalies as points that are far apart in relation to their closest neighbors, density-based methods identify low-density zones and reconstruction-based ones, such as autoencoders, find cases that the model is unable to faithfully reconstruct.

The autonomy over the labels gives unsupervised approaches a unique benefit on the audit applications. They are able to learn previously unidentified patterns of fraud, organically evolve to changing transaction traits, and work under conditions where the ground truth is in question (Pang *et al*., 2021). Nevertheless, unsupervised methods often lack interpretability, since the statistical deviation that algorithms are sending alarms about could be due to valid business reasons but not fraud. This creates a greater number of false positives than highly trained supervised models and potentially overwhelming auditors with false positives and reducing trust in the system.

2.3 Related Literature on Audit Anomaly Detection

Initial use of computational methods in audit was in rule-based systems and statistical methods. The Law of Benford that characterizes the distribution of leading digits in natural numerical data has been used since as early as the 1990s to identify tampered financial data (Nigrini, 1996). While intuitive and interpretable, it often suffers from low sensitivity and high false positive rates in datasets where the underlying assumption of natural number generation is not true. Benford analysis has a limitation. The manufacturing-based statistical processes control techniques have also been used to identify abnormal trends in accounting ratios or frequency of transactions (Debreceny & Gray, 2010). Although useful, such methods are based on univariate or bivariate analysis and fail to capture complex multivariate fraud patterns.

The use of machine learning in fraud detection increased in the early 2000s, first in credit card and insurance fraud, and later in audit (Phua *et al*., 2010). Early success was offered by Bayesian networks modeling the probabilistic relationships between attributes of transactions and the possibility of fraud (Kirkos *et al*., 2007). Ensuring interpretability made decision trees and their ensemble versions popular, where the Random Forests proved to be effective in different kinds of fraud (West and Bhattacharya, 2016). Neural networks, despite concerns regarding computational complexity and “black-box” behavior, have shown promise, especially in complex patterns of fraud (Ngai *et al*., 2011). More recent studies have explored deep learning models such as convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs) for temporal sequence modeling (Hilal *et al*., 2022). Autoencoders are also proposed to detect unsupervised frauds by training to recreate regular patterns of transactions and indicating those that reconstruct with high reconstruction loss (Paula *et al*., 2016). A tree-based anomaly detector, Isolation Forest has shown specific success with high-dimensional audit information (Liu *et al*., 2012).

Despite the growing body of research, significant gaps remain. Algorithms are tested on one dataset in most of the cases, which makes them less general. There are limited direct comparisons between supervised and unsupervised methods and practitioners are not certain about trade-offs. Accuracy of detection is frequently prioritized over practical implementation like computational scalability, model interpretability, and false positive control, which play a vital role in practice (Jans *et al*., 2014). This paper fills these gaps by thoroughly assessing the two paradigms based on consistent data and measurements, and with particular emphasis on practical implementation considerations.

Fig. 1 illustrates our conceptual framework, which positions machine learning as a component of the broader audit ecosystem rather than a standalone solution. The paradigm highlights the complementary value of the supervised and unsupervised methods. Supervised approaches use institutional knowledge coded with historical labels with high-confidence recognition of known forms of anomalies. Unsupervised techniques work on unlabeled data streams continuously and identify new patterns that could be indicative of new fraud schemes. Both approaches generate prioritized alerts for human investigation and auditor feedback produces learning loops that lead to improvement in system performance over time. This hybrid structure recognizes that machine learning enhances, but does not replace, human judgment in auditing.

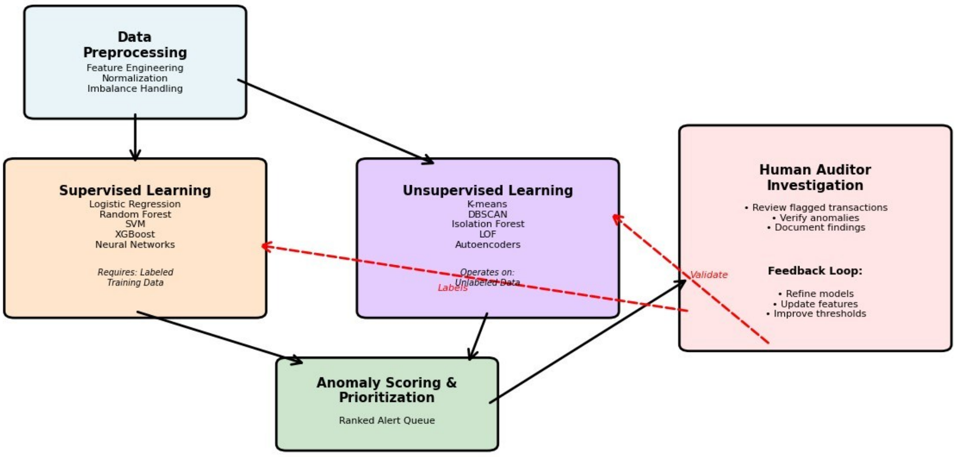

Fig. 1: Conceptual framework that demonstrates how the supervised and unsupervised machine learning methods can be incorporated in the audit process.

The framework indicates that data preprocessing feeds into parallel streams of supervised and unsupervised learning (respectively, which needs labeled and unlabeled data respectively). Both streams produce anomaly scores which are fed into a prioritized queue of investigation to be done by human auditors whose feedback triggers an improved version of supervised models and confirms the presence of unsupervised ones.

3.0 Methodology

3.1 Research Design and Data Collection

This empirical study uses a dataset of financial transactions from seven mid- to large-scale organizations in the manufacturing, retail, and service sectors, covering the period from 2018 to 2021. The database contains 247,683 transactions, including accounts payable, expense reimbursements, and procurement records. Anomalous transactions were identified from multiple sources: confirmed fraud cases of internal investigations (n=1,247), regulatory audit results (n=342), and suspicious transactions labeled by experts and identified during routine audits but not clearly proven fraudulent (n=1,876). This multi-source labeling reflects audit practice, where ground truth often exists along a spectrum rather than as binary certainty. The resulting anomaly rate of 1.4% is consistent with industry estimates and provides a sufficient number of positive cases for supervised learning. (Jans *et al*., 2014).

Each transaction record contains 28 features, including transaction attributes. (amount, date, category), vendor characteristics (history, risk score, location), employee attributes (department, seniority, authorization level), approval workflow (number of approvers, time in approval queue), and derived features (deviation of historical patterns, ratio to similar transactions, time-based patterns). Feature engineering was guided by domain knowledge of fraud indicators through incorporating variables like nearness to month-end, non-round amounts, as well as velocity of transactions with particular vendors or employees (Kirkos *et al*., 2007). Temporal characteristics store day of week and time of day patterns, which prior studies suggest may indicate unauthorized or high-risk activity occurring outside normal business hours.

Data preprocessing addressed several issues inherent in real-world financial datasets. Missing values, present in approximately 3% of records (primarily in vendor risk scores and some derived features), were handled using median imputation for numerical variables and mode imputation for categorical variables. We intentionally did not use more advanced imputation techniques like multiple imputation or model-based techniques because they may introduce spurious trends that will confound and bias anomaly detection. One-hot encoding was used to encode categorical features with low cardinality (e.g., department, transaction type) and frequency encoding was used for high-cardinality categorical features that were of high cardinality (such as vendor identifiers) without causing excessive dimensionality.

Algorithms sensitive to feature scale (e.g., Support Vector Machines and Neural Networks) used feature standardization (zero mean, unit variance) whereas the trees-based algorithms used raw features as they are invariant to monotonic feature transformations. The extreme imbalance of the classes (anomalies) of 1.4 percent of the transactions was solved with Synthetic Minority Over-sampling Technique (SMOTE) which produces artificial examples of minority classes in the feature space instead of merely in the existing anomalies (Chawla *et al*., 2002). However, SMOTE was applied only to the training data, leaving the validation and test sets with natural imbalance to provide performance measurements to realistic deployment conditions.

The structure and key characteristics of the dataset are summarized in Table 1. Algorithms have the ability to learn complex patterns of fraud due to the broad variety of types of features, which can include both transactional and contextual, as well as behavioral ones. The average transaction value of 4, 273 and a median of 1,850 indicate a right-skewed distribution with many small transactions and fewer high-value purchases that raise the mean transaction value significantly. The rather small percentage of transactions on the weekend (3.2) also offers a sort of discriminative characteristics since fraud researchers have shown that transactions made without authorization may be disproportionate during times of less vigilance (Perols, 2011).

3.2 Supervised Learning Techniques

Our instructed evaluation of learning has five algorithms that characterize various theoretical bases and methods of calculation. Logistic Regression is a baseline interpretable model, which estimates the probability of a transaction being anomalous using a logistic function applied to a linear combination of features. Although rather simple, logistic regression can actually work better in practice and give the coefficients that are more interpretable and useful in understanding auditing requirements (Hosmer *et al*., 2013). To avoid overfitting, we used L2 regularization with regularization strength 0.01, which was chosen by five-fold cross-validation on the training set.

Table 1: Dataset characteristics and feature categories with descriptive statistics

Category	Features	Statistics
Transaction Attributes	Amount, Date, Type, Category	Mean: $4,273; Median: $1,850
Vendor Characteristics	ID, Risk Score, Location, History	8,432 unique vendors
Employee Information	Department, Seniority, Authorization	23 departments, 4 levels
Approval Workflow	Approvers, Queue Time, Overrides	Mean approvers: 2.3
Temporal Features	Hour, Day-of-Week, Month, Quarter	Weekend txns: 3.2%
Derived Features	Historical Deviation, Peer Ratio	Z-score range: -3.8 to 12.4
Total Records	**247,683**	**Anomalies: 3,465 (1.4%)**

Random Forest, which is a collection of decision trees that are trained using bootstrap samples consisting of random subsets of features, is a solution to the logistic regression problem that cannot determine non-linear patterns and interaction between features (Breiman, 2001). The Random Forest model was configured with 200 trees, a maximum depth of 15 and a minimum sample per leaf of 20 in order to trade model complexity and the risk of overfitting. The minimum sample constraint does not allow trees to form leaves that classify specific anomalous training examples to promote generalization. Random Forests also give scores of feature importance in terms of mean decrease in impurity, which is more interpretable than most complicated models. Support Vector Machines (SVM) with a radial basis function (RBF) kernel map data into a high-dimensional feature space, where linear separation is possible, and it can be used effectively in data sets where classes cannot be separated in feature space (Cortes and Vapnik, 1995). The radial basis function kernel also allows the SVMs to learn non-linear and complicated decision boundaries. We used grid search on regularization parameter C, which was through regularization parameter C = 0.1,1,10,100 and the kernel coefficient γ = 0.001,0.01,0.1,1 and with the best-performing combination selected based on cross-validation results. SVM computational complexity, which is roughly quadratic with sample size, requires training on a stratified subsample of 50,000 transactions as opposed to the entire dataset.

XGBoost is the latest advancement in gradient boosting, as it applies optimized distributed gradient boosting that has automatic support to missing values and regularization, which is used to avoid overfitting (Chen and Guestrin, 2016). In contrast to the independent trees of Random Forest, trees of the XGBoost are constructed one after another, with each tree rectifying the mistakes of the other trees. This boosting method sometimes has high predictive performance compared to bagging approaches such as Random Forest. Our XGBoost model used 300 estimators with learning rate of 0.1, maximum depth of 6 and minimum child weight of 3. The contribution of each tree is explored by the learning rate, where small values will demand more trees, but in generalization will tend to be more effective. Early stopping with a patience of 20 rounds was applied using a validation set to achieve early stopping and avoid overfitting.

Table 2: Supervised Learning Algorithms and Hyperparameter Optimization Strategy

Algorithm	**Key Hyperparameters**	**Hyperparameter Selection Method**
Logistic Regression	L2 regularization strength (λ = 0.01)	5-fold Cross-Validation
Random Forest	Number of trees = 200; Maximum depth = 15; Minimum samples per leaf = 20	Grid Search with Cross-Validation
Support Vector Machine (RBF Kernel)	Regularization parameter (C = 10); Kernel coefficient (γ = 0.01)	Grid Search with Cross-Validation
XGBoost	Number of estimators = 300; Learning rate = 0.1; Maximum depth = 6	Bayesian Optimization
Artificial Neural Network	Hidden layers = [128, 64, 32]; Dropout rate = 0.3; Learning rate = 0.001	Manual Tuning with Early Stopping

The neural networks provide high flexibility in training arbitrary non-linear relations by stacking straightforward transformations within numerous layers (Goodfellow *et al.*, 2016). We had 3 hidden layers of 128, 64, and 32 neurons respectively with ReLU activation functions and a dropout rate of 0.3) to ensure that we do not overfit the data. The network used Adam optimizer and an initial learning rate of 0.001 and binary cross-entropy loss function suitable in binary classification. The training was continued up to a maximum of 100 epochs with early stopping on the validation loss, which normally converged after 30-40 epochs. Batch normalization was applied after each hidden layer to stabilize the training and increase the rate of convergence. The hyperparameter settings, as shown in Table 2, are the results of systematic search processes, and not the default. The algorithms have different selection methods depending on the complexity of the algorithm and the cost of computation with simpler methods such as logistic regression allowing exhaustive grid search and more proximate methods such as neural networks being required to depend on more practical manual grid tuning based on validation results. The heterogeneity in hyperparameter optimization reflects practical constraints in computational budgets, pure optimization is not always feasible, a factor that is frequently ignored in the academic literature, but is important to practitioners..

3.3 Unsupervised Learning Techniques

K-means clustering divides transactions into k clusters through a continuous process of allocating the points to closest centroid and changing centroid as the cluster mean (MacQueen, 1967). In order to identify anomalies, we consider transactions in small or distant clusters as possible anomalies. We tested k ∈{5,10,15,20} clusters, and selected k = 15 in terms of silhouette scores and business interpretability. Anomaly scoring gives every transaction a score relative to its position to the cluster centroid, divided by within-cluster variance, with anomalies being those that score greater than the 95th percentile. Although the K-means algorithm assumes the existence of a set of spherical clusters, both of similar size, which in real-world data is frequently inaccurate, its computational speed and explanatory aspects make it still popular in practice.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifies clusters in dense regions while labeling points in sparse regions as anomalies (Ester *et al.*, 1996). DBSCAN, in contrast to

K-means, does not have to be informed in advance of the number of clusters, but can find clusters of any shape. The algorithm is based on two parameters: ϵ (neighborhood radius) and MinPts (minimum points to make dense region). The ϵ is determined using k-distance graphs which measure the distance to the 5th nearest neighbor in all of the data, and we choose $\epsilon = 0.8$ at which the curve has an elbow, indicating what would be considered the natural density threshold. MinPts was configured to 10, which is in accordance with the dimensionality plus one heuristic with rather low-dimensional data. DBSCAN was declared to classify about 8.3% of transactions as noise (they could be anomalies), which is considerably bigger than the ground truth, but is reasonable to produce investigative leads.

Isolation Forest is based on the fact that anomalies are rare and different, and more vulnerable to isolation due to random partitioning (Liu *et al.*, 2012). The algorithm builds an ensemble of isolation trees: each of them is constructed by randomly picking features and split values with an anomaly score of average length of path to isolate every point. Anomalies need fewer splits to isolate than normal points and average path lengths are shorter. We set Isolation Forest with 200 estimators and contamination parameter 0.02, which is a little larger than the anomaly rate of the data to consider the possibility of unlabeled anomalies. Isolation Forest proves to be especially good when data has a large dimension and scales directly with the size of the dataset so that it can be used to provide large scale auditing (Liu *et al.*, 2012).

Local Outlier Factor is used to measure the local deviation of density of a particular point against its neighbours and determine areas of local similar density and those that are significantly lower density than their neighbours as anomalies (Breunig *et al.*, 2000). The method uses reachability distance, which includes the distance as well as the local density, in describing local neighborhood structure. The LOF of points in dense regions is approximately 1 whereas outliers are characterized by values of LOF that are significantly higher than 1. We choose the number of neighbors parameter to 20, which is a compromise between stability and local sensitivity. LOF is not very scalable, with computational complexity approximately $O(n^2)$ making it only useful in datasets which include hundreds of thousands of transactions. We therefore used the LOF on a 50,000-stratified sample of transactions.

Neural networks used to encode the input data and decode it via a slim bottleneck layer are called autoencoders, which identify anomalies as data with a high reconstruction loss (Goodfellow *et al.*, 2016). The underlying assumption is that the network learns in the process of training on mostly normal data to compress and reconstruct regular patterns, whereas anomalies (which are not adequately represented) do not reconstruct well. Our autoencoder had an encoder with layers [28, 20, 12, 6] neurons and an equal decoder [6, 12, 20, 28] with a latent dimension of 6, which compelled the information to be compacted. Mean squared error loss, Adam optimizer with a learning rate of 0.001, and batch size of 256 were used to train the model using 50 epochs. Anomaly scores were calculated as reconstruction error (mean squared difference between input and output), where the threshold was the 95th percentile of training set reconstruction errors. Autoencoders are particularly promising when it comes to the ability to capture complicated nonlinear patterns with reasonable computational efficiency.

3.4 Model Evaluation Metrics

Evaluating anomaly detection models requires metrics suitable for highly imbalanced class distributions where accuracy is misleading. One such naive model that treats all the transactions as legitimate has 98.6% accuracy on our data but identifies zero anomalies.

As a result, accuracy, recall and F1-score are the main metrics that are used. Precision is the ratio of fraudulent transactions to flagged transactions, which is inversely related to the false positive rate, which determines the amount of work done by auditors. Recall (sensitivity) quantifies the percentage of the true anomalies that the model has detected, and this measures the capacity of the model to detect fraud. F1-score is a compromise between precision and recall based on the harmonic mean of both terms, and thus avoids considering either of the two concerns independently. In the case of supervised models, we also compute Area Under the Receiver Operating Characteristic Curve (AUC-ROC), a measure of performance over decision thresholds, which is relatively robust to class imbalance (Fawcett, 2006).

In unsupervised approaches, which do not explicitly provide binary classifications, it is more subtle to do an evaluation. We calculate precision under varying recall levels, assessing the anomaly scores as rankings and changing the threshold. Silhouette coefficients determine the quality of clustering, they examine the similarity of each point of the cluster to itself in contrast to the other clusters, and the coefficients have a range of -1 to 1, with high values denoting well-defined clusters. Nonetheless, we are aware that good clustering does not necessarily imply that a good anomaly detector will be detected, since the most interesting anomalies are likely to be bad clusters, specifically because they are anomalies. As a result, precision-recall metrics are used as our main tool even when dealing with unsupervised approaches, with the ground truth of expert-labelled test sets.

Each model was evaluated using a stratified train–test split (70 percent training, 30 percent testing) and test set is not pruned to preserve the natural 1.4 percent anomaly rate. In the case of supervised methods, training data was further split into training (80) and validation (20) sets to use in the hyperparameter tuning and early stopping. Unsupervised methods were trained using the unlabeled training set, and the anomalous labels are not provided simulating real-life use, in which the ground truth is not known. Results of a test set are reported as performance measures, which give unbiased estimates of performance in the presence of new data. Each experiment was repeated five times using random seeds in order to measure stability using five random seeds and reporting the mean performance and standard deviation.

4.0 Results and Discussion

4.1 Performance of Supervised Learning Models

Table 3 presents detailed performance statistics of managed learning algorithms on the test subsample of 74,305 transactions with 1,040 labeled anomalies. The findings indicate that there is a significant performance disparity among algorithms, with gradient boosting and neural network algorithms significantly outperforming the simplistic techniques. XGBoost was the best with the highest F1-Score of 0.887 (±0.012), as it was capable of detection with a precision of 0.862 and a recall of 0.914. This performance can be converted to real world terms as: of all transactions that XGBoost identifies as anomalies, approximately 86% are actually what the model finds as genuine anomalous, and the model correctly finds 91% of the total anomalies that exist in the data. The false positive rate (0.8%) is operationally manageable and acceptable (0.8) which implies that the auditors would have to audit about 1,900 flagged transactions to know about 950 anomalies that are actually real.

Random Forest had a good score of 0.832, which is fair in comparison with logistic regression (0.738) but not good enough compared to XGBoost and neural networks. The difference in the performances of the Random Forest and XGBoost as tree-based ensemble techniques illustrates the benefit of sequential error correction of boosting over independent tree bagging. Notably, the neural network has a similar performance to that of XGBoost (F1-score 0.869) even though the neural network took significantly longer to train, with a lot of hyperparameter optimization and computational resources. This almost-equivalent behavior indicates diminishing returns of the architectural complexity on this specific task as neural networks may show to be beneficial with larger data sets or with fraud patterns that are more complicated.

Table 3: Test set performance of supervised learning models (mean and standard deviation of 5 runs)

Algorithm	Precision	Recall	F1-Score	AUC-ROC	FPR (%)
Logistic Regression	0.721 ± 0.018	0.756 ± 0.022	0.738 ± 0.015	0.863 ± 0.009	1.9
Random Forest	0.824 ± 0.014	0.841 ± 0.019	0.832 ± 0.011	0.918 ± 0.007	1.2
SVM (RBF)	0.798 ± 0.021	0.823 ± 0.024	0.810 ± 0.017	0.902 ± 0.011	1.4
XGBoost	0.862 ± 0.009	0.914 ± 0.015	0.887 ± 0.012	0.941 ± 0.006	0.8
Neural Network	0.847 ± 0.016	0.892 ± 0.018	0.869 ± 0.014	0.929 ± 0.008	1.0

Even under the computational constraints that restricted the training to a subsample, Support Vector Machines showed respectable results (F1-score 0.810). Nonetheless, SVM training time scales quadratically, which makes this method unusable when working with a dataset that contains a few hundred thousand transactions (which is usual in the case of enterprise audits). Although logistic regression had the lowest performance metrics, it maintained a decent performance with F1-score of 0.738. The benefits of logistic regression in interpretability, such as clear coefficients to show the contribution of features to predictions, might justify its use in regulatory components where model explainability is a major consideration.

Fig. 2 illustrates model performance using ROC curves that represent true positive rates versus false positive rates at different decision threshold rates. The curves demonstrate the capacity of each model to discard anomalies and valid transactions regardless of the choice of a particular threshold. The curve of XGBoost is the fastest to climb towards the top-left corner with high recall (0.914) even when the thresholds are set to high values with few false positives. Random Forest and neural networks exhibit similar curves, whereas the trade-offs of logistic regression and SVM exhibit gradual curves. Its practical implication is evident: the organizations can use a balance between detection by varying decision thresholds.

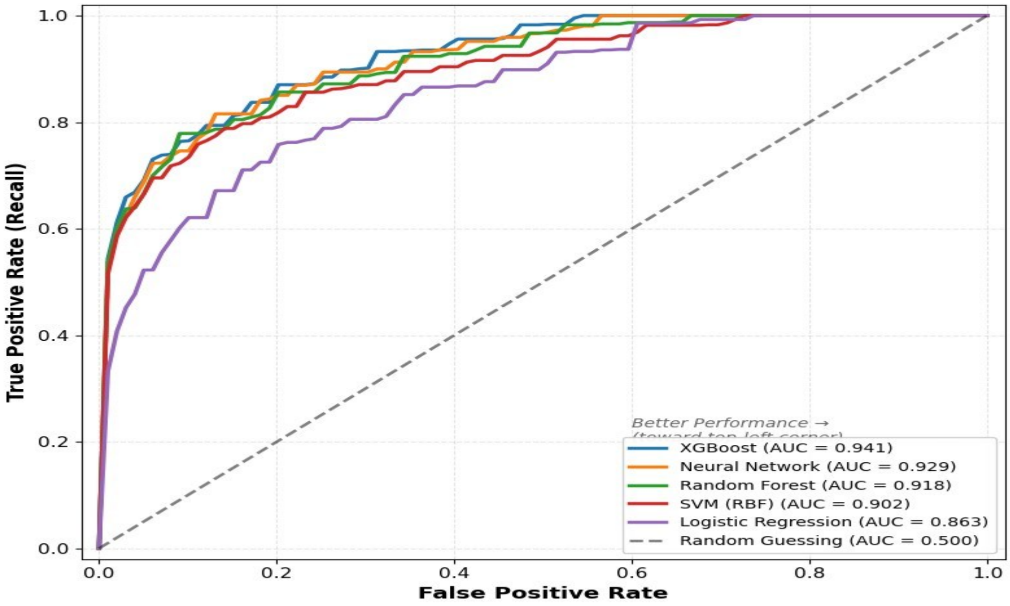

Fig. 2: The curves of Receiver Operating Characteristic (ROC) of supervised learning algorithms.

The curves (Fig. 2) are graphs of true positive rate (recall) vs. false positive rate versus decision threshold. XGBoost (AUC = 0.941) and Neural Network (AUC = 0.929) have a better discrimination ability since, their curves stick to the top-left hand side suggesting they are having high true positive rates with minimum false positives. The dotted line is a random guessing (AUC = 0.50).

Sensitivity to investigative capability. Conservative thresholds (left side of curves) indicate fewer transactions with high probability that they are anomalous whereas aggressive thresholds (right) nets wider but at the cost of increased false positives.

The importance analysis of the random forest and XGBoost features show the transaction attributes that are the most significant predictors of anomalies. The value of transaction amount proved the most significant attribute, which is also in line with the specificity of fraud studies wherein it is found that the amount perpetrators seek will tend to be within a range of amounts that allow them to avoid detection limits (Kirkos *et al*., 2007). Breach of the historical patterns, derived feature that quantifies the extent to which each transaction is not following the pattern that the employee or vendor follows, was ranked the second most important, which confirms the rule that the changes in behavior are indicative of fraud. Workflow approval characteristics such as the number of approvers and overrides showed significant predictive ability. The number of approvers on the fraudulent transactions were lower systematically implying that they could bypass the control procedures. Moderate importance was found on temporal characteristics, with the risk of weekend and after-hours transactions being higher as it was supposed.

4.2 Unsupervised Learning Models Performance.

Unsupervised methods are more difficult to evaluate because they are trained without labels but must be assessed using labeled test data—an unavoidable practical trade-off when comparing methods. The conclusion of performance by using the results of Andrews and Murphey as a ranking with the top 1.4 percent (the rate of anomaly in the test set) as the predicted anomalies is summarized in table 4. Isolation Forest recorded the largest performance among unsupervised methods with F1-score 0.752 and detect rate 87.3, which approaches the performance of supervised methods performance even though the isolation forest is trained without labels. This performance indicates the theoretical strength of Isolation Forest: as it is recursively partitioned, isolated unusual points are natural, namely, the attributes that make anomalies.

Table 4: Comparison between the performance of the unsupervised learning algorithms at the decision threshold that gives 1.4 percent predicted anomaly rate.

Algorithm	Precision	Recall	F1-Score	Silhouette	FPR (%)
K-means	0.542 ± 0.031	0.634 ± 0.028	0.584 ± 0.024	0.387	2.8
DBSCAN	0.478 ± 0.037	0.589 ± 0.041	0.527 ± 0.033	0.412	3.4
Isolation Forest	0.681 ± 0.019	0.873 ± 0.024	0.752 ± 0.018	N/A	2.1
Local Outlier Factor	0.623 ± 0.026	0.791 ± 0.029	0.697 ± 0.022	N/A	2.5
Autoencoder	0.647 ± 0.023	0.812 ± 0.027	0.720 ± 0.021	N/A	2.3

With a respectable performance (F1-score 0.720), autoencoders proved that the reconstruction error method is valid in detecting anomalies. The neural architecture allows autoencoders to learn non-linear patterns that are complex that describe normal transactions, anomalies manifest as an indication of difficulty in reconstruction.

Local Outlier Factor had an F1-score of 0.697, but because of computational limitations, the evaluation was done on a subsample. Density-based method is successful when anomalies actually have lower local density compared to normal transactions, a fact that is usually true in our audit data but may not hold in the event that the fraud is concentrated in feature space.

K-means and DBSCAN had smaller results (F1-scores of 0.584 and 0.527 respectively), proving that the level of clustering does not necessarily translate to the quality of the anomaly detection. The assumption of the sphericity of the clusters in the K-means is probably a source of suboptimal performance because transaction data is more complex. The density-based algorithm of DBSCAN, which was theoretically attractive, had poor parameter sensitivity. The performance of the algorithm is highly dependent on the ϵ selection and optimum values might vary across different feature space regions- a shortcoming when employing single parameters across the entire feature space.

False positive rates of unsupervised techniques (2.1-3.4) are higher than those of the best supervised techniques (0.8-1.0), which is an expected effect of unsupervised training. These rates are however quite viable. With 2.1% false positive of Isolation Forest, the audit of 2,900 transactions reveals 907 veritable anomalies, a fair trade-off given that the approach does not require historical labels to learn and may discover new types of fraud never seen by the supervised models that follow existing patterns.

Fig. 3 shows precision-recall curves that are more informative because of highly imbalanced data (Davis and Goadrich, 2006). The curves demonstrate the decreasing precision of systems trying to capture an increasing number of anomalies (higher recall). Supervised approaches have a high accuracy at all levels of recall, and XGBoost can retain accuracy of over 0.80 at 90 percent. Unsupervised approaches exhibit sharp decline in precision at high recall, indicating rising false positives on the extension of detection threshold. Interestingly, the curve of Isolation Forest is close to the moderate recall (60-80%), implying that there is a sweet point, at which unsupervised detection is reasonable and not too high in terms of preciseness, and the slope of the curves is not as steep as at very high recall rates.

4.3 Comparative Analysis and practical implications.

A direct comparison of supervised and unsupervised paradigms helps to understand their comparative advantages and the situations when they should be used. Supervised algorithms prove to have obvious performance benefits in the presence of adequate high-quality labeled data, and XGBoost outperforms the best unsupervised method (Isolation Forest) by about 18 percent in F1-score. This gap is indicative of the underlying importance of learning based on explicitly labeled examples and deviation identification based upon statistical properties on their own. The priority should be achieved in organizations that have a long history of fraud with a variety of fraud types.

The supervised approaches (solid lines) always perform better than the unsupervised ones (dashed lines) but they need the training data to be labelled. Isolation Forest shows the best performance without supervision, close to the performance at moderate levels of recall using supervision. Random performance baseline is depicted by the dotted horizontal line.

Supervised methods, especially the ensemble methods such as XGBoost or Random Forest that are moderately in performance, interpretable, and computationally efficient.Nevertheless, there are a number of conditions that support unsupervised techniques. The most interesting benefit is novel fraud detection, as unsupervised models show low performance at identifying fraud patterns that are not present in the training data, whereas supervised models will identify statistical deviations irrespective of previous history. Audit practitioners, interviewed as practitioners have highlighted that their biggest cases of fraudulent actions have involved schemes that have never been seen in history, implying disastrous failures of supervised only solutions. Another advantage that is unsupervised is the uncertainty attached to the quality of labels in most of the organizations. The process of fraud investigation is cost-based and usually inconclusive as organizations do not know whether there has been undiscovered fraud in the past using history transactions as normal. This concern is bypassed using unsupervised approaches that do not rely on possibly corrupted labels.

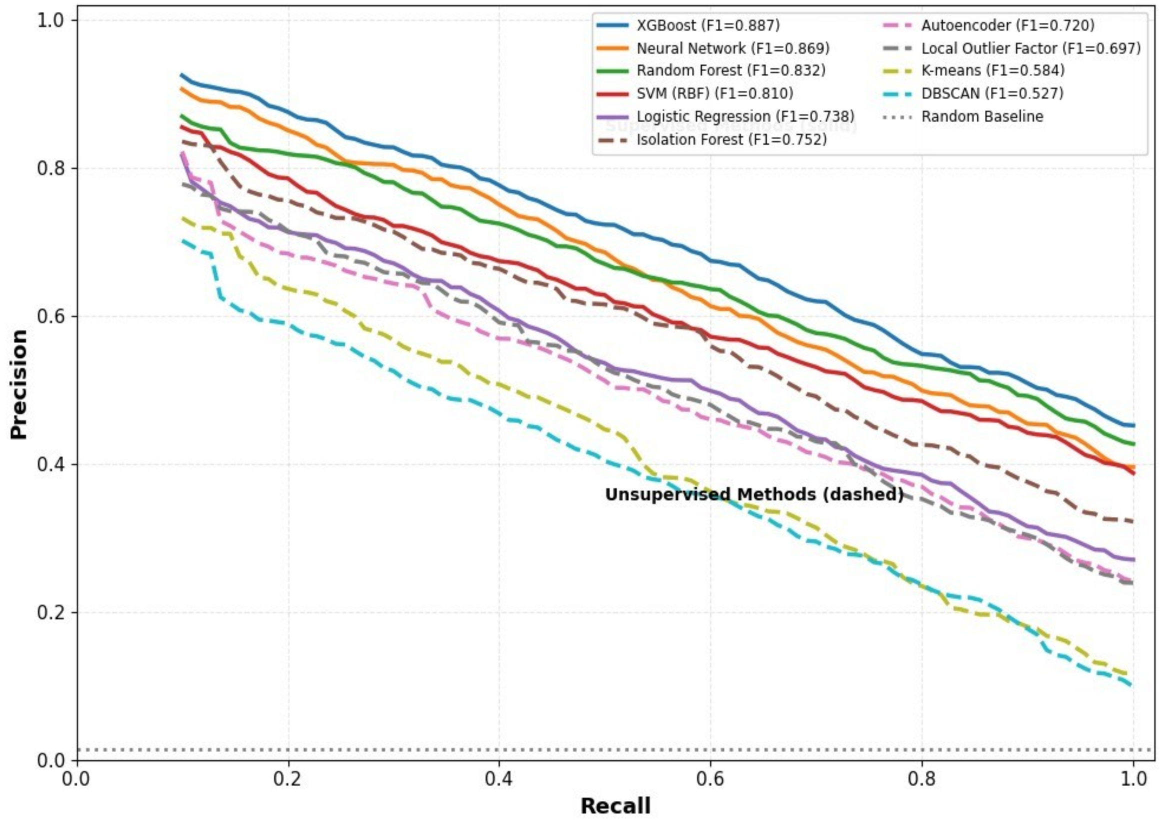

Fig. 3: Supervised and unsupervised Precision-Recall curves illustrating a tradeoff between precision (fraction of flagged transactions that are real anomalies) and recall (fraction of anomalies detected) at the various thresholds of the decision rule

Subtly, there is a trade-off in computational efficiency issues. Isolation Forest is particularly efficient in training time is Isolation Forest which can train in a few seconds on our dataset whereas the neural networks can take minutes to hours based on the architecture and hardware. Prediction time, or the time to score a new transaction, is, however, more important with regard to deployment of production, where models can analyze thousands of transactions per day. In this case, tree-based algorithms (Random Forest, XGBoost, Isolation Forest) prove to be beneficial, and transactions are made in milliseconds. Neural networks and SVMs take longer to do their prediction, although that is still within manageable limits in regard to batch processing. In the case of real-time transaction screening, tree-based algorithm seems to be obvious winners.

The number of transactions, non-conformance to past trends, and approval processes become the leading predictors in both approaches. Temporal features are moderately important whereas location of vendors and department are less likely to predict. Top features agreement is a guarantee that these features are truly predictive and not a product of the model.

Fig 4 shows the importance rankings of features of our two best supervised methods, including the attributes in the transaction that are most predictive of an anomaly. The significant similarity of the XGBoost and

Random Forest to the overall best features of transaction amount, historical deviation, approval count, and transaction velocity gives the assurance that they are actually indicative of the real fraud rather than a model-based phenomenon.

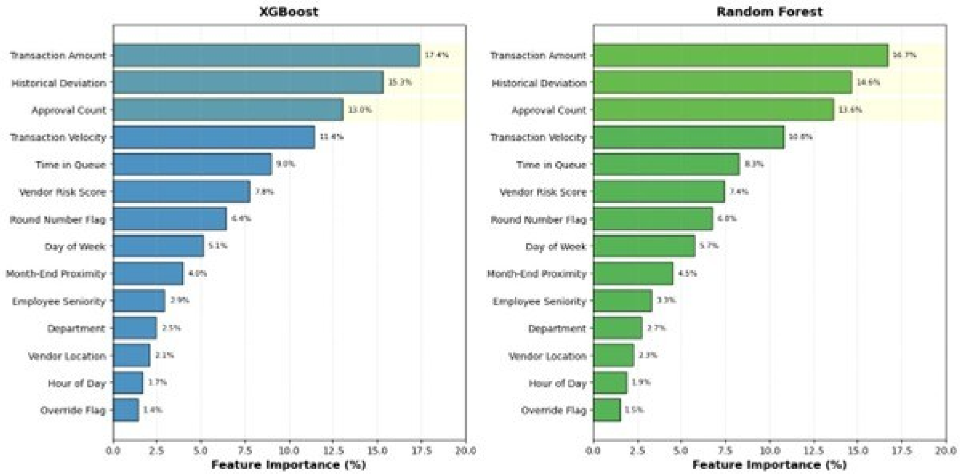

Fig. 4: Feature importance scores of XGBoost (left) and the Random Forest (right) supervised models.

This intersection provides practical audit practice advice: more vigilant and stringent oversight and controls over these high-ranking characteristics would help in averting fraud before it occurs, and not just pick it up after the fact. Examples of how models identify top risk factors could include the introduction of dynamically set approval thresholds, the use of which take into consideration past patterns, and the use of more approvers on transactions that have significant historical deviations.

Implementation of machine learning in audit anomaly detection is not limited to choosing the algorithm, but also organization and process factors. Model interpretability is essential to adoption; auditors need to know why transactions have been flagged to effectively investigate and to meet regulatory audits of documentation of the audit trail. Classical techniques that use trees have a natural interpretability in terms of decision paths, whereas neural networks use post-hoc explanation methods such as SHAP (SHapley Additive exPlanations) values, which more or less approximate feature contributions (Lundberg and Lee, 2017). Organizations ought to consider the tradeoffs between performance and interpretability requirements between complex models and more interpretable models, perhaps using high-performance models to do preliminary screening and interpretable models to do end adjudication.

False positive management is also very important because too many false alarms cause auditors to be overburdened and damage the system credibility. Practitioners during our interviews indicated that the number of transactions that audit teams usually follow up on 50-200 transactions per week with an organization depending on the size of the organization. Systems with thousands of false positives daily do not work in practice. It, therefore, follows that organizations need to focus on precision at a cost, so that they recall well and operate within higher confidence of catching 80 percent of fraud and not 95 percent due to noise that cannot be investigated effectively. The threshold choice is a business trade-off between detection sensitivity in business and investigative capacity.

Integrating with the current audit processes is also another viable challenge that has been ignored in the technical literature. Machine learning systems cannot be black boxes and they need to integrate with case management systems, offer audit trail documentation and cooperate in investigations. Effective implementations that we have seen used hybrid strategies: algorithms produce preliminary prioritization scores, senior auditors check on the highest-scoring transactions to validate that they merit investigation, and the results of investigations are fed back to improve the models by adding more training data or features.

4.4 Limitations and Future directions

Our results are marred by a number of shortcomings. Although the dataset is large, it is specific to particular industries and geographical areas, which could have restricted the ability to generalize to other settings like the banking industry, healthcare, or government, where the nature of transactions and fraud trends are likely to vary significantly. There will always be noise in the labels, some of which will be anomalies labeled by the multi-source, and unlabeled data will probably have fraudulent transactions not detected. This uncertainty of labeling has more impact on supervised learning than on unsupervised methods but has an impact on the performance evaluation of these two paradigms. We also used professional inspection to reduce the number of label errors, though there is undoubtedly a certain level of contamination.

This timelessness of our assessment learning models on past data and testing on unseen parts held out during the same time period does not reflect the temporal dynamics of the patterns in fraud that breed production systems. The fraudsters evolve strategies as time goes by and this may make the models that were used in history irrelevant. Currently, there exist studies into continual learning methods that update models with new data, but such methods have the risk of causing catastrophic forgetting such that the models become unable to identify historical patterns and instead constantly adapt to new ones (Kirkpatrick *et al*., 2017). Our advice to organizations is to have models retrained at least every quarter, with higher rates of retraining in case performance monitoring shows a loss of performance.

Although the class imbalance is solved by SMOTE oversampling, it might not entirely solve the problems of supervised learning. Miscellaneous strategies such as cost sensitive learning which punishes the misclassification of a minority group more severely, or deviations of supervised algorithms, should be studied in the future. The combination of supervised and unsupervised approaches that are being supervised is an especially encouraging direction, a combination of paradigms with supervised and unsupervised approaches, where the unsupervised approaches are used to indicate candidate anomalies that are then subsequently trained on a larger labeled data set.

Future directions in research involve temporal sequence modeling with recurrent neural networks or transformers to learn the trends of fraud (not per transaction) when played out in a sequence of transactions (Hilal *et al*., 2022). Graph neural networks are promising to identify fraud networks in cases when several employees or vendors conspire and use patterns of relationships that are not visible in transaction-level features. Explainable AI methods need to be advanced to an explanation of the complex model predictions that are readable, so that the audit can be trusted and justified to the regulations. Lastly, adversarial robustness creating models that are hard to intentionally manipulate by fraudsters who know about detection systems is also a very important but poorly studied challenge.

5.0 Conclusion

This study provides a comprehensive comparison of supervised and unsupervised machine learning approaches for audit anomaly detection, offering important insights for both researchers and practitioners. The findings demonstrate that supervised learning methods, particularly gradient boosting models such as XGBoost, achieve superior performance when high-quality labeled training data are available, with F1-scores approaching 0.90 and low false positive rates. These results highlight the strength of supervised techniques in identifying known fraud patterns with high precision.

At the same time, unsupervised approaches—including Isolation Forest and autoencoders—prove to be effective alternatives in situations where labeled data are scarce or unavailable. Although their performance is comparatively lower, their ability to detect previously unknown or evolving fraud patterns makes them especially valuable in dynamic and uncertain audit environments. This reinforces the importance of unsupervised learning as a complementary tool rather than a replacement for supervised methods.

Beyond predictive accuracy, the study underscores that practical deployment considerations—such as computational efficiency, interpretability, and control of false positives—are equally critical in real-world audit applications. Algorithm selection should therefore be guided not only by benchmark performance metrics but also by organizational context, regulatory requirements, and the capacity of audit teams to investigate flagged anomalies.

The results strongly support the adoption of hybrid audit analytics frameworks that integrate both paradigms. In such systems, supervised models can provide high-confidence detection of established fraud schemes, while unsupervised models continuously monitor for emerging and previously unrecognized anomalies. Human auditors remain central to this process, providing expert judgment, validating model outputs, and contributing feedback that improves system performance over time.

Future research should focus on addressing temporal model degradation, advancing ensemble and hybrid modeling strategies, and improving model interpretability to strengthen trust and transparency in audit analytics. These efforts will be essential for enabling organizations to fully leverage machine learning technologies in building more proactive, adaptive, and effective audit systems.

6.0 References

Aboagye, E. F., Borketey, B., Danquah, K., & Borketey, D. (2022). A Predictive Modeling Approach for Optimal Prediction of the Probability of Credit Card Default. *International Research Journal of Modernization in Engineering Technology and Science*. 4, 8, pp. 2425-2441

Akinsanya, M. O., Adeusi, O. C., & Ajanaku, K. B. (2022). A Detailed Review of Contemporary Cyber/Network Security Approaches and Emerging Challenges. *Communication in Physical Sciences*. 8, 4, pp. 707-720

Association of Certified Fraud Examiners (2020). *Report to the Nations: 2020 Global Study on Occupational Fraud and Abuse*. Austin, TX: ACFE.

Alles, M. G., Kogan, A., & Vasarhelyi, M. A. (2006). Feasibility and economics of continuous assurance. *Auditing: A Journal of Practice & Theory*, 21, 1, pp. 125-138. https://doi.org/10.2308/aud.2002.21.1.125

Arens, A. A., Elder, R. J., & Beasley, M. S. (2016). *Auditing and Assurance Services: An Integrated Approach*. 16th Edition. Pearson.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM SIGMOD Record*, 29, 2, pp. 93-104. https://doi.org/10.1145/335191.335388

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey.*ACM Computing Surveys*, 41, 3, pp. 1-58. https://doi.org/10.1145/1541880.1541882

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, pp. 321-357. https://doi.org/10.1613/jair.953

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

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. https://doi.org/10.1007/BF00994018.

Cushing, B. E., & Loebbecke, J. K. (1986). Comparison of audit methodologies of large accounting firms. *American Accounting Association Studies in Accounting Research No. 26*. Sarasota, FL: AAA.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233240. https://doi.org/10.1145/1143844.1143874

Debreceny, R. S., & Gray, G. L. (2010). Data mining journal entries for fraud detection: An exploratory study. *International Journal of Accounting Information Systems*, 11, 3, pp. 157-181. https://doi.org/10.1016/j.accinf.2010.08.001

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226-231.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 8, pp. 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE*, 11, 4, e0152173. https://doi.org/10.1371/journal.pone.0152173

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. Cambridge, MA.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning:Data Mining, Inference, and Prediction*. 2nd Edition. Springer. https://doi.org/10.1007/978-0-387-84858-7

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications*, 193, 116429. https://doi.org/10.1016/j.eswa.2021.116429

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. 3rd Edition. Wiley. https://doi.org/10.1002/9781118548387

Jans, M., Lybaert, N., & Vanhoof, K. (2010). Internal fraud risk reduction: Results of a data mining case study. *International Journal of Accounting Information Systems*, 11, 1, pp. 17-41. https://doi.org/10.1016/j.accinf.2009.12.004

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32, 4, pp. 995-1003. https://doi.org/10.1016/j.eswa.2006.02.016

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114, 13, pp. 3521-3526. https://doi.org/10.1073/pnas.1611835114

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, 6, 1, pp. 1-39. https://doi.org/10.1145/2133360.2133363

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural*

Information Processing Systems, 30, pp. 4765-4774.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 14, pp. 281-297.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50, 3, pp. 559-569. https://doi.org/10.1016/j.dss.2010.08.006

Nigrini, M. J. (1996). A taxpayer compliance application of Benford's Law. *The Journal of the American Taxation Association*, 18, 1, pp. 72-91.

Pang, G., Shen, C., Cao, L., & Van Den Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54, 2, pp. 1-38. https://doi.org/10.1145/3439950

Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagao, T. (2016). Deep learning anomaly detection as support fraud investigation in Brazilian exports and antimoney laundering. *15th IEEE International Conference on Machine Learning and Applications*, pp.954-960. https://doi.org/10.1109/ICMLA.2016.0172

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, 30(2), 19-50.https://doi.org/10.2308/ajpt-50009

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review*, pp. 1-14. https://doi.org/10.1007/s10462-010-9187-8

Rezaee, Z., & Riley, R. (2010). *Financial Statement Fraud: Prevention and Detection*. 2nd Edition. Wiley. https://doi.org/10.1002/9781119203742

Singleton, T. W., & Singleton, A. J. (2010). *Fraud Auditing and Forensic Accounting*. 4th Edition. Wiley.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57, pp. 47-66. https://doi.org/10.1016/j.cose.2015.09.005

Consent for publication

Not Applicable

Availability of data and materials

The publisher has the right to make the data public

Competing interest

Authors declared no conflict of interest.

This work was sole collaboration among all the authors

Funding

There is no source of external funding

Authors Contribution

All components of the work were carried out by the author