TabNet Interpretable Deep Structure for Adaptive Recognition of Audit Voucher Anomalies

Lixuan Sun
School of Economics and Management, Fuzhou University, Fuzhou 350108, Fujian, China
Corresponding author: Lixuan Sun (e-mail: sunlixuan2025@163.com)

ABSTRACT This study proposes an interpretable deep learning framework for adaptive recognition of audit voucher anomalies based on the TabNet architecture. The framework addresses the dual challenges of decision transparency and environmental adaptability in intelligent auditing systems. Structured audit voucher data are processed through sequential attention layers to generate traceable feature importance masks and contribution paths. A sliding timewindow mechanism is introduced to accumulate historical feature importance distributions, while a dynamic threshold calibration strategy based on statistical characteristics enables adaptive anomaly boundary adjustment under concept drift conditions. The proposed framework integrates feature selection, temporal distribution analysis, and anomaly scoring to achieve transparent and adaptive anomaly recognition. Experimental results demonstrate that the detection accuracy reaches 92.3% for amount logic conflicts and 93.1% for supplier-related anomalies, while all anomaly categories achieve F1-scores above 87%. In addition, the verification time for supplier anomalies is reduced from 19.6 min to 8.4 min, resulting in a 57.1% improvement in audit verification efficiency. The average importance of transaction amount features reaches 0.883 in financial audit scenarios, confirming the model's ability to adaptively focus on critical business characteristics. The proposed framework can be integrated with wireless data acquisition infrastructures, intelligent financial information systems, and edge-computing-enabled audit platforms to support realtime anomaly monitoring and interpretable decision-making in large-scale financial auditing environments.

INDEX TERMS audit voucher anomaly detection, TabNet, interpretable deep learning, intelligent auditing

I. INTRODUCTION

AS the cornerstone of economic activity supervision, financial auditing focuses on verifying the authenticity and compliance of financial reports. In sectors like manufacturing, this rigorous verification is critical to accurately assess fluctuating inventory valuations and substantial capital investments in specialized production machinery. As legal evidence of transaction records, the identification of anomalies in audit vouchers is directly related to the reliability of financial data [1-3]. The traditional manual sampling method is difficult to cope with massive voucher data due to its low efficiency [4], while the introduction of deep learning technology has improved the level of automation, the lack of traceability in the decision-making process due to the "black box" characteristics of the model seriously restricts the acceptance of auditors [5,6]. The adequacy and appropriateness of audit evidence require that the decision-making basis must be transparent and verifiable, which makes the development of intelligent recognition systems that combine high precision and interpretability an urgent need in the industry [7,8]. In recent years, the lack of interpretability of deep learning models in structured data scenarios has become a key bottleneck hindering their implementation in the auditing field [9,10], such as the frequent occurrence of compliance risk cases caused by model misjudgments in financial industry applications. Thus, in addition to increasing audit efficiency, investigating anomaly recognition techniques that can adjust to changes in the environment and offer clear decision-making pathways can also bolster the reliability of financial data and offer technical assistance for the stability of the capital market.

The process of identifying abnormal audit vouchers is complicated by a number of technical issues, such as the inherent opaqueness of deep learning models, which makes it impossible for auditors to ascertain the reasons behind a model's determination that a particular voucher is "abnormal" [11,12]. The integrity of the audit evidence chain is compromised when an auditor is unable to establish the basis for identifying abnormal vouchers. Temporal dynamics in audit data have a significant impact on the development

of business models as well as the emergence and updating of fraudulent techniques. This phenomenon is known as "concept drift" and can quickly make any static model obsolete [13,14]. Additionally, the structure of the voucher data is highly multidimensional and heterogeneous in nature; for example, it is comprised of mixed feature values, categorical values, and time-series values. Creating meaningful associations among these attributes using typical feature engineering methods is difficult [15,16]. There is considerable uncertainty regarding how an "anomaly" should be defined in different domains, and no universally accepted quantitative standard exists for defining an "anomaly". This combination results in significant inconsistencies in the quality of the labels created for use in model training. Moreover, the consistent mismatch between model outputs, particularly when compared to expert opinion, presents a significant barrier to the advancement and adoption of audit intelligence. Lastly, the presence of data distribution bias in financial time-series datasets, particularly with respect to the generalizability of a model built on those data, is likely to substantially impact the overall accuracy of the findings generated by the model.

The problem of explainable artificial intelligence is being explored in the research community by developing techniques that provide insight into the algorithmic decisions made by AI systems. SHAP (Shapley Additive Explanations) provides insights into contributions made by features but increases exponentially as the number of features increase, therefore failing to accommodate real-time audit requirements [17,18]. Additionally, LIME (Local Interpretable Model-Agnostic Explanations) creates local explanations that are not globally consistent and do not comprehensively represent the decision logic of the model [19,20]. Some research have proposed dynamic data processing approaches based on sliding window methodologies; however, most of these methodologies apply only to traditional time-series models, provide limited accuracy, and lack integrated interpretability modules. While models developed by ensemble learning frameworks enhance robustness, they further obfuscate decision pathways as decision pathways are created from the interaction of multiple models [21,22]. Collectively, these methods do not effectively provide a balance between the predictive capability of deep learning models and the transparency needed for auditing decisions, specifically the inability for the interpretation module to adjust automatically to changes in data distribution, resulting in limited utility in real-world auditing settings.

To resolve the challenges of non-transparent audit decision-making and the inability of deep-learning-based anomaly recognition to properly adjust to changing environments, this study proposes a novel framework. This dynamic calibration mechanism is specifically designed to adjust the valuation of features, allowing the model to adaptively respond to shifting financial landscapes. The framework is built upon the TabNet model, which processes credential information, employs a sequential attention mechanism to produce traceable feature contribution sequences, and offers a combinatorial timeseries statistical approach to identifying anomalies while permitting real-time adjustments to the boundaries of the anomaly detection process based on the data of the environment being analyzed. The dynamic calibration mechanism retains the explicit nature of the process by which features are selected; however, it allows models to dynamically optimize their ability to identify anomaly detection points based upon the provision of new samples from the data of the audits being completed, thus creating a methodology for the adaptation of audit practices that provides decision transparency as well as an adaptive ability to respond to shifts in the data of the audit environment.

II. ADAPTIVE INTERPRETABLE ANOMALY RECOGNITION FRAMEWORK BASED ON TABNET

A. TABNET FEATURE SELECTION MECHANISM FOR AUDIT VOUCHERS

Audit voucher data presents high-dimensional structured characteristics, including heterogeneous features such as numerical transaction amounts, timestamps and categorical business types, voucher sources, etc. The TabNet model uses sequential attention mechanism to parse such mixed feature structures, achieving explicitization of the feature selection process [23,24]. The input layer of the model receives the preprocessed voucher feature vector x∈Rd, where d represents the feature dimension, and the batch normalization layer eliminates dimensional differences. Category features are transformed into continuous vector representations through embedding layers and participate in subsequent calculations in a unified dimensional space with numerical features. Figure 1 shows the complete architecture process of feature selection for audit vouchers.

As shown in Figure 1, this architecture processes heterogeneous data through a dual branch path. Numerical features are batch-normalized, and categorical features are embedded and converted before being jointly input into a sequential attention iteration mechanism. The model coordinates the mask generator and feature converter in each iteration, with the former generating sparse feature masks and the latter performing nonlinear transformations. The outputs of the two are fused and updated with GLU (Gated Linear Unit) to update the hidden state. The iterative process continues until the preset maximum number of steps, ultimately accumulating feature importance and outputting a traceable decision path.

The sequential attention mechanism separates key features through N_{steps} iterative decision-making process, generating a sparse feature mask $M[t]$ for each iteration t.

$$M[t] = \text{Sparsemax}\left(W_m^T h[t] + b_m\right) \tag{1}$$

The Sparsemax function generates differentiable sparse output, where $W_m \in R^{h\times d}$ and $b_m \in R^d$ are learnable parameters of the mask generator, $h[t] \in R^h$ is the hidden state at step t, and h represents the dimension of the hidden

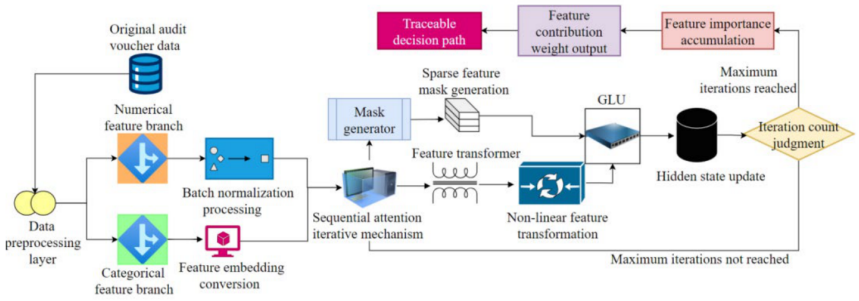

FIGURE 1. Audit Voucher Feature Selection Architecture

layer. This mask dynamically weights key audit features such as abnormal amount patterns or non working time transactions.

The feature transformation module performs nonlinear transformation on the selected features of the mask [25-26], and outputs a decision representation a[t]:

$$a[t] = \mathrm{GLU}(\mathrm{BN}([x \odot M[t], d[t-1]]W_a + b_a)) \quad (2)$$

$\odot$ represents element level multiplication, BN represents batch normalization operation, $[\cdot,\cdot]$ represents vector concatenation, $d[t-1]$ represents the decision representation of the previous step, $W_a \in R^{2d\times a}$ and $b_a \in R^a$ are transformation layer parameters, and a is the output dimension. The model iteratively accumulates feature importance weights to form a traceable decision path. The feature importance in step t is proportional to the mask value, and the final importance weight is the sum of contributions from all steps:

$$I_{\mathrm{final}} = \sum_{t=1}^{N_{\mathrm{steps}}} M[t] \odot \mid a[t] \mid \quad (3)$$

A parameter exists that adjusts the sparsity of the mask. This parameter provides a variable input based on the various complexities of an audit scenario, selecting the ratios of the features for each iteration. The feature importance sequence Ifinal feeds into the dynamic calibration module, enabling transparent decision-making. The model architecture preserves feature contributions, ensures audit evidence chain integrity, and maintains deep learning's capacity to capture high-dimensional feature interactions. The feature selection process assigns weights according to data distribution patterns, eliminating reliance on pre-determined guidelines for anomaly identification.

B. DYNAMIC ACCUMULATION AND DRIFT QUANTIFICATION OF FEATURE IMPORTANCE

Audit voucher data has significant time series characteristics, and the voucher generation process follows specific business cycle rules. A fixed sliding time window captures temporal dynamics in feature importance from TabNet output. The window size is determined based on the audit business cycle, reflecting the time span of a typical voucher processing cycle. Continuously accumulate the distribution of feature importance within the window and establish a statistical model of feature contribution in the time dimension.

The optimal window size w* is determined by analyzing the autocorrelation function of the feature importance sequence. The window size is set to the smallest time lag where the autocorrelation coefficient drops below 0.2, ensuring the window captures meaningful temporal patterns while maintaining sensitivity to concept drift. This approach is theoretically grounded in time series analysis, where the window size must balance the trade-off between capturing sufficient historical information and responding to changes in the data distribution. The optimal window size w* satisfies:

$w^* = \min\{k : |r(k)| < 0.2\}$, where $r(k)$ represents the autocorrelation coefficient at lag k for the feature importance sequence. This mathematical formulation provides a rigorous basis for determining the sliding window size, ensuring it aligns with the characteristic time scale of audit business cycles while maintaining responsiveness to concept drift events. The sliding time window mechanism divides the feature importance sequence into continuously overlapping subsequences, each containing w feature importance weights for consecutive time points [27-28]. $\mu_j(t)$ is calculated recursively:

$$\mu_j(t) = \mu_j(t-1) + \frac{I_j(t) - I_j(t-w)}{w} \quad (4)$$

$\mu_j(t)$ represents the average feature importance of the jth feature at time t, $I_j(t)$ represents the feature importance weight value of the jth feature at the current time t, $I_j(t-w)$ represents the feature importance weight value at the forefront of the window at time $t-w$, and w is the sliding window size. This recursive formula effectively reduces computational complexity and avoids recalculating the mean of the entire window every time.

The cumulative importance standard deviation $\sigma_j(t)$ of the jth feature in the sliding window at time t is also calculated recursively:

$$\sigma_j(t) = \left(\frac{w-1}{w} \times \sigma_j^2(t-1) + \frac{\left(I_j(t) - \mu_j(t-1)\right) \times \left(I_j(t) - \mu_j(t-w)\right)}{w} + \frac{-\left(I_j(t-w) - \mu_j(t-1-w)\right) \times \left(I_j(t-w) - \mu_j(t-w)\right)}{w} \right)^{1/2} \quad (5)$$

$\sigma_j(t)$ represents the standard deviation of the feature importance of the jth feature at time t, while $\mu_j(t)$ and $\mu_j(t-1)$ represent the mean feature importance at the current and previous time, respectively. This recursive formula maintains computational efficiency while accurately reflecting the degree of dispersion of feature importance distribution. Figure 2 presents the importance distribution of 10 audit voucher features within 30 time windows.

The supplier code features has increased from 0.57 to 0.83 across the 20-30 window from Figure 2. This reflects the business cycle characteristics at the end of the month. The fluctuating characteristics associated with transaction amounts for a company correspond with windows five, 15 and 25 and indicates company fund flow timing activities as they are related to transaction amounts. The characteristics associated with the voucher origins changed as a company's use of new business channels evolved from 0.23 to 0.42. Supplier codes were rated 0.12 higher than the previously documented level of importance and transaction amounts were documented with a 0.03 decline, thus indicating that conceptual drift occurred in windows 12 and 20. The temporal distribution shown above can also be utilised as a statistical base to enable the formation of a methodology that enables the continual adjustment of both the dynamic threshold calibrations and the anomaly definition limits according to the latest distributions of data; in short, allowing the auditors to fully capture dynamic changes in feature contributions related to an audit voucher anomaly recognition process.

The statistical measures described previously are used to quantify the temporal variation of feature importance for use as a statistical foundation for identifying anomalies. Average feature importance is an indicator of how much, on average, each feature contributes during a given "window" of time, while the standard deviation of feature importance is an indicator of how much variability exists in the contribution over time. Changes in the mean and standard deviation of feature importance indicate significant shifts in the distribution of feature importance, which can indicate a drift in the data distribution. Through a sliding window approach, the statistics produced are solely based on the most recent w data points, where w represents the window size.

C. DYNAMIC THRESHOLD CALIBRATION ALGORITHM BASED ON STATISTICS

The threshold algorithm establishes dual anomaly boundaries using statistical characteristics from the sliding window. The statistical characteristics for the dynamic threshold calibration algorithm consist of the mean and standard deviation of the feature importance distribution of the sliding window data. The framework implements symmetric thresholds to capture significant deviations across the entire distribution spectrum. The upper threshold identifies features with unexpectedly elevated importance weights while the lower threshold detects features with abnormally diminished importance weights.

$$\begin{aligned} \theta_i^{upper}(t) &= \mu_i(t) + \alpha \cdot \sigma_i(t), \\ \theta_i^{lower}(t) &= \mu_i(t) - \alpha \cdot \sigma_i(t) \end{aligned} \tag{6}$$

$\theta_i^{upper}(t)$ and $\theta_i^{lower}(t)$ represent the upper and lower dynamic anomaly detection thresholds of the i-th feature at time t, $\mu_i(t)$ is the mean feature importance of the i-th feature at time t, $\sigma_i(t)$ is the standard deviation of feature importance of the i-th feature at time t, and α is the threshold offset coefficient. The threshold shift coefficient α is determined by an optimization problem that minimizes the total error rate, which simultaneously considers the false positive rate and false negative rate constraints under concept drift conditions. Based on Gaussian mixture model analysis of feature importance distribution, $\alpha = 2.0$ ensures that the false alarm rate is less than 5% and remains sensitive to concept drift. This value satisfies the optimization condition of $\partial(\mathrm{FP}+\mathrm{FN})/\partial\alpha = 0$, where FP and FN represent the false positive and false negative probabilities respectively.

The anomaly determination process quantifies bidirectional deviations from expected importance ranges through two difference metrics that measure upper and lower threshold violations. The upper deviation metric captures features exceeding the upper threshold while the lower deviation metric identifies features falling below the lower threshold:

$$\begin{aligned} \delta_i^{upper}(t) &= w_i(t) - \theta_i^{upper}(t), \\ \delta_i^{lower}(t) &= \theta_i^{lower}(t) - w_i(t) \end{aligned} \tag{7}$$

$\delta_i^{upper}(t)$ and $\delta_i^{lower}(t)$ represent the upper and lower anomaly detection differences of the i-th feature at time t, $w_i(t)$ is the feature importance weight value of the i-th feature at time t, and $\theta_i^{upper}(t)$ and $\theta_i^{lower}(t)$ are the upper and lower dynamic anomaly detection thresholds respectively. Positive values in either difference metric indicate significant deviation from expected importance ranges.

The threshold calibration algorithm uses a recursive approach to maintain computational efficiency, updating the statistical measures within the sliding window each time a new data point arrives [29,30]. Figure 3 shows the response characteristics of the dynamic threshold calibration algorithm in concept drift events.

As shown in Figure 3, time points (1-11) illustrate stable period 1 within a fluctuation range of feature importance weights (0.43-0.51). The fluctuating dynamic threshold is in the range of (0.59-0.62) and is always less than the corresponding weights. At time point 12, it is possible to identify a concept drift (a change in the way data is generated; this event marks the occurrence of an anomaly), and feature importance weight increases while still being less than the dynamic threshold until time point 19. At time point 20, a second concept drift can be signified by the fact that the feature importance weight (0.64) has exceeded the dynamic threshold (0.61). This event activates the dynamic threshold adjustment mechanism. Following this concept drift event, feature importance weight continues to remain above the dynamic threshold from time points (21-28),

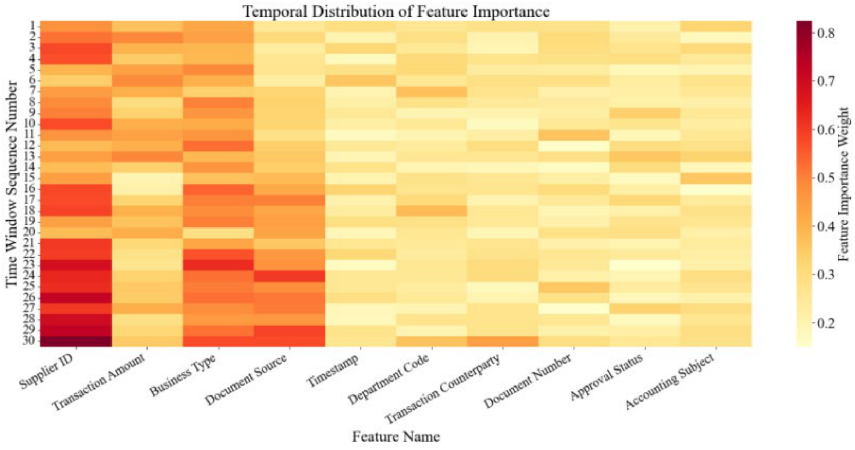

FIGURE 2. Time series distribution of feature importance

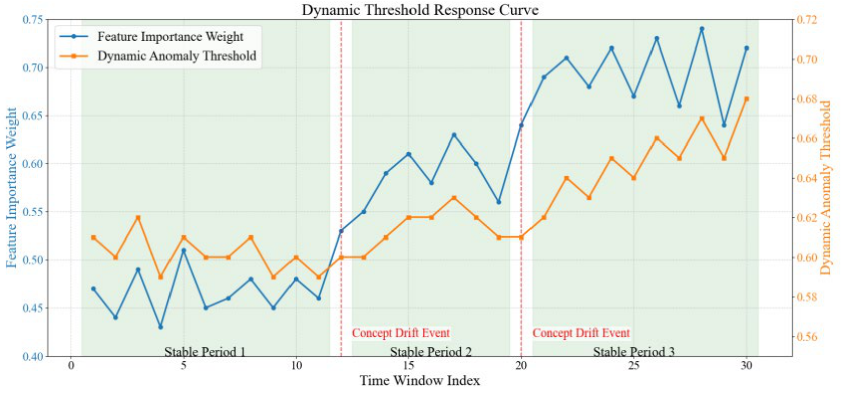

FIGURE 3. Dynamic Threshold Response Curve

signalling that a structural change occurred in the feature distribution during the time period of the concept drift. The lag between the time of change in feature importance and the time at which the dynamic threshold is adjusted reflects the fact that feature importance weights are cumulative in nature, confirming that the anomaly judgement boundary can be adjusted dynamically based on the mean/standard deviation.

The calibration process adjusts boundaries after concept drifts through statistical parameter recalculation [31,32]. The threshold for anomaly identification is thereby based on the mean and standard deviation of the changes in feature importance. As the feature distribution changes through time, thresholds can adjust automatically. The threshold calibration process allows for the flexibility of anomaly recognition to adapt to the changing environment without the use of predefined rules; it adjusts the threshold values based on the real-time characteristics of the credential data distribution. In this way, the threshold calibration maintains a constant, consistent logic in anomaly determination throughout the lifetime of the credential data.

D. ABNORMAL RECOGNITION OUTPUT MECHANISM WITH DECISION TRANSPARENCY

The real-time comparison process between the feature importance weight sequence and the dynamic threshold constitutes the final decision-making stage of anomaly recognition. The current feature importance weight is compared with the dynamic threshold calculated from the sliding time window on a feature by feature basis to form a feature level anomaly judgment difference. The feature importance weight sequence is cumulatively analyzed to obtain the contribution of each feature to the current voucher anomaly determination, and the dynamic threshold reflects the normal contribution range of specific features in recent audit business.

The difference in anomaly determination quantifies bidirectional deviations between the current feature contribution and the expected importance range defined by upper and lower thresholds [33,34]:

$$\begin{aligned} D_i^{upper}(t) &= w_i'(t) - \theta_i^{upper}(t), \\ D_i^{lower}(t) &= \theta_i^{lower}(t) - w_i'(t) \end{aligned} \tag{8}$$

$D_i^{upper}(t)$ and $D_i^{lower}(t)$ represent the upper and lower anomaly detection differences of the i-th feature at time t, $w_i'(t)$ represents the feature importance weight value of the

i-th feature at time t, and $\theta_i^{upper}(t)$ and $\theta_i^{lower}(t)$ represent the upper and lower dynamic anomaly detection thresholds of the i-th feature at time t. The overall degree of abnormality of the quantified voucher with abnormal comprehensive score incorporates both elevated and diminished importance anomalies:

$$S(t) = \sum_{i=1}^{n} \beta_i \cdot \max\left(D_i^{upper}(t), D_i^{lower}(t), 0\right) \quad (9)$$

S(t) represents the overall abnormal score of the voucher at time t, n represents the total number of features, and β_i is the feature weight adjustment coefficient, reflecting the differences in domain importance of different features in audit anomaly determination. The max function selects the larger of the two directional deviations or zero, ensuring only actual threshold violations contribute to the anomaly score. The anomaly tagging function produces binary results based on the comprehensive anomaly score:

$$M(t) = \begin{cases} 1, & \text{if } S(t) > \theta, \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

When the time index is equal to t, the value of the anomaly indicator M(t) is equal to 1, the value of the normal indicator is equal to 0, and the judgment threshold of the anomaly indicator is represented by θ. Along with the generation of the anomaly indicators, the system also produces feature contribution paths which consist of sequences of feature importance weights, the amount of variation between the determined anomalies, and the percentage of contribution that each feature has on the overall anomaly score, generating a basis for traceability for decision making processes.

Feature contribution paths preserve the original feature names associated with their weighted scores to create accountability in the decision making process through transparent verifiable information. The system provides real-time generation of anomaly indicators and generates a comprehensive decision path based upon the comparison of feature importance weight score values as well as dynamic thresholds for that specific time period. It provides the ability for auditors to trace back to the original indicator data and to review the basis for the anomaly decision at the feature level. The feature contribution path data is saved in a standardized format and includes key information, including timestamps, feature name, importance weight value, dynamic threshold, and the difference in judgment for the anomaly, to support the chain of evidence in audits. Figure 4 shows the complete process of the anomaly recognition output mechanism.

As shown in Figure 4, it consists of two main branches: the anomaly detection branch and the feature contribution path generation branch. The anomaly determination branch receives the current feature importance weight and dynamic threshold as inputs, and generates anomaly markers through comparison; The feature contribution path generation branch integrates the original feature name, feature importance weight sequence, anomaly judgment difference, and the contribution ratio of each feature to the overall anomaly score to form a structured output. The two branches are connected through a decision path tracking module to ensure synchronous generation of anomaly markers and feature contribution explanations.

III. EXPERIMENTAL SETUP

A. AUDIT VOUCHER DATASET AND ANOMALY ANNOTATION

Based on the Electronic Data Gathering, Analysis, and Retrieval System (EDGAR) of the US Securities and Exchange Commission, public financial report data covers three major industry sectors: manufacturing, finance, and retail. Table 1 presents the core parameters of the dataset.

TABLE 1. Dataset Parameters

Parameter	Numerical values	Illustrate
Industry distribution	Manufacturing 35%, Finance 40%, Retail 25%	Classified by standard industry classification codes
Sample size	350,000 voucher records	Valid samples after cleaning
Feature Dimension	47	Numerical type 23, Categorical type 21, Time series features 3
Data partitioning	Training set, validation set, test set	7:1:2

As shown in Table 1, the industry distribution is divided according to the standard industry classification code, with manufacturing accounting for 35%, finance accounting for 40%, and retail accounting for 25%. The sample size of 350000 voucher records is valid data after cleaning. There are a total of 47 feature dimensions, among which 23 numerical features include financial indicators such as transaction amounts, 21 categorical features cover attributes such as business types, and 3 temporal features are used to capture business cyclicality. The dataset is divided into training set, validation set, and testing set in a ratio of 7:1:2 to ensure the preservation of time series characteristics.

B. MODEL PARAMETERS AND COMPARATIVE EXPERIMENTAL CONFIGURATION

The size of the sliding time window is 15 time units. The cyclical nature of audit voucher generation is well captured by the window length. The batch size is set to 128, the mask sparsity adjustment parameter is fixed at 0.3, and the TabNet feature selection iteration is set to five times. The optimizer employs the Adam algorithm, and the cosine annealing strategy attenuates the learning rate from 0.001 to 0.0001 [35,36]. To satisfy the demands of categorical feature representation, the feature embedding dimension is consistently set to 8. The dynamic threshold calibration algorithm uses a threshold offset coefficient of 2.0.

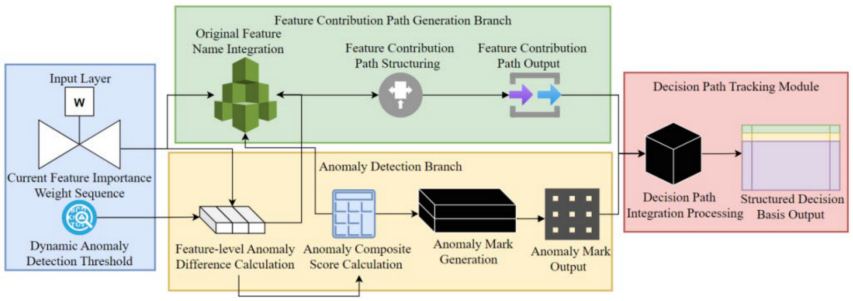

FIGURE 4. Abnormal recognition output and feature contribution path generation

IV. RESULT ANALYSIS

A. COMPARISON OF ANOMALY RECOGNITION PERFORMANCE

The primary challenge in audit voucher anomaly recognition is striking a balance between decision transparency and deep learning model detection accuracy. The performance of the TabNet adaptive recognition framework for five high-frequency audit voucher anomalies—amount logic conflicts, time logic anomalies, business type anomalies, supplier anomalies, and transaction frequency anomalies—is methodically assessed in this study. The main manifestation of abnormal amount logic conflicts is financial logic contradictions such as imbalanced borrowing and lending. Time logic anomalies reflect the mismatch between voucher timestamps and business cycles. Business type anomalies reflect the deviation between transaction nature and enterprise business scope. Supplier anomalies involve abnormal associations between supplier qualifications and transaction behavior. Transaction frequency anomalies refer to fluctuation patterns where transaction frequency exceeds the normal business scope. Accuracy measures the overall correct recognition ability of the model, recall reflects the ability to capture abnormal samples, and accuracy characterizes the reliability of recognition results. The F1 score serves as a harmonic average comprehensive evaluation model for accuracy and recall, and these indicators together constitute a multidimensional anomaly detection capability evaluation system. Table 2 presents the quantitative evaluation results of five types of anomaly detection.

TABLE 2. Multi-dimensional evaluation of TabNet model anomaly detection performance

Type	Accuracy (%)	Recall rate (%)	Precision rate (%)	F1 score (%)
Amount logic conflict	92.3	88.7	90.5	89.6
Time logic anomaly	85.6	91.2	87.2	89.2
Business type anomaly	89.8	86.5	88.3	87.4
Supplier anomaly	93.1	87.4	91.8	89.5
Abnormal transaction frequency	87.5	90.1	86.7	88.4

As shown in Table 2, the accuracy of detecting logical conflict anomalies in amounts reached 92.3%, with a recall rate of 88.7%, indicating that the model can effectively identify financial logic conflict anomalies, which is consistent with the core position of numerical features in audit vouchers. The accuracy rate of supplier anomaly detection is as high as 93.1%, with an accuracy rate of 91.8%. This verifies the temporal variation characteristics of the importance of supplier coding features, indicating that the model is highly sensitive to supplier related anomalies. For time logical anomalies, the recall is at a high of 91.2%, which indicates that the model can identify time-related anomalies. Additionally, the business-type anomaly detection accuracy is at 89.8%, which is indicative of the complexity of categorical variables as well as the large variances in business-type audit scenarios. The F1 scores across all anomaly categories exceeded 87%, which indicates the balance between recall and precision has been successfully established from the use of the sliding time window-calibrated dynamic feature importance approach, based on which it can be concluded that the anomaly detection results as accurate and reliable.

B. CONSISTENCY ANALYSIS BETWEEN INTERPRETABLE OUTPUT AND AUDIT EVIDENCE CHAIN

To meet the qualifications of professional verification, the decision-making process behind the identification of audit voucher anomalies must be transparent. In this section, this study compares the feature contributions generated by interpretable deep learning models with those derived from twelve certified public accountants having fifteen years average professional experience, six from international accounting firms, who conducted independent feature importance assessments following ISA(International Standards on Auditing) guidelines through a three-round Delphi process with geometric mean calculation determining final expert weights. The objective is to determine if there is consistency between the algorithmic decision-making process and current industry standards. The research focuses on three forms of high frequency anomalies: anomalies associated with logic conflicts due to amounts that violate fundamental accounting principles (e.g. loan balance), supplier anomalies due to discrepancies between the supplier's qualifications and the supplier's actions, and time logic anomalies due to differences between the timing of business operations and the timing of financial transactions. Feature importance weights quantify each variable's contribution to anomaly detection. A value close to 1.0 suggests that the feature dominates in decision-making. Figure 5 shows the comparative analysis results of model calculation weights and expert

evaluation weights in three typical abnormal scenarios, and the feature contribution path consistency verification framework.

Figure 5 (A) shows that the transaction amount in the amount logic conflict anomaly is highly consistent with the weight of the loan balance feature (all coefficients are>0.8), confirming the core position of accounting identity in audit judgment; The weight of the voucher type feature model (0.55) is slightly higher than that of expert evaluation (0.51), reflecting the sensitivity of the algorithm to formal compliance. Figure 5 (B) shows the difference in the weight of supplier codes in supplier anomaly recognition (model 0.87, expert 0.83), while the expert weight of association relationship features (0.58) is higher than that of the model (0.52), indicating that manual judgment focuses more on business substance rather than surface identification. Figure 5 (C) reveals that the trading time weights in time logic anomalies are highly consistent (model 0.83, expert 0.86), and the seasonal factor expert evaluation (0.65) exceeds the model calculation (0.62), reflecting the compensation mechanism of professional experience for cyclical fluctuations. The correlation coefficients of core feature weights (transaction amount, supplier code, transaction time) in the three types of anomalies all exceed 0.8, confirming the systematic consistency between the decision path generated by TabNet and the audit professional logic, providing empirical basis for the reliable deployment of the algorithm in audit practice.

C. EVALUATION OF THE EFFECTIVENESS OF ADAPTIVE MECHANISMS IN MITIGATING CONCEPT DRIFT

System adaptability depends on the threshold mechanism's response to concept drift during anomaly identification. The impact mechanism of window parameters on the adaptability of anomaly detection boundaries is revealed in this section by methodically examining the changes in threshold adjustment lag time, threshold convergence time, and error after stabilization under various sliding window sizes. The system's response time to modify the judgment boundary following the concept drift event is reflected in the threshold adjustment lag time. The time needed for the threshold to stabilize is represented by the threshold convergence time, and the error following stabilization measures how much the final threshold deviates from the ideal value. These three indicators together constitute the core dimension for evaluating the timeliness of the dynamic threshold calibration mechanism, where the difference between the lag time and the convergence time directly determines the system's adaptation speed to environmental changes, and the error after stabilization affects the accuracy of anomaly identification. The theoretical minimum error is achieved when w=15, which is consistent with the 15 time unit characteristic business cycle length derived from the feature importance sequence autocorrelation analysis. The observed error curve follows the theoretically predicted quadratic relationship for window size, confirming the mathematical validity of the sliding window approach.

Figure 6 (a) shows that the threshold adjustment lag time increases with the increase of the window, with an average of 2.99 time points for 10 windows, 4.95 time points for 15 windows, 6.86 time points for 20 windows, 8.97 time points for 25 windows, and 11.42 time points for 30 windows. This nonlinear growth is due to the concept drift response delay caused by the cumulative characteristics of sliding window statistics. The convergence time of the threshold in Figure 6 (b) shows a similar trend, with an average of 7.01 time points for 10 windows, 10.4 time points for 15 windows, 13.87 time points for 20 windows, 16.01 time points for 25 windows, and 19.56 time points for 30 windows, consistent with the time relationship of concept drift events. Figure 6 (c) reveals that the error is minimized at 15 windows after stabilization, with an average of 0.055, 0.163 at 10 windows, 0.075 at 20 windows, 0.095 at 25 windows, and 0.125 at 30 windows. The calibration mechanism of threshold offset coefficient dynamically adjusts the anomaly judgment boundary, enabling the system to respond to environmental changes in a timely manner after concept drift, while maintaining the continuity and consistency of anomaly recognition, reflecting the adaptability of dynamic threshold calibration algorithm in audit voucher anomaly recognition.

D. THE ROLE OF MODEL INTERPRETABILITY IN IMPROVING AUDIT DECISION-MAKING EFFICIENCY

One important metric for assessing the usefulness of intelligent audit systems is audit verification efficiency. This section quantitatively examines how the feature importance output mechanism optimizes the manual review process of abnormal vouchers and focuses on the substantive impact of model interpretability on auditor verification. The workload of auditors in examining system tagging anomalies is directly reflected in the verification time, a fundamental efficiency indicator; The degree of process optimization brought about by interpretable output is quantified by the efficiency improvement rate of verification; the accuracy of the verification guarantees that the efficiency improvement does not come at the expense of compromising the reliability of the results. To evaluate these indicators for the system, the processing time of five typical anomalies by auditors during the actual verification process was recorded, and the results are shown in Table 3.

As shown in Table 3, the validation method using feature importance output resulted in a decrease in the average validation time for various types of anomalies. The validation time for supplier anomalies decreased from 19.6 minutes to 8.4 minutes, with an efficiency improvement rate of 57.1%. The reduction in verification time for abnormal business types is 49.4%. The complexity of categorical features involved in such anomalies and the variability of business types in audit scenarios still require a lot of manual judgment in the verification process. The validation

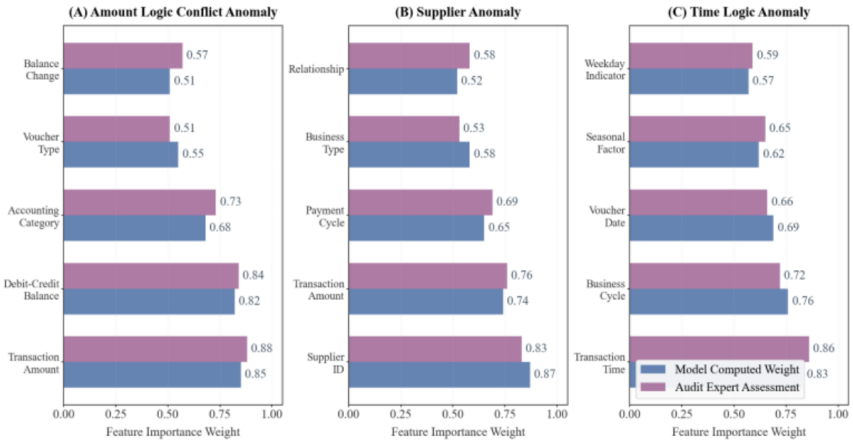

FIGURE 5. Verification of Consistency between Feature Contribution Paths and Audit Standards under Three Typical Audit Anomaly Modes. Figure 5 (A) Abnormal amount logic conflict; Figure 5 (B) Supplier Exception; Figure 5 (C) Time logic exception

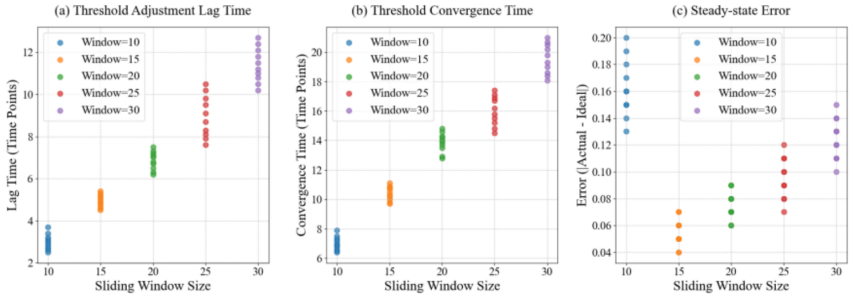

FIGURE 6. Influence of Sliding Window Size on the Timeliness of Dynamic Threshold Response. Figure 6 (a) Threshold adjustment lag time; Figure 6 (b) Threshold convergence time; Figure 6 (c) Error after stabilization

TABLE 3. Comparative Analysis of Audit Verification Efficiency

Type	Average time taken by traditional verification methods (min)	The average time taken by the verification method in this article (min)	Verification efficiency improvement rate (%)	Verification accuracy (%)
Amount logic conflict	18.7	9.2	50.8	96.5
Time logic anomaly	22.4	10.5	53.1	94.8
Business type anomaly	25.3	12.8	49.4	93.2
Supplier anomaly	19.6	8.4	57.1	95.7
Abnormal transaction frequency	21.8	11.2	48.6	94.1

accuracy of all types of anomalies remains above 93%, and the interpretability output mechanism improves efficiency without reducing the reliability of the results. The decision path, generated by comparing feature importance weights with dynamic thresholds, helps auditors quickly identify anomaly causes, eliminating the need for comprehensive feature reviews in traditional methods. The improvement rate of verification efficiency is related to the consistency of feature contribution paths, with significant supplier anomalies (57.1%) and temporal logic anomalies (53.1%). The consistency between the decision path generated by the algorithm and the audit professional logic directly determines the verification efficiency. The feature importance weight sequence, anomaly judgment difference, and contribution ratio of each feature to the overall anomaly score output by the model provide traceable technical support for the audit evidence chain, effectively reducing manual verification time.

E. GENERALIZATION ABILITY OF THE FRAMEWORK IN MULTI-INDUSTRY AUDIT SCENARIOS

The audit voucher anomaly recognition system needs to adapt to the business characteristics of different industry scenarios. This section explores the adaptive mechanism of the model to industry characteristics by comparing the importance distribution characteristics of key audit features in the three major fields of manufacturing, finance, and retail. The supply chain structure in the manufacturing industry is complex, and supplier coding features play a central role in voucher verification; The financial industry has strict requirements for the accuracy of transaction amounts; The retail industry is significantly affected by the sales cycle. Figure 7 presents the fluctuation characteristics of the importance of each feature in different industries, showing a comparison of the importance distribution of five key audit features: transaction amount, transaction time, supplier code, voucher type, and business type in three major

industry scenarios.

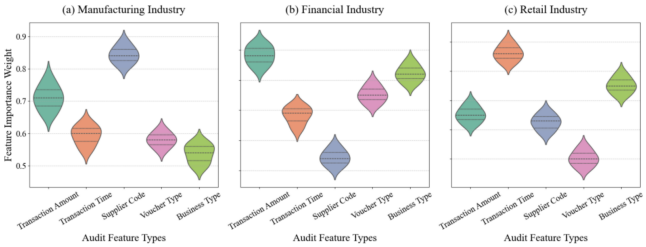

FIGURE 7. Comparison of Importance Distribution of Key Features in Audit Scenarios of Different Industries. Figure 7 (a) Manufacturing Industry; Figure 7 (b) Financial Industry; Figure 7 (c) Retail Industry

In Figure 7 (a), the average importance of supplier coding features in the manufacturing industry scenario is 0.843, which is higher than other features. The distribution pattern is narrow and concentrated, indicating the stable dependence of the industry on supplier qualification verification. Figure 7 (b) shows that in the financial industry, the average importance of transaction amount features is highest at 0.883, with a wide distribution range, reflecting the high sensitivity of the financial industry to abnormal amounts but with significant fluctuations. Figure 7 (c) shows that the average importance of transaction time features in the retail industry scenario is 0.862, with a compact distribution, reflecting the decisive impact of sales cycles on retail business and verifying the high recall rate of time logic anomalies. The differences in the distribution of characteristic importance among different industries map the essential characteristics of the business. The complexity of the manufacturing supply chain results in the highest average importance of supplier coding characteristics, while the liquidity requirements of the financial industry lead to a significant difference in the average importance of transaction amount characteristics compared to other types of characteristics. The cyclical nature of retail business results in the highest average importance of transaction time features, indicating that the model dynamically adjusts the focus of feature attention based on audit scenarios, maintaining environmental adaptability and professional reliability for anomaly recognition.

V. CONCLUSION

This article introduces a TabNet adaptive recognition framework based on dynamic calibration of sliding time window feature importance, aiming to address the dual challenges of environmental adaptability and decision transparency in anomaly identification. According to the experimental results, supplier anomalies can be detected with 93.1% accuracy and logical conflicts in amounts can be detected with 92.3% accuracy. The verification process now takes 8.4 minutes instead of 19.6 minutes. The manufacturing sector has an average importance of 0.843 for supplier code features, the financial sector has an average importance of 0.883 for transaction amount features, and the retail sector has an average importance of 0.862 for transaction time features. The consistency between the feature contribution path and audit standards is confirmed by the correlation coefficient exceeding 0.8 between the feature importance weight and the audit expert evaluation weight on core features. In order to adapt to shifts in the audit business cycle, this framework automatically modifies the focus of feature attention in various industry scenarios.

AUTHOR CONTRIBUTIONS

Lixuan Sun designed, collected and analyzed the data, and drafted the manuscript. Lixuan Sun conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Lixuan Sun participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately inves tigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] N. T. Popoola, "Big Data-Driven Financial Fraud Detection and Anomaly Detection Systems for Regula-Tory Compliance and Market Stability," Int. J. Comput. Appl. Technol. Res, vol. 12, no. 9, pp. 32-46, 2023, doi: 10.7753/IJCATR1209.1004.

[2] S. C. Friday, C. I. Lawal, D. C. Ayodeji, and A. Sobowale, "Reviewing the effectiveness of digital audit tools in enhancing corporate transparency," International Journal of Advanced Multidisciplinary Research and Studies, vol. 6, no. 4, pp. 1679-1689, 2024, doi: 10.62225/2583049X.2024.4.6.4099.

[3] A. Al-Omush, A. Almasarwah, and A. Al-Wreikat, "Artificial Intelligence in Financial Auditing: Redefining Accuracy and Transparency in Assurance Services," EDPACS, vol. 70, no. 6, pp. 1-20, 2025, doi: 10.1080/07366981.2025.2459490.

[4] Q. Chen, J. Lin, X. Chen, and X. Wang, "Research on the Application of RPA+ AI Technology in the Con-struction of Paperless Intelligent System for Finance in Colleges and Universities," Journal of Mod-ern Social Sciences, vol. 1, no. 2, pp. 368-383, 2025, doi: 10.5281/zenodo.14553755.

[5] S. K. Afadzinu, L. D. David, I. Pothaczky Racz, and F. Jemimah, "The impact of technological innovations on audit transparency, objectivity, and assurance in the digital era," Journal of Infrastructure Policy and Development, vol. 8, no. 14, pp. 8241-8262, 2024, doi: 10.24294/jipd8241.

[6] G. Chaudhary, "Unveiling the black box: Bringing algorithmic transparency to AI," Masaryk University Journal of Law and Technology, vol. 18, no. 1, pp. 93-122, 2024, doi: 10.5817/MUJLT2024-1-4.

[7] D. Goswami, "ADVANCING THREAT DETECTION THROUGH ARTIFICIAL INTELLI-GENCE AND MACHINE LEARNING ENHANCED CYBERSECURITY AUDITS," Ameri-can Journal of Scholarly Research and Innovation, vol. 4, no. 1, pp. 428-457, 2025, doi: 10.63125/gb5s3f54.

[8] D. Brand, M. E. Hoffmann, and J. Van der Merwe, "Development of effective methods and tools for the auditing AI algorithms by Supreme Audit Institutions," JeDEM-eJournal of eDemocracy and Open Government, vol. 17, no. 3, pp. 106-140, 2025, doi: 10.29379/jedem.v17i3.1049.

[9] H. Zhong, D. Yang, S. Shi, L. Wei, and Y. Wang, "From data to insights: the application and challenges of knowledge graphs in intelligent audit," Journal of Cloud Computing, vol. 13, no. 1, pp. 114-140, 2024, doi: 10.1186/s13677-024-00674-0.

[10] V. Ganapathy, "AI in auditing: A comprehensive review of applications, benefits and challenges," Shodh Sari-An International Multidisciplinary Journal, vol. 2, no. 4, pp. 328-343, 2023, doi: 10.59231/SARI7643.

[11] F. T. H. Lo, "The paradoxical transparency of opaque machine learning," AI & SOCIETY, vol. 39, no. 3, pp. 1397-1409, 2024, doi: 10.1007/s00146-022-01616-7.

[12] E. SAHiN, N. N. Arslan, and D. Ozdemir, "Unlocking the black box: an in-depth review on interpreta-bility, explainability, and reliability in deep learning," Neural Computing and Applications, vol. 37, no. 2, pp. 859-965, 2025, doi: 10.1007/s00521-024-10437-2.

[13] O. Ilori, N. T. Nwosu, and H. N. N. Naiho, "Advanced data analytics in internal audits: A conceptual framework for comprehensive risk assessment and fraud detection," Finance & Accounting Re-search Journal, vol. 6, no. 6, pp. 931-952, 2024, doi: 10.51594/farj.v6i6.1213.

[14] D. Pelosi, D. Cacciagrano, and M. Piangerelli, "Explainability and Interpretability in Concept and Data Drift: A Systematic Literature Review," Algorithms, vol. 18, no. 7, pp. 443-525, 2025, doi: 10.3390/a18070443.

[15] D. Zhou and J. He, "Rare Category Analysis for Complex Data: A Review," ACM Computing Surveys, vol. 56, no. 5, pp. 1-35, 2023, doi: 10.1145/3626520.

[16] S. I. Kampezidou, A. Tikayat Ray, A. P. Bhat, O. J. Pinon Fischer, and D. N. Mavris, "Fundamental components and principles of su-pervised machine learning workflows with numerical and categorical data," Eng, vol. 5, no. 1, pp. 384-416, 2024, doi: 10.3390/eng5010021.

[17] Z. Wang, "Adaptive Ensemble Learning Framework with SHAP-Based Feature Optimization for Financial Anomaly Detection," Artificial Intelligence and Machine Learning Review, vol. 5, no. 1, pp. 51-66, 2024, doi: 10.69987/AIMLR.2024.50105.

[18] Y. Gebreyesus, D. Dalton, D. De Chiara, M. Chinnici, and A. Chinnici, "AI for automating data center operations: model ex-plainability in the data centre context using shapley additive explanations (SHAP)," Electronics, vol. 13, no. 9, pp. 1628-1647, 2024, doi: 10.3390/electronics13091628.

[19] S. R. A. Parisineni and M. Pal, "Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations," International Journal of Data Science and Analytics, vol. 18, no. 4, pp. 457-466, 2024, doi: 10.1007/s41060-023-00458-w.

[20] A. Chinnaraju, "Explainable AI (XAI) for trustworthy and transparent decision-making: A theoret-ical framework for AI interpretability," World Journal of Advanced Engineering Technology and Sciences, vol. 14, no. 3, pp. 170-207, 2025, doi: 10.30574/wjaets.2025.14.3.0106.

[21] T. A. Shaikh, T. Rasool, P. Verma, and W. A. Mir, "A fundamental overview of ensemble deep learning models and applications: systematic literature and state of the art," Annals of Operations Research, no. 1, pp. 1-77, 2024, doi: 10.1007/s10479-024-06444-0.

[22] M. Sakib, S. Mustajab, and M. Alam, "Ensemble deep learning techniques for time series analysis: a comprehensive review, applications, open issues, challenges, and future directions," Cluster Computting, vol. 28, no. 1, pp. 73, 2025, doi: 10.1007/s10586-024-04684-0.

[23] D. Zegarra Rodriguez, O. Daniel Okey, S. S. Maidin, E. Umoren Udo, and J. H. Kleinschmidt, "Attentive transformer deep learning al-gorithm for intrusion detection on IoT systems using automatic Xplainable feature selection," Plos one, vol. 18, no. 10, Art. no. e0286652-e0286676, 2023, doi: 10.1371/journal.pone.0286652.

[24] Abdullah, M. Ateeb Ather, O. Kolesnikova, and G. Sidorov, "Detection of Biased Phrases in the Wiki Neu-trality Corpus for Fairer Digital Content Management Using Artificial Intelligence," Big Data and Cognitive Computing, vol. 9, no. 7, pp. 190-221, 2025, doi: 10.3390/bdcc9070190.

[25] W. Liu, K. Liu, W. Sun, G. Yang, K. Ren, X. Meng, et al., "Self-supervised feature learning based on spectral masking for hyperspectral image classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, no. 1, pp. 1-15, 2023, doi: 10.1109/TGRS.2023.3310489.

[26] Z. Huang, X. Jin, C. Lu, Q. Hou, M. M. Cheng, D. Fu, et al., "Contrastive masked autoencoders are stronger vision learners," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 4, pp. 2506-2517, 2023, doi: 10.1109/TPAMI.2023.3336525.

[27] N. Mazumdar and P. K. D. Sarma, "Sequential pattern mining algorithms and their applications: a technical review," International Journal of Data Science and Analytics, vol. 20, no. 3, pp. 1683-1726, 2025, doi: 10.1007/s41060-024-00659-x.

[28] S. Naz, M. H. Khan, and M. S. Farid, "Enhancing fall detection using multimodal time-series sensory data and VLAD encoding: a framework: S," Naz et al. The Journal of Supercomputing, vol. 81, no. 8, pp. 890, 2025, doi: 10.1007/s11227-025-07370-z.

[29] E. A. Asiamah, N. K. Akrasi-Mensah, P. Odame, E. Keelson, A. S. Agbemenu, E. T. Tchao, et al., "A storage-efficient learned indexing for blockchain systems using a sliding window search enhanced online gradient descent," The Journal of Supercomputing, vol. 81, no. 1, pp. 1-29, 2025, doi: 10.1007/s11227-024-06805-3.

[30] C. Chen, J. Liang, W. Sun, G. Yang, and X. Meng, "An automatically recursive feature elimination method based on threshold decision in random forest classification," Geo-spatial information science, vol. 28, no. 4, pp. 1494-1519, 2025, doi: 10.1080/10095020.2024.2387457.

[31] E. B. Gulcan and F. Can, "Unsupervised concept drift detection for multi-label data streams," Artificial Intelligence Review, vol. 56, no. 3, pp. 2401-2434, 2023, doi: 10.1007/s10462-022-10232-2.

[32] D. Lukats, O. Zielinski, A. Hahn, and F. Stahl, "A benchmark and survey of fully unsupervised concept drift detectors on real-world data streams," International Journal of Data Science and Analytics, vol. 19, no. 1, pp. 1-31, 2025, doi: 10.1007/s41060-024-00620-y.

[33] M. A. Belay, S. S. Blakseth, A. Rasheed, and P. Salvo Rossi, "Unsupervised anomaly detection for IoT-based multivariate time series: Existing solutions, performance analysis and future directions," Sensors, vol. 23, no. 5, pp. 2844-2867, 2023, doi: 10.3390/s23052844.

[34] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, "Deep learning for time series anomaly detection: A survey," ACM Computing Surveys, vol. 57, no. 1, pp. 1-42, 2024, doi: 10.1145/3691338.

[35] J. Yang and Q. Long, "A modification of adaptive moment estimation (adam) for machine learning," Journal of Industrial and Management Optimization, vol. 20, no. 7, pp. 2516-2540, 2024, doi: 10.3934/jimo.2024014.

[36] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimiza-tion," Neural Computing and Applications, vol. 35, no. 23, pp. 17095-17112, 2023, doi: 10.1007/s00521-023-08568-z.