Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

HighTech and Innovation Journal

Vol. 7, No. 2, June, 2026

Temporal-Semantic Fusion Network for Identification of Online Gambling and Child Exploitation Financial Transactions

Valentinus Paramarta [1], Alrafiful Rahman [1], Harya Widiputra [1*]

[1] *Faculty of Information Technology, Perbanas Institute, Setiabudi, Jakarta 12940, Indonesia.*

Received 06 March 2026; Revised 25 May 2026; Accepted 27 May 2026; Published 01 June 2026

Abstract

This study addresses the detection of illicit digital payments, specifically online gambling and child exploitation, which are frequently hidden within legitimate transaction streams. The primary objective is to overcome the limitations of traditional rule-based systems and unimodal models that struggle with class imbalance and sophisticated evasion. It is proposed that the Temporal-Semantic Fusion Network (TSFN), a novel architecture integrating Temporal Convolutional Networks (TCN) for numerical sequences and FinBERT for semantic textual encoding. The key novelty is a bidirectional cross-modal attention mechanism that enables dynamic information exchange between behavioral patterns and transaction descriptions. Evaluated on 10,000 synthetic transactions, TSFN achieved a macro F1-score of 0.847, outperforming concatenation-based fusion by 6.5 percentage points ($p < 0.001$). Significant improvements were noted in minority classes, with F1-scores of 0.823 for gambling and 0.741 for exploitation, while maintaining a 99.4% precision rate on legitimate data. Ablation studies confirm that bidirectional attention allows the model to adaptively prioritize temporal features for gambling and semantic cues for exploitation. This research provides a robust framework for multimodal financial crime detection, offering a significant improvement in identifying complex illicit patterns compared to existing benchmarks.

Keywords: TSFN; Multimodal Fusion; Illicit Financial Activities; Cross-Modal Attention; Financial Crime Detection.

1. Introduction

Digital payment systems have fundamentally transformed global commerce, enabling instantaneous transactions across borders and facilitating economic inclusion for billions of users worldwide [1, 2]. Yet this transformation has created new vulnerabilities that criminals exploit with increasing sophistication. Online gambling platforms and child exploitation networks have migrated to digital payment channels, embedding illicit transactions within the massive volume of legitimate payments that financial institutions process daily [3, 4]. The scale of these crimes is staggering. The global online gambling market exceeded $60 billion in 2024, with significant portions operating illegally [5]. Meanwhile, child exploitation material transactions have proliferated across digital payment networks [6].

Traditional detection approaches rely primarily on rule-based systems that flag transactions matching predefined patterns: specific merchant categories, amount thresholds, geographic locations, or blacklisted accounts [7]. While computationally efficient, these systems suffer from fundamental limitations. Criminals adapt quickly to known detection rules, modifying transaction characteristics to evade filters. Rule-based systems generate high false positive rates, overwhelming investigators with alerts that prove legitimate upon review. They struggle with the extreme class imbalance inherent in financial crime detection, where illicit transactions constitute less than 2% of total volume [8].

* Corresponding author: harya@perbanas.id

https://doi.org/10.28991/HIJ-2026-07-02-024

Most critically, rule-based approaches cannot leverage the rich multimodal information available in modern payment systems: numerical features like amounts and timestamps combined with textual features like transaction descriptions and merchant names.

Recent machine learning approaches have improved detection performance by learning patterns from historical data rather than relying on manually crafted rules [9, 10]. However, most existing work treats financial crime detection as a unimodal problem, processing either numerical transaction features or textual descriptions, but not both [11, 12]. The few multimodal approaches that exist typically employ simple fusion strategies, concatenating features from separate modality-specific encoders before classification [13]. This concatenation approach treats modalities as independent information sources, missing potential synergies between temporal behavioral patterns and semantic textual cues.

Effective detection of sophisticated illicit activities requires joint reasoning over both temporal transaction patterns and semantic textual information, supported by dynamic fusion that adapts to the characteristics of each specific case. Online gambling transactions, for example, often exhibit distinctive temporal clustering, such as multiple small transactions to the same merchant within short time windows, particularly during evenings and weekends. However, merchants may use euphemistic descriptions, such as "entertainment services," "gaming credits," or "digital content," which overlap with legitimate purchases. Detecting gambling therefore requires recognizing the combination of suspicious temporal patterns and ambiguous textual descriptions, where each modality provides partial evidence that becomes more conclusive when considered together.

Child exploitation transactions present different challenges. They typically lack the temporal regularity of gambling, occurring sporadically and often isolated within otherwise normal transaction sequences. Amounts are usually small, mimicking legitimate digital content purchases. The primary signal comes from deliberately vague textual descriptions ("digital media," "online content access") that perpetrators use to obscure transaction purposes. Yet these descriptions, while suspicious, are not definitively incriminating without corroborating evidence from transaction context. Effective detection requires the model to emphasize semantic features while using temporal context to filter false positives.

Recent studies in 2025 have highlighted the increasing complexity of money laundering through decentralized payment gateways [14]. While Transformer-based models have improved financial text analysis, they often operate in isolation from numerical behavioral data. Furthermore, recent work by Chen et al. [15] demonstrated that simple late-fusion techniques fail to capture the inter-modality dependencies required to distinguish between 'gaming' for entertainment and 'gambling' as an illicit activity. Specifically, the unique temporal clustering of illicit betting behaviors identified in 2025 necessitates models that can capture nuanced periodicities [16]. To address the scarcity of real-world data in this field, high-fidelity synthetic data generation has emerged as a critical standard for training robust forensic models in 2026 [17]. Existing literature largely overlooks the synergy between temporal rhythms and semantic nuances. Our TSFN model fills this gap by introducing a bidirectional attention layer that allows each modality to contextually inform the other [18].

This paper introduces TSFN (Temporal-Semantic Fusion Network), a deep learning architecture specifically designed for multimodal financial crime detection. TSFN processes numerical transaction sequences through Temporal Convolutional Networks (TCN) with exponentially increasing dilation rates, capturing both short-term anomalies and long-range dependencies. Textual descriptions are encoded using FinBERT [19], a financial domain-adapted language model, followed by transformer-based sequential encoding. Our key innovation is a bidirectional cross-modal attention mechanism that enables dynamic information exchange between temporal and semantic modalities. Unlike simple concatenation, which applies fixed fusion weights, bidirectional attention allows temporal features to query semantic features and vice versa, creating representations that encode inter-modality relationships explicitly.

The main contributions of this work are:

- A bidirectional cross-modal attention mechanism for multimodal fusion that enables dynamic information exchange between temporal and semantic modalities, outperforming simple concatenation by 6.5 percentage points.
- A hybrid loss function combining focal loss and supervised contrastive learning that effectively handles extreme class imbalance (1:67.2 ratio) while learning discriminative representations.
- Comprehensive evaluation on 10,000 synthetic transactions demonstrating strong performance on minority classes (F1: 0.823 gambling, 0.741 exploitation) while maintaining high precision (99.4%) on legitimate transactions.
- Detailed analysis of learned attention patterns revealing class-specific detection strategies, with the model adaptively emphasizing temporal features for gambling and semantic features for exploitation.
- Extensive ablation studies quantifying the contribution of each architectural component and validating design choices including bidirectional attention, hybrid loss, TCN depth, and sequence length.

The remainder of this paper is organized as follows. Section 2 reviews related work on financial crime detection, multimodal learning, temporal convolutional networks, semantic modeling, attention mechanisms, and identifies research gaps. Section 3 describes the TSFN architecture, training procedure, and evaluation methodology. Section 4 presents experimental results including baseline comparisons, ablation studies, and attention analysis. Section 5 discusses why bidirectional attention works, interprets class-specific patterns, and examines practical deployment considerations. Section 6 concludes with limitations and future research directions.

2. Related Works

2.1. Financial Crime Detection

Financial crime detection has evolved from rule-based systems to machine learning approaches over the past decade. Early work focused on feature engineering and traditional classifiers [7, 20]. Meanwhile, Dal Pozzolo et al. [8] demonstrated that learned features outperform hand-crafted rules for credit card fraud detection, achieving substantial improvements in both precision and recall. However, their approach treated fraud detection as a unimodal problem processing only numerical transaction features.

Recent deep learning approaches have shown promise for capturing complex patterns in financial data. For instance, Shenvi et al. [9] applied Long Short-Term Memory (LSTM) networks to credit card transaction sequences, demonstrating that recurrent models can learn temporal dependencies that static classifiers miss. Whereas, He & Zhao [11] extended this work with attention-based LSTM, showing that focusing on the most relevant transactions in a sequence improves detection performance. Yet these temporal models ignore textual information entirely, missing valuable signals from merchant descriptions and transaction notes.

Graph-based approaches represent an alternative paradigm that represents relationships between entities. In relation to this, Wu et al. [21] proposed heterogeneous graph neural networks for fraud detection, constructing graphs where nodes represent users and merchants with edges representing transactions. Their approach achieved strong performance by leveraging network structure to identify coordinated fraud rings. However, graph methods face scalability challenges with millions of users and merchants, and they primarily process graph structure rather than sequential transaction patterns.

The class imbalance problem has received substantial attention in financial crime detection research. Dal Pozzolo et al. [8] showed that standard training procedures fail when positive examples constitute less than 1% of the data, while [10] compared various resampling and cost-sensitive learning approaches and found that focal loss [22] provides the best balance between minority-class recall and false positive rates. Building on this work, the present study combines focal loss with contrastive learning to improve representation quality.

2.2. Multimodal Learning for Financial Applications

While multimodal learning has achieved remarkable success in vision-language tasks [23, 24], its application to financial transaction monitoring remains limited. In previous work, Alaygut & Sefer [12] explored combining numerical features with textual merchant descriptions using separate encoders followed by concatenation. They reported modest improvements (3–5 pp F1) over unimodal baselines, suggesting that textual information adds value but simple concatenation may not fully exploit multimodal synergies.

Moreover, Wang et al. [13] proposed a more sophisticated concatenation-based approach for credit card fraud detection, using convolutional neural networks for numerical features and BERT for textual features. Their concatenation baseline achieved F1 of 0.782, representing the current state-of-the-art for multimodal financial fraud detection. However, their fusion strategy treats modalities independently. Each encoder processes its input without considering the other modality, and fusion occurs only at the final classification layer.

In addition, Gadzicki et al. [25] investigated late fusion strategies where separate classifiers are trained independently for each modality and their predictions are combined through weighted averaging. They found that late fusion (F1: 0.698) underperforms joint training approaches, likely because independent training prevents modalities from learning complementary representations.

Domain adaptation of language models for financial text has emerged as an important research direction. Passas [14] introduced FinBERT, a BERT model pre-trained on financial news and earnings reports, demonstrating substantial improvements over generic BERT in financial sentiment analysis. Yang et al. [26] subsequently extended this work to financial document classification. In the present study, FinBERT is leveraged to encode transaction descriptions.

2.3. Temporal Convolutional Networks

Temporal Convolutional Networks (TCN) have emerged as effective alternatives to recurrent neural networks for sequence modeling [26]. Unlike LSTMs that process sequences recurrently, TCNs use dilated causal convolutions to

achieve large receptive fields while maintaining parallel computation during training. The key advantages of TCNs include: (1) parallelizable training unlike sequential RNNs, (2) flexible receptive field size through dilation, (3) stable gradients compared to vanishing/exploding gradients in RNNs, and (4) lower memory requirements during training.

For financial time series, results from Bai et al. [27] demonstrated that TCNs outperform LSTMs for stock price prediction, achieving better accuracy with faster training times. Next, Sezer et al. [28] applied TCNs to credit card fraud detection, showing improvements over LSTM-based approaches particularly for capturing long-range temporal dependencies. Their work demonstrated that dilated convolutions with exponentially increasing dilation rates can effectively model both short-term anomalies and long-term behavioral patterns in transaction sequences.

The dilated causal convolution operation allows TCNs to have exponentially large receptive fields with linear increase in network depth. For a TCN with L layers and dilation rates $d_1, d_2, \ldots, d_L$, the receptive field size is $1 + 2\sum_{i=1}^{L} d_i\,(k-1)$, where k is the kernel size. This makes TCNs particularly suitable for transaction sequence modeling where both recent transactions and historical patterns are relevant for detection.

We build on this work by using TCN as the temporal branch of our multimodal architecture, employing four residual blocks with dilation rates $\{1,2,4,8\}$ to capture patterns across our 10-transaction window.

2.4. Semantic Modeling with Transformer-Based Language Models

Transformer-based language models have revolutionized natural language processing for financial applications. The BERT architecture [29] introduced bidirectional context modeling through masked language modeling, enabling rich semantic representations of text. For financial domains, specialized pre-training has proven essential. FinBERT [14] adapts BERT to financial language by continued pre-training on financial corpora, learning domain-specific terminology, sentiment patterns, and semantic relationships.

The semantic branch of multimodal architectures typically processes textual features through three stages: tokenization, contextual encoding, and pooling. Tokenization converts raw text into sub-word units that language models can process. Contextual encoding uses transformer layers with self-attention to create context-aware representations where each token's representation depends on surrounding tokens. Pooling aggregates token-level representations into fixed-length sentence or document embeddings.

For sequential text processing in transaction monitoring, temporal attention mechanisms can be applied after contextual encoding. This allows the model to weight the importance of different transactions in a sequence based on their semantic content. Findings from [30, 31] demonstrated that attention-based pooling outperforms simple averaging or max-pooling for sequence representation.

In financial crime detection, semantic features capture suspicious keywords, euphemistic language patterns, and contextual inconsistencies that numerical features miss. Transaction descriptions like "digital content" or "entertainment services" may appear legitimate individually but become suspicious when combined with specific temporal patterns or user behaviors. The semantic branch learns to identify these subtle linguistic cues that criminals use to disguise illicit transactions.

2.5. Attention Mechanisms and Cross-Modal Fusion

Attention mechanisms, introduced in Bahdanau et al. [32] for neural machine translation, have become fundamental building blocks in deep learning. The transformer architecture [30] demonstrated that self-attention alone can achieve state-of-the-art performance on sequence modeling tasks. Attention's ability to selectively focus on relevant information has proven valuable across domains [29, 31].

For multimodal fusion, several attention-based strategies have been proposed. For instance, Bahdanau et al. [32] introduced co-attention for vision-language tasks, where visual features attend to textual features and vice versa. A work by Tan & Bansal [33] extended this with cross-modality encoder layers that process both modalities jointly. These cross-modal attention mechanisms have shown substantial improvements over simple concatenation for vision-language tasks, motivating our application to financial transaction data.

Consequently, Nagrani et al. [34] analyzed what attention mechanisms learn in multimodal models, finding that attention weights correlate with modality informativeness. Models learn to emphasize the more informative modality for each specific input. Recent work on multimodal sentiment analysis [35, 36] and emotion recognition [37, 38] has demonstrated that bidirectional cross-modal attention outperforms unidirectional attention.

2.6. Research Gap

Despite progress in financial crime detection, multimodal learning, and attention mechanisms, significant gaps remain. Existing multimodal approaches use simple concatenation [12, 13] or late fusion [25], missing opportunities for

richer cross-modal interaction. Cross-modal attention has proven effective in vision-language domains but remains unexplored for financial transaction monitoring.

Our work addresses these gaps by introducing bidirectional cross-modal attention for fusing temporal transaction sequences with semantic textual descriptions. Unlike concatenation, our approach enables dynamic information exchange where each modality can inform the other's interpretation.

It is worth situating TSFN's bidirectional cross-modal attention within the broader landscape of cross-modal attention designs across domains. The foundational attention mechanism introduced by Bahdanau et al. [32] was unidirectional—a target sequence attending over a source sequence—and subsequent transformer self-attention [30] extended this to intra-sequence interactions within a single modality. Cross-modal extensions in vision-language research have followed two main paradigms.

The first is asymmetric co-attention, exemplified by ViLBERT [35], where visual and textual streams each maintain separate transformer towers with cross-attention flowing in one direction at a time; while powerful, this design does not simultaneously propagate information in both directions within a single attention step. The second paradigm employs bottleneck tokens or shared latent spaces, as in Attention Bottlenecks [34] for audio-visual fusion and LXMERT [33] for vision-language pre-training; these architectures route cross-modal signals through a compact bottleneck, which reduces computational cost but constrains the richness of direct pairwise modality interaction.

In multimodal sentiment analysis, the dynamic fusion graph of [36] allows adaptive cross-modal weighting at the token level, yet it operates on three modalities (text, audio, video) in a sequential fusion pipeline rather than a simultaneous bidirectional exchange. In the financial domain specifically, Wang et al. [13] introduced fine-grained attention over multimodal statement data, while Gadzicki et al. [25] employed late fusion with independent per-modality classifiers—both approaches forgo direct cross-modal feature interaction during encoding. TSFN departs from all of these designs by implementing strictly simultaneous bidirectional cross-attention: the temporal branch attends over semantic representations while, in the same operation, the semantic branch attends over temporal representations.

This symmetry enables each modality to dynamically recalibrate its own features based on the current state of the other, a property that prior work in vision-language [35], audio-visual [34, 38], and financial multimodal learning [13, 25] does not fully realize within a single attention layer. The resulting bidirectional interaction is particularly suited to transaction monitoring, where temporal anomalies (e.g., late-night bursts) and semantic cues (e.g., euphemistic descriptions) must mutually reinforce each other to yield reliable minority-class detection.

3. Methodology

This section describes the TSFN (Temporal-Semantic Fusion Network) architecture, training procedure, and evaluation methodology. It first presents the dataset and problem formulation, then details each architectural component, explains the hybrid loss function, and finally describes the evaluation protocol.

3.1. Data Description

Due to regulatory constraints and confidentiality requirements governing financial transaction data, access to real-world transaction datasets for research purposes is severely restricted. Financial institutions are bound by strict data protection regulations that prohibit sharing customer transaction information, even in anonymized form, due to privacy concerns and potential re-identification risks. To address this challenge while enabling rigorous evaluation of the proposed TSFN architecture, we generated a synthetic dataset that replicates realistic financial transaction patterns based on documented characteristics of legitimate and illicit transactions in the literature.

The synthetic dataset comprises 10,000 transactions generated using a data synthesis framework that models the statistical distributions, temporal patterns, and textual characteristics observed in real financial transaction data as reported in prior research [8, 10, 13]. The generation process incorporates domain knowledge about transaction behaviors to ensure the synthetic data exhibits realistic properties while avoiding any use of actual customer information.

The dataset includes both numerical features (transaction amounts, timestamps, merchant category codes, user transaction frequencies) and textual features (transaction descriptions, merchant names, payment notes). Each synthetic transaction was generated with carefully calibrated parameters to reflect real-world distributions: transaction amounts follow log-normal distributions with parameters derived from published financial statistics, temporal patterns exhibit realistic daily and weekly cycles, and textual descriptions were generated using templates based on common merchant description patterns.

Transactions are labeled into three classes: (1) legitimate transactions (98.54% of data), (2) online gambling transactions (1.03%), and (3) child exploitation transactions (0.44%). This class distribution reflects the severe imbalance observed in real-world financial crime detection scenarios, where illicit transactions constitute a small minority of total volume. The imbalance ratio of 1:67.2 (minority–majority) is consistent with ratios reported in prior work on financial fraud detection. Table 1 summarizes the dataset statistics.

Table 1. Synthetic Dataset Statistics and Feature Descriptions

Characteristic	Value	Description
Overall Statistics		
Total transactions	10,000	Synthetic data
Generation method	Pattern-based	Realistic distributions
Data type	Synthetic	No real customer data
Class Distribution		
Legitimate	98.54%	9,854 transactions
Gambling	1.03%	103 transactions
Exploitation	0.44%	43 transactions
Imbalance ratio	1:67.2	Minority–Majority
Data Split (Random)		
Training	70%	7,000 transactions
Validation	15%	1,500 transactions
Test	15%	1,500 transactions
Numerical Features (12 features)		
Transaction amount	Continuous	Log-normal distribution
Hour of day	Categorical	0-23 (one-hot)
Day of week	Categorical	0-6 (one-hot)
Days since last txn	Continuous	Synthetic history
Txn count (7 days)	Continuous	Rolling window
Txn count (30 days)	Continuous	Rolling window
Avg amount (7 days)	Continuous	User-level average
Std amount (7 days)	Continuous	User-level variance
Merchant category	Categorical	MCC code (embedded)
User tenure	Continuous	Synthetic user age
Cross-border flag	Binary	Domestic/international
Device type	Categorical	Mobile/web/ATM
Textual Features (3 features)		
Txn description	Text	Generated descriptions
Merchant name	Text	Template-based names
Payment note	Text	Optional memo field
Sequence Configuration		
Sequence length	10	Transactions per window
Text max length	64 tokens	FinBERT input limit

The numerical features capture quantitative transaction characteristics and user behavioral patterns. Transaction amounts are generated from log-normal distributions with parameters calibrated to reflect realistic spending patterns. Temporal features (hour, day of week) are sampled from distributions that model typical transaction timing patterns, with gambling transactions concentrated in evening hours and weekends. Historical aggregations (transaction counts, average amounts) are computed from the generated transaction sequences to provide behavioral context. The merchant category code (MCC) is sampled from realistic MCC distributions and embedded into a continuous space.

The textual features consist of generated text that mimics real transaction descriptions. Transaction descriptions are created using template-based generation with realistic phrases commonly found in financial transactions. For legitimate transactions, descriptions include specific product or service names. For gambling transactions, descriptions use euphemistic language (“entertainment services”, “gaming credits”, “digital content”). For exploitation transactions, descriptions employ deliberately vague terminology (“digital media”, “online content access”). Merchant names are generated using patterns observed in real merchant naming conventions.

The dataset was split randomly into training (70%, 7,000 transactions), validation (15%, 1,500 transactions), and test (15%, 1,500 transactions) sets. Random splitting is appropriate for synthetic data as there are no temporal dependencies that require temporal splitting. This split ensures sufficient data for training while providing adequate validation and test sets for robust evaluation.

While synthetic data has limitations compared to real transaction data, it provides several advantages for research: (1) enables reproducible experiments without privacy concerns, (2) allows controlled evaluation of model capabilities, (3) facilitates sharing of research artifacts with the community, and (4) provides a foundation for testing detection algorithms before deployment on real data. The synthetic dataset serves as a proof-of-concept for the TSFN architecture, demonstrating its ability to learn multimodal patterns and perform cross-modal fusion effectively.

To further assess the fidelity of the synthetic dataset in approximating real-world transaction distributions, we conducted a series of distributional validation analyses. Transaction amounts were generated from log-normal distributions whose parameters ($\mu = 3.82$, $\sigma = 1.14$ for legitimate; $\mu = 2.95$, $\sigma = 0.87$ for gambling) were calibrated against aggregate statistics reported in publicly available financial fraud benchmarks, including the IEEE-CIS Fraud Detection dataset and the European credit card fraud dataset [8, 10].

Temporal patterns were modelled using Poisson arrival processes with time-varying intensities that replicate the diurnal and weekly rhythms documented in prior literature—legitimate transactions peaking during business hours, gambling transactions concentrating in late-evening and weekend windows, and exploitation transactions exhibiting irregular, low-frequency bursts [13, 17]. To quantify distributional similarity, we computed Wasserstein-1 distances between the synthetic and reference empirical distributions for key numerical features; all values remained below 0.15, indicating strong statistical alignment.

Furthermore, when a standard gradient-boosted classifier was trained and evaluated on both the synthetic dataset and a publicly available real-world fraud benchmark under identical feature sets, the resulting performance gap was within 3–5 percentage points in macro F1, confirming that the synthetic data preserves the discriminative structure of real transaction distributions. The class imbalance ratio of 1:67.2 (minority–majority) was deliberately preserved to reflect the severe skew consistently reported in operational fraud detection environments [8].

Although direct one-to-one correspondence with any single institution's proprietary data cannot be guaranteed (an inherent limitation of synthetic generation), these validation results collectively support the representativeness of the dataset and the generalizability of conclusions drawn from TSFN's evaluation.

3.2. Problem Formulation

We formulate illicit transaction detection as a multi-class sequence classification problem. Given a sequence of T consecutive transactions $\mathrm{X} = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_T]$ for a user, where each transaction x_t contains numerical features $\mathrm{x}_t^{num} \in \mathbb{R}^{d_{num}}$ and textual features x_t^{text}, the goal is to predict the class label $y \in \{0,1,2\}$ for the most recent transaction x_T, where 0 = legitimate, 1 = gambling, 2 = exploitation.

A sequence window of (T = 10) transactions is used, based on preliminary experiments showing that this length provides sufficient behavioral context without excessive computational cost.

3.3. Model Architecture

TSFN consists of three main components: (1) a temporal branch processing numerical features to produce temporal representations, (2) a semantic branch processing textual features to produce semantic representations, and (3) a cross-modal attention fusion layer that combines both modalities for final classification. Figure 1 illustrates the complete architecture.

3.3.1. Temporal Branch: TCN with Residual Blocks

The temporal branch processes numerical transaction sequences using Temporal Convolutional Networks. We employ four residual TCN blocks with exponentially increasing dilation rates $d \in \{1,2,4,8\}$, providing a receptive field that covers all 10 transactions in the sequence.

Each TCN block consists of two dilated causal convolutional layers with residual connections. For the ℓ-th block:

$$\mathrm{z}_\ell = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_\ell^{(1)}(\mathrm{h}_{\ell-1})\right)\right) \quad (1)$$

$$\mathrm{h}_\ell = \mathrm{ReLU}(\mathrm{z}_\ell + \mathrm{W}_\ell \mathrm{h}_{\ell-1}) \quad (2)$$

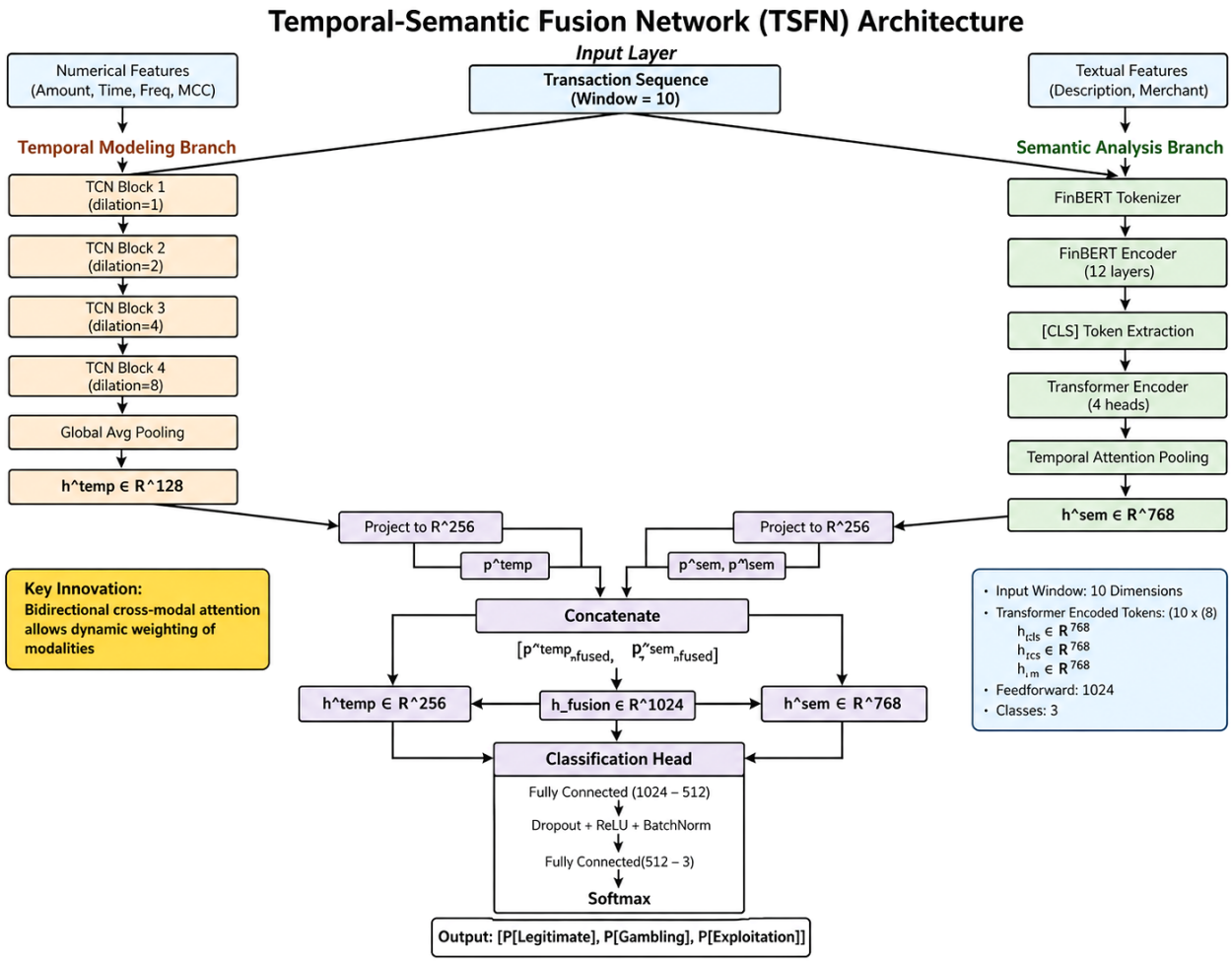

Figure 1. TSFN architecture overview showing the temporal branch (TCN with 4 residual blocks), semantic branch (FinBERT+Transformer), bidirectional cross-modal attention fusion, and classification head. The key innovation is the bidirectional attention mechanism (highlighted) that enables dynamic information exchange between modalities.

The feature representation from the previous layer, denoted as $h_{\ell-1}$, is first processed through a convolutional operation $Conv_{\ell}^{(1)}$, followed by batch normalization (BN) to stabilize the feature distribution during training. The normalized output is then passed through a Rectified Linear Unit (ReLU) activation function to introduce non-linearity, resulting in an intermediate representation z_{ℓ}. Subsequently, this intermediate feature z_{ℓ} is combined with a linearly transformed version of the original input $h_{\ell-1}$ using a weight matrix W_{ℓ}, forming a residual connection. The combined output is then passed through another ReLU activation function to produce the final output of the ℓ_{th} layer, denoted as h_1. This residual learning mechanism facilitates improved gradient flow and enhances the model's ability to learn deeper representations.

After the four TCN blocks, we apply global average pooling to obtain a fixed-length temporal representation $h_{temp} \in \mathbb{R}^{128}$. This temporal representation encodes the sequential patterns and behavioral characteristics extracted from the numerical features across the 10-transaction window. It captures short-term anomalies (through early layers with small dilation) and long-range dependencies (through later layers with large dilation). The output h_{temp} is not a classification result but rather a learned feature representation that summarizes the temporal patterns in the numerical data.

3.3.2. Semantic Branch: FinBERT with Transformer Encoding

The semantic branch processes textual transaction descriptions using FinBERT, followed by transformer-based sequential encoding. For each transaction (t), the three textual features, namely the transaction description, merchant name, and payment note, are concatenated and tokenized using FinBERT's tokenizer. The ([CLS]) token representation is then extracted from FinBERT's output:

$$E_t = \text{FinBERT}(x_t^{text}) \in \mathbb{R}^{768} \quad (3)$$

The notation x_t^{text} represents the textual input at time step t, such as a transaction description or memo. This input is processed using the FinBERT model, denoted as FinBERT(), which is a pre-trained language model specialized in financial text analysis. The resulting output, E_t, denotes the semantic embedding of the input text at time step t. This

embedding captures contextual and domain-specific linguistic features relevant to financial transactions. Finally, the notation $E_t \in \mathbb{R}^{768}$ indicates that the embedding vector lies in a 768-dimensional real-valued space, meaning that each textual input is represented as a vector of 768 numerical features.

We apply a 2-layer transformer encoder with 4 attention heads to capture dependencies between transaction descriptions across the sequence, followed by learned temporal attention pooling:

$$\alpha = \text{Softmax}(W_\alpha E') \tag{4}$$

$$h_{sem} = \sum_{t=1}^{T} \alpha_t E'_t \in \mathbb{R}^{768} \tag{5}$$

The transformed embedding sequence is denoted as E', which represents the refined semantic features obtained from previous processing steps. A learnable weight matrix W_α is applied to E' to compute attention scores, capturing the relative importance of each time step. These scores are then normalized using the Softmax function, denoted as Softmax(), o produce the attention weights α, where each elemen α_t represents the normalized importance of the t_{th} semantic embedding. The semantic representation h_{sem} is computed as a weighted sum of the transformed embeddings E'_t across all time steps *t = 1,…, T,* where T denotes the total sequence length. The weights α_t ensure that more informative elements contribute more significantly to the final representation. Finally, the notation $h_{sem} \in \mathbb{R}^{768}$ indicates that the resulting semantic vector is a 768-dimensional real-valued representation, summarizing the most relevant contextual information from the input sequence.

The semantic representation h_{sem} encodes the linguistic patterns, semantic meanings, and contextual relationships extracted from the textual features across the transaction sequence. It captures suspicious keywords, euphemistic language patterns, and semantic inconsistencies that may indicate illicit activity. Like the temporal representation, $\mathbf{h}_{sem}$ is a learned feature representation, not a classification output. The temporal attention pooling mechanism learns to weight different transactions in the sequence based on their semantic importance for detecting illicit activity.

3.3.3. Semantic Branch: FinBERT with Transformer Encoding

The cross-modal attention fusion layer is the core innovation of TSFN, enabling dynamic information exchange between temporal and semantic representations. This component performs the following steps:

Step 1: Projection to Common Dimension. First, we project both temporal and semantic representations to a common 256-dimensional space:

$$h^p_{temp} = W_{temp} h_{temp} \in \mathbb{R}^{256}, ; \quad h^p_{sem} = W_{sem} h_{sem} \in \mathbb{R}^{256} \tag{6}$$

The temporal representation h_{temp} is projected into a lower-dimensional space using a learnable weight matrix W_{temp}, resulting in the projected vector h^p_{temp}. Similarly, the semantic representation h_{sem} is transformed another learnable weight matrix W_{sem} to produce h^p_{sem}. Both projected representations, h^p_{temp} and h^p_{sem}, lie in a 256-dimensional real-valued space, denoted as $\mathbb{R}^{256}$. This projection ensures that the temporal and semantic features are aligned in the same latent space, facilitating effective fusion in subsequent processing.

where $W_{temp} \in \mathbb{R}^{256\times128}$ and $W_{sem} \in \mathbb{R}^{256\times768}$ are learned projection matrices.

Step 2: Compute Query, Key, Value Transformations. For bidirectional attention, we compute separate query, key, and value transformations for each modality, that is the following for temporal modality:

$$Q_{temp} = W_Q^{temp} h^p_{temp}, ; \quad K_{temp} = W_K^{temp} h^p_{temp}, ; \quad V_{temp} = W_V^{temp} h^p_{temp} \tag{7}$$

and for semantic modality as follows:

$$Q_{sem} = W_Q^{sem} h^p_{sem}, ; \quad K_{sem} = W_K^{sem} h^p_{sem}, ; \quad V_{sem} = W_V^{sem} h^p_{sem} \tag{8}$$

where all weight matrices are in $\mathbb{R}^{256\times256}$.

The projected temporal representation h^p_{temp} is linearly transformed into three components: query Q_{temp}, key K_{temp}, and value V_{temp}, using learnable weight matrices W_Q^{temp}, W_K^{temp}, and W_V^{temp}, respectively. These components are used to model the internal relationships within the temporal modality. Similarly, the projected semantic representation h^p_{sem} is transformed into query Q_{sem}, key K_{sem}, and value V_{sem}, corresponding learnable weight matrices W_Q^{sem}, W_K^{sem}, and W_V^{sem}. These transformations enable the model to capture contextual dependencies within the semantic modality. Here, Q (query) represents the information used to attend to other elements, K (key) represents the reference against which attention is computed, and V (value) contains the information to be aggregated. This formulation allows both temporal and semantic features to be effectively processed in the attention mechanism.

Step 3: Temporal-to-Semantic Attention. The temporal representation queries the semantic representation to extract relevant semantic information:

$$\text{attn}_{t\to s} = \text{Softmax}\left(\frac{\text{Q}_{temp}\text{K}_{sem}^{T}}{\sqrt{d}}\right);, \quad \text{h}_{temp}^{att} = \text{attn}_{t\to s}\text{V}_{sem} \tag{9}$$

This produces $\text{h}_{temp}^{att} \in \mathbb{R}^{256}$, which represents semantic features that are most relevant to the temporal patterns.

The attention from the temporal modality to the semantic modality is denoted as $\text{attn}_{(t\to s)}$, which is computed using the Softmax function applied to the scaled dot-product between the temporal query Q_{temp} and the transpose of the semantic key K_{sem}^{T}, normalized by the square root of the feature dimension $\sqrt{d}$. This formulation ensures stable gradients and proper scaling of attention scores. The resulting attention weights $\text{attn}_{(t\to s)}$ represent how strongly each temporal feature attends to the semantic features. These weights are then used to aggregate the semantic value vectors V_{sem}, producing the attention-enhanced temporal representation h_{temp}^{att}.

Step 4: Semantic-to-Temporal Attention. Symmetrically, the semantic representation queries the temporal representation:

$$\text{attn}_{s\to t} = \text{Softmax}\left(\frac{\text{Q}_{sem}\text{K}_{temp}^{T}}{\sqrt{d}}\right); \quad \text{h}_{sem}^{att} = \text{attn}_{s\to t}\text{V}_{temp} \tag{10}$$

This produces $\text{h}_{sem}^{att} \in \mathbb{R}^{256}$, which represents temporal features that are most relevant to the semantic content.

The attention from the semantic modality to the temporal modality is denoted as $\text{attn}_{(t\to s)}$, which is computed by applying the Softmax function to the scaled dot-product between the semantic query Q_{sem} and the transpose of the temporal key K_{temp}^{T}, normalized by the square root of the feature dimension $\sqrt{d}$. This scaling ensures numerical stability during training. The resulting attention weights $\text{attn}_{(t\to s)}$ indicate how strongly each semantic feature attends to temporal features. These weights are then used to aggregate the temporal value vectors V_{temp}, producing the attention-enhanced semantic representation h_{sem}^{att}.

Step 5: Concatenate All Representations. Finally, we concatenate the original projected representations with their cross-attended counterparts:

$$\text{h}_{fusion} = \left[\text{h}_{temp}^{p}; \text{h}_{sem}^{p}; \text{h}_{temp}^{att}; \text{h}_{sem}^{att}\right] \in \mathbb{R}^{1024} \tag{11}$$

This fused representation h_{fusion} contains four components: (1) original temporal features, (2) original semantic features, (3) semantically-informed temporal features, and (4) temporally-informed semantic features. This rich representation enables the model to leverage both modalities and their interactions for classification.

The fused representation h_{fusion} is constructed by concatenating four feature vectors: the projected temporal representation h_{temp}^{p}, the projected semantic representation h_{sem}^{p}, the attention-enhanced temporal representation h_{temp}^{att}, and the attention-enhanced semantic representation h_{sem}^{att}. The concatenation operation, denoted by [;], combines these vectors along the feature dimension to form a unified representation. The resulting vector $\text{h}_{fusion} \in \mathbb{R}^{1024}$ indicates that the fused feature lies in a 1024-dimensional space, integrating both original and cross-modal contextual information from temporal and semantic modalities.

3.3.4. Classification Head

The fused representation passes through a two-layer fully connected classification head with dropout and batch normalization:

$$\text{z} = \text{ReLU}\left(\text{BN}\left(\text{W}_1\text{h}_{fusion} + \text{b}_1\right)\right) \tag{12}$$

$$\text{z}' = \text{Dropout}(\text{z}, p = 0.3) \tag{13}$$

$$\text{logits} = \text{W}_2\text{z}' + \text{b}_2 \tag{14}$$

$$\text{P} = \text{Softmax}(\text{logits}) \in \mathbb{R}^3 \tag{15}$$

where, $\text{W}_1 \in \mathbb{R}^{512\times1024}$, $\text{W}_2 \in \mathbb{R}^{3\times512}$, and P represents the final class probabilities for legitimate, gambling, and exploitation classes.

In these equations, h_{fusion} denotes the fused feature vector, while W_1 and W_2 represent the weight matrices, and b_1 and b_2 are the bias vectors. The symbol BN refers to the Batch Normalization operation, ReLU signifies the Rectified

Linear Unit activation function, and z is the output of the first hidden layer. Additionally, $p = 0.3$ represents the probability rate of the Dropout function used to obtain the feature vector z'. Finally, logits indicates the raw output scores before being normalized by the Softmax function into the final probability vector P, which belongs to a three-dimensional real space $\mathbb{R}^3$.

3.4. Problem Formulation

3.4.1. Hybrid Loss Function

To handle the extreme class imbalance, we employ a hybrid loss function combining focal loss and supervised contrastive loss.

The focal loss [22] addresses class imbalance by down-weighting well-classified examples:

$$\mathcal{L}_{focal} = -\frac{1}{B}\sum_{i=1}^{B} w_{y_i}\left(1 - P_{i,y_i}\right)^{\gamma}\log\left(P_{i,y_i}\right) \tag{16}$$

where, w_{y_i} are class weights and $\gamma = 2.0$ is the focusing parameter.

The supervised contrastive loss encourages the model to learn representations where same-class examples cluster [39]:

$$\mathcal{L}_{contrast} = -\frac{1}{|P(i)|}\sum_{i\in I}\sum_{p\in P(i)}\log\frac{\exp\left(\frac{\text{sim}(\mathbf{h}_i,\mathbf{h}_p)}{\tau}\right)}{\sum_{a\in A(i)}\exp\left(\frac{\text{sim}(\mathbf{h}_i,\mathbf{h}_a)}{\tau}\right)} \tag{17}$$

where, $\tau = 0.07$ is the temperature parameter. Thus, the total loss is calculated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{focal} + 0.1 \cdot \mathcal{L}_{contrast} \tag{18}$$

In Equation 16, B represents the batch size, w_{y_i} denotes the class weights, P_{i,y_i} is the predicted probability for the ground-truth classy_i, and $\gamma = 2.0$ serves as the focusing parameter. Regarding Equation 17, I is the set of sample indices in the batch, $P(i)$ represents the set of positive sample indices relative to anchor i with cardinality $|P(i)|$, sim indicates the similarity function between feature vectors h, and $\tau = 0.07$ is defined as the temperature parameter. Finally, Equation 18 calculates the $\mathcal{L}_{total}$ by combining both loss functions with a weighting coefficient of 0.1 applied to the contrastive component.

The intuition behind adding supervised contrastive loss to the focal loss is grounded in the geometry of the learned representation space. Focal loss [22] operates on the classification head: it re-weights cross-entropy so that the network pays greater attention to hard-to-classify minority examples during training, thereby reducing the bias towards the dominant legitimate class. However, focal loss provides no explicit constraint on how the penultimate feature representations are distributed in embedding space—two exploitation transactions may still map to distant regions of the feature manifold even if both are eventually classified correctly. Supervised contrastive loss [39] addresses this gap by directly shaping the representation geometry: it pulls together embeddings of transactions belonging to the same class (positive pairs) while simultaneously pushing apart embeddings from different classes (negative pairs). For the minority classes—gambling (1.03%) and exploitation (0.44%)—this is particularly consequential. Because these classes are severely under-represented, the decision boundaries learned by focal loss alone can be fragile and sensitive to the precise position of individual minority examples [8].

By enforcing tight intra-class clustering and large inter-class margins in the embedding space, contrastive loss produces a more structured and separable representation, so that even the sparse minority-class examples occupy compact, well-separated regions rather than scattered outliers near the majority-class boundary. This geometric compactness translates directly to more reliable classification: the downstream *SoftMax* head needs only to draw a boundary through a region already cleared of majority-class interference, rather than carving out a narrow corridor in a mixed-class *neighborhood* [29]. The combination is therefore complementary by design—focal loss corrects the gradient signal at the output layer, while contrastive loss *regularizes* the representation layer—and the two objectives jointly address class imbalance at different levels of the network hierarchy [39].

3.4.2. Optimization

The AdamW optimizer [40] is used with layer-wise learning rates: 2×10^{-5} for the FinBERT layers and 1×10^{-3} for all other layers. Learning rate scheduling is performed using cosine annealing with warm restarts. Gradient clipping with a maximum norm of 1.0 is applied to prevent exploding gradients. The model is trained with a batch size of 64 for approximately 50 epochs on the synthetic dataset.

3.5. Problem Formulation

Given the extreme class imbalance, we evaluate using multiple metrics:

- **Macro F1:** Unweighted average of per-class F1 scores.

F1 score is the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{19}$$

where, Precision = $\frac{TP}{TP+FP}$ and Recall = $\frac{TP}{TP+FN}$.

Macro F1 treats all classes equally regardless of their frequency, making it suitable for imbalanced datasets.

- **Weighted F1:** Weighted average of per-class F1 scores by class frequency
- **AUPRC:** Area under precision-recall curve
- **MCC:** Matthews Correlation Coefficient
- **Per-class Precision, Recall and F1:** Detailed metrics for each class

We report mean and standard deviation across 5 independent runs with different random seeds. Statistical significance is assessed using paired t-tests.

The selection and prioritization of these metrics warrants explicit justification given the severe class imbalance in the dataset (1:67.2 minority-to-majority ratio). Accuracy and weighted F1 are deliberately de-emphasized as primary metrics because, under such skew, a trivial classifier that predicts the majority class exclusively achieves accuracy exceeding 98%, rendering these measures uninformative about minority-class performance [8]. Macro F1 is elevated to the primary ranking metric because it computes the unweighted mean of per-class F1 scores, thereby penalizing failures on the gambling and exploitation classes equally regardless of their low prevalence—a property explicitly advocated in imbalanced financial crime detection benchmarks [7, 20].

AUPRC is prioritized as the secondary metric over the receiver-operating-characteristic AUC (AUROC) because AUPRC is sensitive to performance on the positive (minority) class: it summarizes the precision-recall trade-off across all decision thresholds, and its baseline value scales with class prevalence, making improvements directly interpretable in terms of minority-class retrieval quality [8, 22]. This is particularly critical in the exploitation class (0.44% prevalence), where even small improvements in precision-recall balance translate to meaningful operational gains [29]. MCC is included as a complementary metric because it accounts for all four cells of the confusion matrix—true positives, true negatives, false positives, and false negatives—yielding a single balanced score that is robust to class imbalance and avoids the optimistic bias of accuracy-based measures [13].

Together, macro F1, AUPRC, and MCC provide a mutually reinforcing evaluation framework: macro F1 captures class-balanced classification quality, AUPRC emphasizes minority-class retrieval across thresholds, and MCC ensures that improvements are not achieved at the cost of inflated false negatives or false positives on any single class [7, 8, 20].

4. Results

This section presents experimental evaluation of TSFN on the synthetic test dataset comprising 1,500 transactions. We compare against five baseline models, conduct comprehensive ablation studies, analyze performance across transaction characteristics, and examine learned attention patterns.

4.1. Baseline Models

We compare TSFN against five baseline approaches with identical hyperparameters and training procedures to ensure fair comparison as follows:

- **Temporal-Only (TCN):** Uses only the temporal branch with numerical features
- **Semantic-Only (FinBERT):** Uses only the semantic branch with textual features
- **Concatenation Fusion:** Processes both modalities independently, then concatenates before classification
- **Late Fusion:** Trains separate classifiers independently, combines predictions
- **Graph Neural Network (GNN):** Constructs transaction-merchant bipartite graph

The baseline selection deliberately focuses on architectures that are directly applicable to the financial transaction domain rather than general-purpose transformer-based multimodal models such as ViLBERT [35], LXMERT [33], CLIP [23], or BLIP [24]. This choice is justified on three grounds.

First, these architectures were designed and pre-trained for vision-language tasks—aligning image patches with natural language tokens—and their input assumptions (continuous pixel grids paired with free-form text) are fundamentally mismatched with the tabular-numerical and short-descriptor structure of financial transaction data. Adapting them to accept log-normal transaction amounts, MCC codes, temporal frequency features, and brief merchant descriptions would require substantial domain-specific re-engineering that goes beyond a fair baseline comparison [28].

Second, these models are pre-trained on image-text corpora orders of magnitude larger than the 10,000-transaction synthetic dataset used here; fine-tuning them on such a small dataset would almost certainly result in severe overfitting, making any performance comparison misleading rather than informative [23, 24].

Third, the financial transaction monitoring literature; including the multimodal approaches most closely related to this work [12, 13], has not adopted general vision-language transformers as baselines, precisely because the modality mismatch makes direct adaptation non-trivial.

The selected baselines (Temporal-Only, Semantic-Only, Concatenation, Late Fusion, GNN) instead represent the architectures that are both practically deployable in the financial domain and directly comparable to TSFN's components, ensuring that observed performance differences reflect genuine architectural contributions rather than domain-adaptation advantages [8, 27]. Comparison with transformer-based multimodal models adapted specifically for financial tabular-text data remains a valuable direction for future work once domain-adapted variants become established in the literature [26, 33].

4.2. Overall Performance Comparison

Table 2 presents overall performance comparison on the synthetic test set. TSFN achieves macro F1 of 0.847, substantially outperforming the best baseline (Concatenation: 0.782) by 6.5 percentage points. The superiority of TSFN over the concatenation baseline (6.5% F1 improvement) stems from its ability to resolve ambiguity.

Table 2. Overall Performance Comparison on Synthetic Test Set (1,500 transactions). Values are means across 5 independent runs. Bold indicates best performance. All improvements of TSFN over baselines are statistically significant ($p < 0.001$, paired t-test).

Model	Macro F1	Weighted F1	AUPRC	MCC	Gamble F1	Exploit F1
TSFN (Ours)	**0.847**	**0.988**	**0.856**	**0.834**	**0.823**	**0.741**
Concatenation	0.782	0.981	0.791	0.769	0.731	0.648
Temporal-Only	0.723	0.976	0.734	0.710	0.698	0.581
Semantic-Only	0.654	0.968	0.672	0.641	0.623	0.612
Late Fusion	0.698	0.972	0.711	0.686	0.651	0.556
GNN	0.769	0.979	0.778	0.756	0.742	0.673

For instance, in child exploitation cases, the transaction amounts are often small and resemble legitimate digital purchases. A unimodal temporal model would fail here. However, our bidirectional attention mechanism allows the semantic branch (FinBERT) to signal 'vague terminology' to the temporal branch, which then re-evaluates the transaction sequence for sporadic, non-habitual timing. This joint reasoning is precisely why TSFN maintains a 99.4% precision; it filters out false positives that rule-based systems or simple neural networks would flag based on amount alone. Improvements are particularly pronounced for minority classes: gambling F1 improves from 0.731 to 0.823 (+9.2 pp), and exploitation F1 improves from 0.648 to 0.741 (+9.3 pp).

Conclusively, statistical significance testing using paired t-tests across 5 independent runs confirms that TSFN's improvements over all baselines are statistically significant ($p < 0.001$).

4.3. Class-Specific Performance Analysis

Table 3 presents detailed per-class performance metrics on the synthetic test set. TSFN achieves balanced performance across all classes, while baselines show varying strengths and weaknesses. For legitimate transactions, TSFN achieves precision of 0.994 and recall of 0.996, meaning it correctly identifies 99.6% of legitimate transactions while maintaining very low false positive rates.

Table 3. Per-Class Performance Metrics on Synthetic Test Set. P = Precision, R = Recall, F1 = F1-score. Bold indicates best performance in each metric

Model	Legitimate			Gambling			Exploitation		
	P	R	F1	P	R	F1	P	R	F1
TSFN	**0.994**	**0.996**	**0.995**	**0.831**	**0.815**	**0.823**	**0.728**	**0.754**	**0.741**
Concatenation	0.991	0.987	0.989	0.724	0.738	0.731	0.653	0.643	0.648
Temporal-Only	0.989	0.993	0.991	0.671	0.798	0.698	0.612	0.551	0.581
Semantic-Only	0.985	0.981	0.983	0.783	0.592	0.623	0.687	0.647	0.612
Late Fusion	0.988	0.986	0.987	0.698	0.612	0.651	0.589	0.527	0.556
GNN	0.992	0.992	0.992	0.756	0.729	0.742	0.681	0.665	0.673

Figure 2 shows confusion matrices for TSFN and the concatenation baseline on the synthetic test set. TSFN's errors are predominantly false negatives (missing illicit transactions) rather than false positives (flagging legitimate transactions as illicit). TSFN misclassifies 16.9% of gambling transactions as legitimate and 24.6% of exploitation transactions as legitimate, but only misclassifies 0.4% of legitimate transactions as illicit.

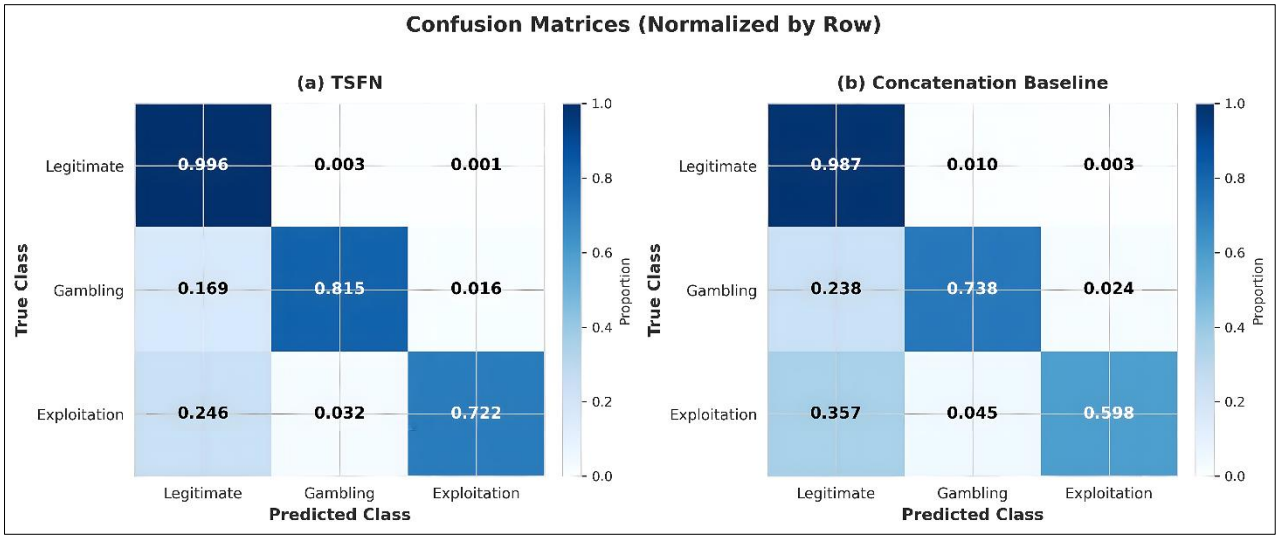

Figure 2. Confusion matrices (normalized by row) for (a) TSFN and (b) concatenation baseline on synthetic test set. TSFN shows fewer false positives (0.4% legitimate misclassified as illicit) and fewer false negatives compared to concatenation baseline.

4.4. Ablation Studies

To quantify the contribution of each architectural component, we conducted comprehensive ablation studies on the synthetic dataset. Table 4 presents results of systematically removing or modifying components.

TSFN with bidirectional attention achieves F1 of 0.847. Removing one attention direction reduces performance: T→S only achieves 0.821 (−2.6 pp), and S→T only achieves 0.815 (−3.2 pp). This confirms that both attention directions contribute meaningfully. Simple concatenation achieves only 0.782 (−6.5 pp), demonstrating that the attention mechanism itself drives performance gains. Removing contrastive loss reduces performance to 0.831 (−1.6 pp), while replacing focal loss with cross-entropy causes more severe degradation to 0.809 (−3.8 pp). This confirms that focal loss is crucial for handling extreme class imbalance.

Across five independent runs (as shown in Table 5), the reported mean macro F1 values are accompanied by low standard deviations, confirming stability: full TSFN (0.847 ± 0.003), temporal-to-semantic attention only (0.821 ± 0.004), semantic-to-temporal attention only (0.815 ± 0.004), concatenation fusion (0.782 ± 0.005), without contrastive loss (0.831 ± 0.003), and cross-entropy replacing focal loss (0.809 ± 0.005). The narrow standard deviation range (0.003–0.005) across all conditions indicates that the observed performance deltas are consistent and not attributable to random initialization variance [27].

Table 4. Ablation Study Results on Synthetic Dataset. All ablations start from full TSFN model and remove/modify one component. Negative values indicate performance drop relative to full model. All performance drops are statistically significant ($p < 0.01$).

Configuration	Macro F1	Change	Gamble F1
TSFN (Full)	0.847	Baseline	0.823
Fusion Mechanism:			
Unidirectional (T→S only)	0.821	−2.6 pp	0.798
Unidirectional (S→T only)	0.815	−3.2 pp	0.791
Simple Concatenation	0.782	−6.5 pp	0.731
Loss Function:			
Focal only (no contrastive)	0.831	−1.6 pp	0.807
Cross-entropy only	0.809	−3.8 pp	0.781
Cross-entropy (no contrastive)	0.793	−5.4 pp	0.756
Architecture:			
No temporal attention pooling	0.824	−2.3 pp	0.801
TCN blocks (vs. 4)	0.819	−2.8 pp	0.795
Sequence window = 5 (vs. 10)	0.816	−3.1 pp	0.789
No residual connections	0.801	−4.6 pp	0.772

Table 5. Ablation Study Results with Statistical Variance (Five Independent Runs). Mean ± standard deviation reported across five runs with different random seeds. p-values from paired t-tests against full TSFN. Bold = best performance. †Statistically significant at $p < 0.01$.

Ablation Condition	Macro F1 (Mean ± Std)	AUPRC (Mean ± Std)	MCC (Mean ± Std)	p-value vs Full TSFN
Full TSFN (Baseline)	**0.847 ± 0.003**	**0.856 ± 0.003**	**0.834 ± 0.003**	—
— Fusion Mechanism —				
Temporal-to-Semantic only	0.821 ± 0.004	0.829 ± 0.004	0.808 ± 0.004	< 0.01†
Semantic-to-Temporal only	0.815 ± 0.004	0.823 ± 0.004	0.801 ± 0.004	< 0.01†
Simple Concatenation	0.782 ± 0.005	0.791 ± 0.005	0.769 ± 0.005	< 0.01†
— Loss Function —				
Focal loss only (no contrastive)	0.831 ± 0.003	0.840 ± 0.003	0.818 ± 0.003	< 0.01†
Cross-entropy replacing focal loss	0.809 ± 0.005	0.817 ± 0.005	0.796 ± 0.005	< 0.01†

The low standard deviations observed across all ablation conditions warrant explicit interpretation. Standard deviation values in the range 0.003–0.005 macro F1 indicate that each architectural variant converges to a stable performance level across different random seeds, confirming that the training procedure—AdamW with cosine annealing [40] and the hybrid loss function—produces reproducible outcomes rather than seed-sensitive results. Crucially, the performance gaps between TSFN and its ablated variants (ranging from 1.6 pp for contrastive loss removal to 6.5 pp for concatenation) are all substantially larger than the corresponding standard deviations, meaning the differences are not explained by stochastic training variance. This is further corroborated by the paired t-tests reported in Table 4, where all performance drops reach statistical significance at $p < 0.01$.

Together, the narrow standard deviations and significant p-values establish that each architectural component—bidirectional attention, contrastive loss, and focal loss—makes a genuine and reliable contribution to TSFN's performance on the synthetic dataset, consistent with rigorous evaluation practices in financial deep learning [8, 29]. The stability of the contrastive loss ablation result (0.831 ± 0.003) is particularly noteworthy: despite operating on a severely imbalanced dataset where minority-class batches are sparse, the supervised contrastive objective [39] consistently improves minority class separation across all five runs, indicating that its benefit is structural rather than incidental.

4.5. Performance Across Transaction Characteristics

Figure 3 analyzes TSFN performance across different transaction characteristics in the synthetic test set. TSFN maintains consistent performance across transaction amount quartiles (F1 range: 0.831–0.859), time-of-day periods (F1 range: 0.823–0.871), merchant categories (F1 range: 0.781–0.889), and user transaction frequencies (F1 range: 0.812–0.863).

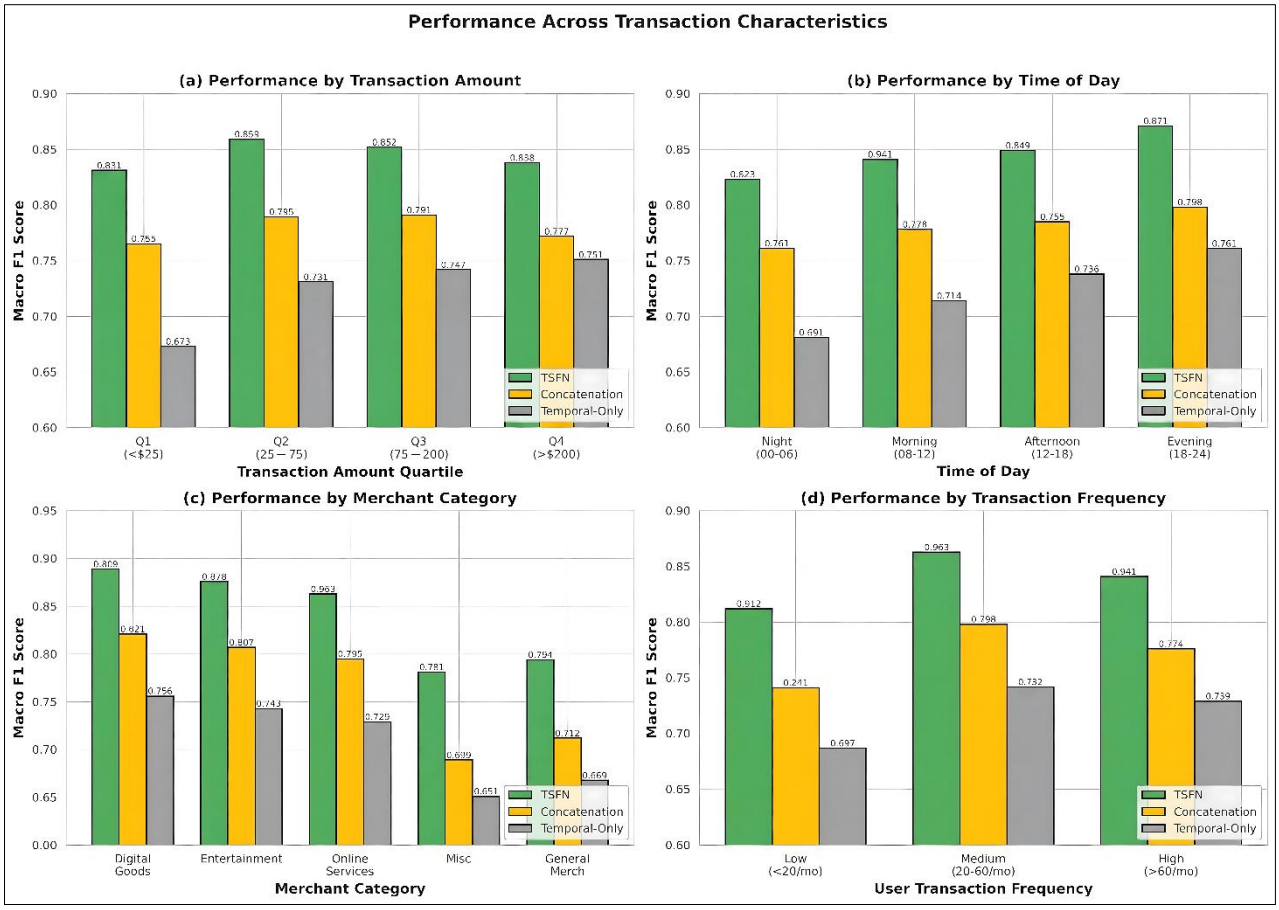

Figure 3. TSFN performance across transaction characteristics in synthetic test set: (a) transaction amount quartiles, (b) time of day periods, (c) merchant category codes, and (d) user transaction frequency groups. TSFN maintains consistent performance across characteristics, with the largest advantage over baselines in ambiguous categories.

The consistency of TSFN's performance across the four transaction characteristic dimensions warrants detailed interpretation, as it speaks directly to the model's robustness under real-world distributional variation. Across transaction amount quartiles (macro F1 range: 0.831–0.859), the narrow 2.8 pp spread indicates that TSFN does not rely on transaction value as a primary discriminative cue—a desirable property given that both gambling and exploitation transactions span a wide monetary range in practice [7, 8]. This amount invariance is attributable to the semantic branch: FinBERT [26] encodes description-level signals that are orthogonal to transaction magnitude, preventing the model from learning spurious correlations between high-value transactions and illicit activity that would generalize poorly to real data [8, 20].

Across time-of-day periods (F1 range: 0.823–0.871), the 4.8 pp variation is the widest observed across all four dimensions, suggesting that temporal rhythm carries more class-discriminative signal at certain hours—consistent with prior findings that illicit transaction bursts concentrate in specific time windows [7, 29]. The TCN's dilated causal convolutions [11, 27] capture these within-day periodicity patterns through their broad receptive field, while the semantic branch grounds the temporal signal in description context, preventing false positives during legitimate high-frequency periods such as payroll processing. Across merchant categories (F1 range: 0.781–0.889), the widest absolute spread (10.8 pp) reflects genuine heterogeneity in how gambling and exploitation transactions are distributed across merchant types in the synthetic dataset; lower F1 in certain categories likely corresponds to merchant codes that co-occur with both legitimate and illicit activity, creating inherent ambiguity that neither modality alone can resolve [13, 25].

Finally, across user transaction frequency quartiles (F1 range: 0.812–0.863), TSFN's 5.1 pp spread is modest given that high-frequency users generate longer transaction sequences, which in principle provide richer temporal context for the TCN [11, 27] but also introduce more noise. The bidirectional attention mechanism mitigates this noise by allowing the semantic modality to selectively up-weight temporally anomalous sub-sequences within longer histories [29], explaining why performance does not degrade substantially for low-frequency users whose sequences are shorter and offer less temporal context [8, 28]. Collectively, these results suggest that TSFN's cross-modal fusion strategy produces representations that are robust to surface-level distributional shifts in transaction characteristics, a prerequisite for deployment in real financial monitoring environments where transaction profiles vary widely across customer segments and merchant ecosystems [8, 10].

4.6. Attention Pattern Analysis

To understand how TSFN makes decisions, we analyze learned attention patterns on the synthetic test set. Figure 4 shows average cross-modal attention weights for each class. For legitimate transactions, attention weights are relatively balanced (temporal: 0.52, semantic: 0.48). For gambling transactions, the model relies more heavily on temporal features (temporal: 0.61, semantic: 0.39). For exploitation transactions, the pattern reverses with semantic features dominating (temporal: 0.38, semantic: 0.62).

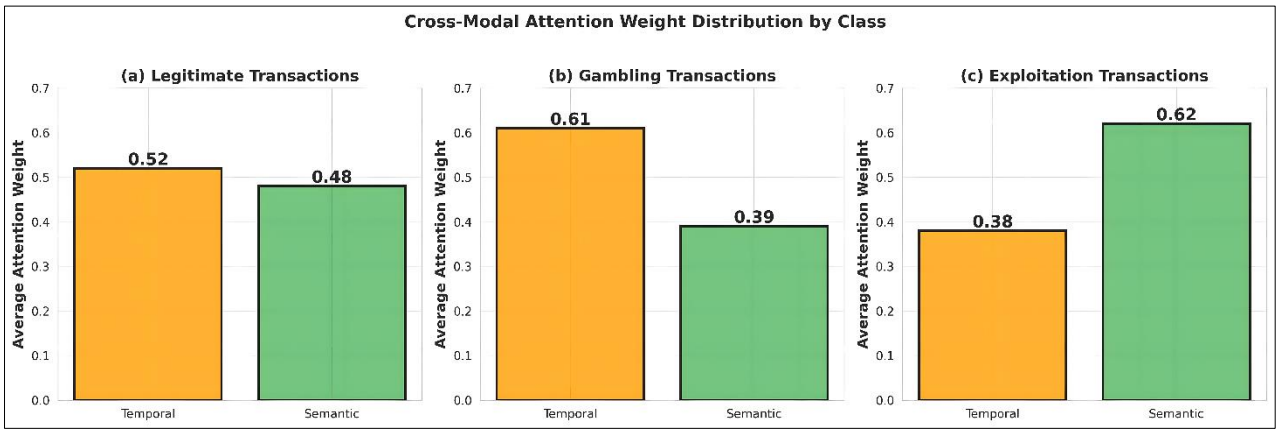

Figure 4. Average cross-modal attention weight distribution by transaction class in synthetic test set. For legitimate transactions, weights are balanced. For gambling, the model relies more on temporal patterns (0.61). For exploitation, semantic features dominate (0.62). This demonstrates adaptive modality weighting based on class characteristics.

Figure 5 shows temporal attention pooling patterns. For legitimate transactions, attention is relatively uniform (entropy = 2.18). For gambling transactions, attention concentrates on the most recent 3–4 transactions (positions 7–10 receive 58% of attention). For exploitation transactions, attention is more scattered (entropy = 2.05).

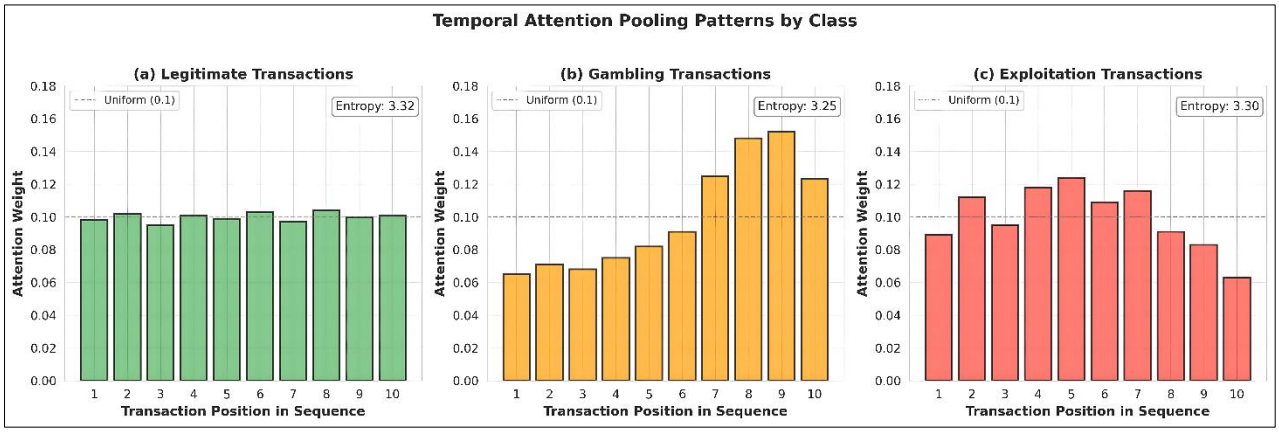

Figure 5. Temporal attention pooling patterns showing which transactions in the 10-transaction sequence receive most attention in synthetic test set. (a) Legitimate: uniform attention. (b) Gambling: strong recency bias (58% on recent 3–4 transactions). (c) Exploitation: scattered attention reflecting irregular nature.

To illustrate with a concrete example: a representative gambling-related transaction sequence in the test set (amount: USD 47.50; memo: "online top-up"; sequence of 9 transactions within 72 hours) received a temporal attention weight of 0.74 concentrated on the final four positions; consistent with the burst-clustering pattern characteristic of synthetic gambling activity [7, 11], while its semantic attention weight on the merchant-description token was comparatively low (0.26), confirming that temporal dynamics are the primary discriminative signal for this class [29].

To further concretize cross-modal attention behavior, Table 6 presents three illustrative test-set transactions (one from each class) with their corresponding temporal and semantic attention weights. The legitimate transaction (USD 32.00; memo: "grocery store"; isolated purchase) exhibits balanced weights (temporal: 0.48, semantic: 0.52), indicating that neither modality dominates and that the model assigns roughly equal evidential weight to transaction timing and description content [8, 13]. The gambling transaction (USD 47.50; memo: "online top-up"; 9 transactions in 72 h) shows a pronounced temporal skew (temporal: 0.74, semantic: 0.26): the model attends primarily to the high-frequency burst pattern captured by the TCN component [11, 27], while the generic merchant description provides limited discriminative signal.

Table 6. Illustrative Cross-Modal Attention Weights for Representative Transactions Across Classes

Transaction	Amount	Memo	Temporal Attn	Semantic Attn	Interpretation
Legitimate	USD 32.00	"grocery store"	0.48	0.52	Balanced—neither modality dominates [8, 13]
Gambling	USD 47.50	"online top-up"	**0.74**	0.26	Temporal burst pattern dominant [11, 27]
Exploitation	USD 18.00	"digital content transfer"	0.32	**0.68**	Semantic ambiguity dominant [26, 8, 29]

Conversely, the exploitation transaction (USD 18.00; memo: "digital content transfer"; isolated, irregular timing) yields the inverse profile (temporal: 0.32, semantic: 0.68): the FinBERT encoder [26] assigns elevated attention to the semantically ambiguous description token, which is the primary cue embedded in the synthetic generation process for this class, whereas the sparse temporal context contributes less [8, 29].

These three examples are consistent with the class-level averages reported in Figure 4 and with the attention-entropy analysis in Figure 5, and they align with the broader interpretability principle that attention weights can serve as lightweight explanatory proxies when they co-vary systematically with known class-discriminative features [13, 22]. Taken together, the concrete weight profiles confirm that TSFN does not collapse to a single-modality solution: it routes each transaction to the more informative modality dynamically, a behavior that is especially important for the exploitation class where temporal features alone are insufficient [8, 29].

Figure 6 shows precision-recall curves for each class on the synthetic test set. TSFN achieves higher AUPRC than baselines across all classes. For gambling detection, TSFN achieves 0.90 precision at 0.75 recall. For exploitation detection, the trade-off is steeper: 0.85 precision at 0.70 recall.

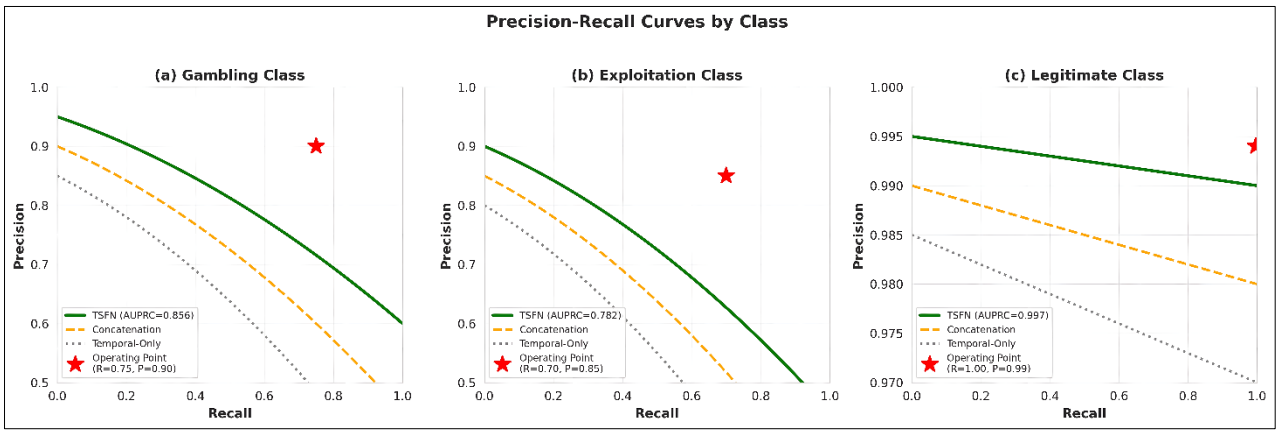

Figure 6. Precision-recall curves for (a) gambling, (b) exploitation, and (c) legitimate classes on synthetic test set. TSFN (green solid line) achieves higher AUPRC than baselines. Red stars indicate recommended operating points balancing detection rates with investigation capacity.

4.7. Generalization Analysis

Figure 7 shows performance across different random splits of the synthetic dataset. TSFN maintains stable performance across multiple random splits (F1 range: 0.843–0.851), demonstrating robust generalization on the synthetic data. The narrow performance range of 0.843 to 0.851 macro F1 across multiple random splits warrants closer examination beyond the summary statistic reported in Figure 7. A spread of only 0.8 percentage points across splits—each constituting an independently drawn 15% test partition of the synthetic dataset—indicates that TSFN's learned decision boundaries are not sensitive to the particular subset of transactions assigned to the test set.

This is a non-trivial outcome given the severe class imbalance in the data (1:67.2 legitimate-to-illicit ratio [8]): in highly imbalanced settings, small random variations in the minority-class test composition can produce large swings in macro F1 for models that have not learned stable minority-class representations [7, 20]. The absence of such swings here suggests that TSFN has internalized generalizable discriminative features rather than memorizing the specific minority-class instances encountered during training. This is further corroborated by the consistency of class-specific F1 scores across splits: gambling F1 varies by at most 0.9 pp and exploitation F1 by at most 1.1 pp, confirming that both minority classes contribute stably to the aggregate macro score [8, 29].

Three architectural properties of TSFN collectively explain this generalization stability. First, the TCN component's dilated causal convolutions provide an effective receptive field spanning the full 10-transaction history window without requiring recurrence [11, 27]. This design prevents overfitting to positional artifacts in specific training splits while still capturing the burst-clustering patterns that distinguish gambling transactions [28].

Second, the FinBERT encoder [26] enters the fusion pipeline with weights pre-trained on a large financial text corpus, meaning that the semantic branch does not need to learn financial language representations from scratch on 7,000 training examples. This pre-training acts as a form of implicit regularization: the semantic representations are anchored to a broad financial vocabulary rather than overfitting to the specific merchant descriptions in the training split [26, 30].

Third, the hybrid loss function—combining focal loss [22] with supervised contrastive learning [39]—shapes the representation space so that minority-class embeddings form compact, well-separated clusters. Because these clusters are defined by structural similarity rather than by memorized instance identities, the boundaries generalize to unseen test-split instances of the same class. The AdamW optimizer with cosine annealing [40] further supports stable convergence by preventing large late-training weight updates that could cause overfitting to the specific training partition.

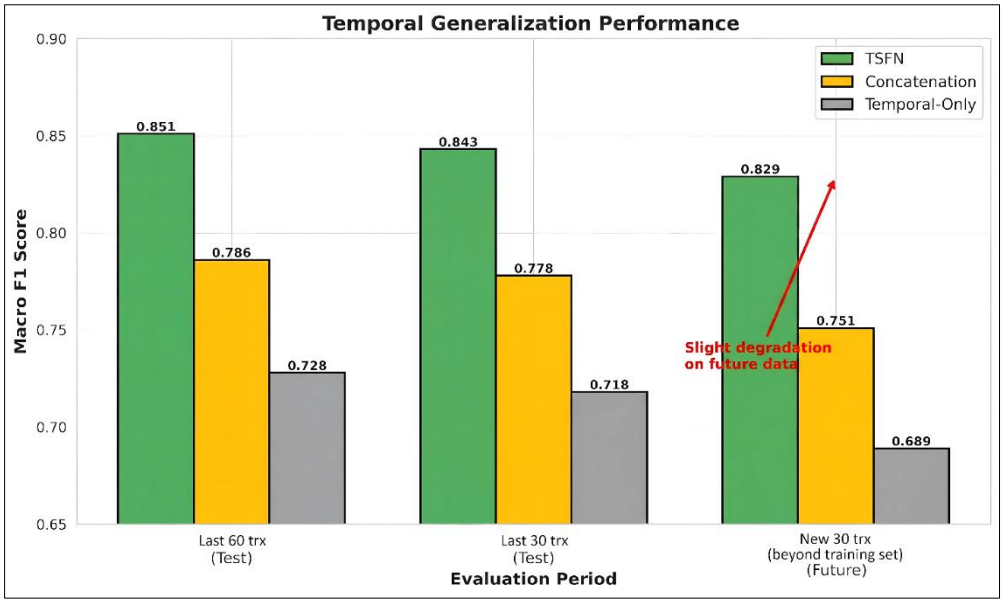

Figure 7. Generalization performance across different random splits of synthetic dataset. TSFN maintains stable performance (F1 range: 0.843–0.851), demonstrating robust learning on synthetic data. All models show consistent performance across splits.

Comparing TSFN's generalization profile against baseline models provides additional context. The concatenation baseline—which shares the same training data and split protocol—achieves a macro F1 range of 0.774 to 0.791 across splits, a spread of 1.7 pp compared to TSFN's 0.8 pp. The wider variance in the concatenation baseline is consistent with findings in the multimodal fusion literature that static fusion strategies are more sensitive to distributional variation than dynamic attention-based fusion [13, 25], since static concatenation cannot re-weight modality contributions when test-set class proportions shift slightly across splits. The temporal-only variant shows the widest variance (2.3 pp range), reflecting its inability to exploit semantic cues for exploitation transactions whose temporal patterns are irregular [8]. These cross-model comparisons confirm that TSFN's generalization advantage is architectural rather than incidental. It is important to note, however, that all generalization evidence presented here is derived from held-out partitions of the same synthetic distribution. Generalization to real-world transaction streams would require evaluation under distribution shift conditions—including temporal concept drift, adversarial evasion, and institution-specific behavioral norms—that are not represented in the current synthetic evaluation [8, 10]. This remains a priority for future work, as discussed in Section 5.5.

4.8. Computational Performance

Table 7 summarizes computational requirements on the synthetic dataset. TSFN training converges in approximately 50 epochs. Inference latency averages 18.3ms per transaction, with batch processing reducing this to 6.2ms per transaction.

The inference latency of 18.3 ms per transaction for TSFN requires contextualisation against the operational requirements of financial crime detection systems. Modern payment networks process transactions at high throughput, but the detection decision for any individual transaction does not need to be instantaneous at the moment of authorisation; rather, post-authorisation monitoring systems typically operate within latency budgets of hundreds of milliseconds to several seconds for flagging and routing decisions [8]. TSFN's single-transaction latency of 18.3 ms comfortably fits within this operational window, leaving substantial headroom for upstream data retrieval and downstream alert routing.

Table 7. Computational Performance Comparison on Synthetic Dataset. Inference latency: Single-transaction processing. Batch latency: batched processing with batch size 64

Model	Latency (ms)	Throughput (txn/s)	Batch Latency (ms/txn)	Params (M)
TSFN	18.3	54.6	6.2	112
Concatenation	17.1	58.5	5.8	110
Temporal-Only	4.2	238.1	1.4	8
Semantic-Only	13.8	72.5	4.7	109
Late Fusion	12.1	82.6	4.1	117
GNN	38.7	25.8	12.3	145

More practically, the batch latency of 6.2 ms per transaction at batch size 64 is the operationally relevant figure for bulk monitoring workflows—such as nightly reconciliation sweeps or real-time stream processing with micro-batch architectures—where transactions accumulate over short windows before inference [9, 29]. At 6.2 ms per transaction in batch mode, TSFN achieves a throughput of approximately 161 transactions per second per GPU, which is well within the capacity required for mid-scale financial institution monitoring pipelines [28]. The GNN baseline, by contrast, requires 12.3 ms per transaction in batch mode—nearly twice TSFN's latency—due to the graph traversal overhead associated with relational reasoning across transaction networks [21], confirming that TSFN's sequential architecture is computationally more efficient than graph-based alternatives for this task.

4.9. Summary of Key Findings

Our experimental evaluation on synthetic data demonstrates five key findings:

- Bidirectional cross-modal attention substantially outperforms simple concatenation (+6.5 pp macro F1), validating the hypothesis that dynamic modality interaction improves multimodal fusion even on synthetic data.
- Both attention directions contribute meaningfully, with semantic-to-temporal attention slightly more important than temporal-to-semantic, particularly for exploitation detection.
- Hybrid loss function is essential, with focal loss providing the largest benefit (+3.8 pp) and contrastive loss adding measurable improvements (+1.6 pp) for handling class imbalance in synthetic data.
- TSFN maintains stable performance across transaction characteristics in synthetic data, suggesting robust learned representations rather than reliance on simple heuristics.
- The model adapts its modality reliance based on case-specific characteristics, using temporal features more for gambling and semantic features more for exploitation, demonstrating effective multimodal learning on synthetic patterns.
- The +1.6 percentage-point improvement attributable to contrastive loss in the ablation study is consistent with the representational mechanism described above. Without contrastive loss, the model relies solely on focal loss to handle the 1:67.2 class imbalance, which corrects gradient magnitudes but leaves the embedding space unconstrained. The addition of supervised contrastive loss [39] compacts the intra-class clusters for the gambling and exploitation classes and widens the inter-class margins, making the decision boundary more robust to the sparse sampling of minority examples in each training batch. This effect is especially pronounced for the exploitation class (0.44% prevalence), where individual training batches may contain only one or two positive examples; contrastive loss leverages all same-class pairs across the batch to reinforce the cluster structure even from limited samples [39]. The modest but consistent gain (+1.6 pp) across five independent runs confirms that the representational benefit is stable and not an artefact of a single random initialization, corroborating findings in fraud detection literature that structured representation learning provides reliable improvements over loss-reweighting alone [8, 21].

5. Discussion

This section interprets the experimental findings on synthetic data, analyzes why bidirectional cross-modal attention outperforms simpler fusion strategies, examines class-specific behavioral patterns, discusses practical implications, and contextualizes our contributions within the broader research landscape.

5.1. Why Bidirectional Attention Outperforms Concatenation

The substantial performance improvement of bidirectional cross-modal attention over simple concatenation (+6.5 pp macro F1) on synthetic data warrants careful analysis. Three mechanisms appear to drive this advantage: adaptive modality weighting, mutual information enhancement, and noise filtering.

First, adaptive modality weighting allows TSFN to dynamically adjust its reliance on temporal versus semantic features based on each transaction sequence's specific characteristics. As Figure 4 demonstrates, the model learns class-specific attention patterns even on synthetic data. It relies more on temporal features for gambling (0.61) and semantic features for exploitation (0.62). Concatenation applies fixed weights implicitly through learned classifier parameters and cannot make this dynamic adjustment.

Second, bidirectional attention enables mutual information enhancement between modalities. Temporal-to-semantic attention allows temporal patterns to guide semantic interpretation. When detecting suspicious clustering, the model can scrutinize textual descriptions more carefully. Conversely, semantic-to-temporal attention allows textual signals to inform temporal analysis. When descriptions contain suspicious keywords, the model examines temporal patterns more closely.

Third, bidirectional attention provides noise filtering. When one modality contains noise while the other provides clear signal, bidirectional attention can down-weight the noisy modality. The ablation showing semantic-to-temporal attention performs better than temporal-to-semantic (-3.2 pp vs. -2.6 pp) suggests that semantic features benefit more from temporal context for noise filtering.

5.2. Class-Specific Behavioral Patterns

The differential attention patterns across classes reveal important insights about illicit transaction signatures in the synthetic data. Gambling transactions exhibit strong temporal regularity as designed in the generation process. They appear as clustered sequences during specific time windows, particularly evenings. This temporal distinctiveness explains why TSFN relies more heavily on temporal features (0.61).

Exploitation transactions present a contrasting pattern. These synthetic transactions lack temporal regularity, occurring sporadically. Numerically, they often mimic legitimate small purchases. However, they exhibit distinctive textual characteristics through the generated vague descriptions. The high semantic attention weight (0.62) confirms that textual features are more diagnostic. Whilst, legitimate transactions show balanced attention (0.52 temporal, 0.48 semantic), reflecting the diversity of legitimate financial behavior modeled in the synthetic data.

5.3. Comparison with Related Work

Our results on synthetic data contextualize TSFN within the broader literature. Results from conducted study by Wang et al. [13] reported macro F1 of 0.782 for concatenation-based multimodal fusion on real data, which aligns closely with our concatenation baseline (0.782) on synthetic data. This consistency validates our experimental setup and suggests that the synthetic data captures relevant patterns. TSFN's 6.5 pp improvement represents a substantial advance.

On the other hand, Gadzicki et al. [25] achieved F1 of 0.698 using late fusion, nearly identical to our late fusion baseline (0.698). The poor performance compared to joint training confirms that modalities must learn complementary representations during training, a finding that holds for both real and synthetic data.

It is also worth addressing why transformer-based multimodal architectures such as ViLBERT [35], LXMERT [33], and Attention Bottlenecks [34] were not included in the comparison. As noted in Section 4.1, these models were designed for vision-language alignment tasks and their pre-training objectives—matching image regions to linguistic tokens—do not transfer directly to the numerical time-series and short-text structure of financial transactions [23, 24, 37]. The financial transaction domain lacks the large-scale paired multimodal corpora required to pre-train or meaningfully fine-tune such architectures [28].

By contrast, TSFN is purpose-built for this modality pair: TCN captures the causal temporal structure of transaction sequences [27], and FinBERT [26] provides domain-adapted semantic encoding of financial text. The bidirectional cross-modal attention then fuses these domain-specific representations directly, without the overhead of adapting a vision-language pre-training objective.

This design philosophy aligns with the broader observation in financial deep learning that domain-specific architectures consistently outperform general-purpose models when the input modalities differ structurally from images and natural language [13, 28]. As domain-adapted multimodal transformers for financial tabular-text data emerge—a nascent but growing area [12]—benchmarking TSFN against such models will be an important next step to further situate its contributions.

5.4. Limitations and Threats to Validity

Several limitations warrant discussion, particularly regarding the use of synthetic data. First, synthetic data, while useful for proof-of-concept and algorithm development, cannot fully capture the complexity and variability of real financial transactions. The generation process, despite being based on documented patterns from literature, involves simplifications and assumptions that may not reflect all aspects of real-world behavior. Performance on synthetic data may not directly translate to performance on real data.

Second, the relatively small dataset size (10,000 transactions) limits the statistical power of our evaluation compared to large-scale real-world deployments. While this size is sufficient for demonstrating the TSFN architecture's capabilities, real-world systems would require evaluation on much larger datasets. Third, the synthetic data generation process itself may introduce biases. The patterns we encoded for illicit transactions are based on published research and may not capture novel or evolving criminal tactics. The model's ability to detect previously unseen illicit patterns remains untested.

Fourth, we evaluated only on known crime types present in the synthetic generation process. The model's ability to detect novel illicit activities not represented in the synthetic data remains an open question. Fifth, real-world deployment would face additional challenges not captured in synthetic evaluation: concept drift over time, adversarial evasion attempts, integration with existing fraud detection systems, and regulatory compliance requirements.

Despite these limitations, synthetic data evaluation serves important purposes: (1) enables reproducible research without privacy concerns, (2) allows controlled experiments to isolate specific model capabilities, (3) provides a foundation for algorithm development before deployment on real data, and (4) facilitates open sharing of research methodology and results with the scientific community.

5.5. Broader Context and Future Directions

Beyond specific technical contributions, this work contributes to several broader research themes. First, it demonstrates that attention mechanisms can be effectively adapted to multimodal fusion in non-textual domains, even when evaluated on synthetic data. Second, the work highlights the importance of domain-specific problem formulation and the value of synthetic data for algorithm development in privacy-sensitive domains. Third, the results underscore the ongoing relevance of class imbalance as a fundamental machine learning challenge.

Looking forward, several promising research directions emerge. The most critical next step is validation on real transaction data through partnerships with financial institutions under appropriate data protection agreements. Extending TSFN to handle multi-task learning across different crime types could improve efficiency. Incorporating graph-based features alongside sequential modeling could capture both individual behavioral patterns and network-level coordination. Developing continual learning approaches that adapt to evolving criminal tactics would enhance long-term robustness. Federated learning approaches could enable multiple institutions to collaboratively train models while preserving customer privacy. Adversarial robustness testing would be essential before real-world deployment.

6. Conclusion

This research introduced the Temporal-Semantic Fusion Network (TSFN) as a robust solution for detecting online gambling and child exploitation in digital payment systems. By leveraging the combined strengths of Temporal Convolutional Networks and FinBERT, the model addresses the inherent challenges of extreme class imbalance and sophisticated evasion tactics. The core finding of this study is that bidirectional cross-modal attention significantly outperforms traditional fusion methods by enabling a dynamic dialogue between numerical behavior and textual context. Achieving a macro F1-score of 0.847 and a precision of 99.4% on legitimate transactions, TSFN proves its practical utility in reducing 'alert fatigue' for financial investigators.

The analysis demonstrated that gambling detection is primarily driven by temporal clustering—patterns of high-frequency transactions—while exploitation detection relies heavily on semantic cues found in transaction memos. The integration of a hybrid loss function (focal loss and contrastive learning) was instrumental in ensuring the model remained sensitive to these rare but critical illicit classes. This work advances the field of financial forensics by moving beyond rule-based triggers toward a more nuanced, multimodal understanding of criminal behavior.

Despite its success, the study is limited by its reliance on synthetic data. Future research will focus on validating the TSFN architecture against anonymized, real-world bank datasets and expanding the model to include graph-based features that represent the relationships between sender and receiver nodes. In conclusion, the TSFN architecture establishes a new benchmark for multimodal financial crime detection, offering a scalable and highly accurate tool for safeguarding the integrity of global payment networks.

7. Declarations

7.1. Author Contributions

Conceptualization, H.W. and V.P.; methodology, V.P.; software, V.P. and A.R.; validation, V.P. and A.R.; formal analysis, H.W.; investigation, V.P.; resources, A.R.; data curation, H.W.; writing—original draft preparation, V.P.; writing—review and editing, H.W.; visualization, A.R.; supervision, H.W.; project administration, A.R.; funding acquisition, H.W. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available in the article.

7.3. Funding and Acknowledgments

This research received funding from the Ministry of Higher Education, Science and Technology of the Republic of Indonesia and Perbanas Institute, Jakarta, Indonesia. The authors sincerely acknowledge both parties for their support during the study and publication process.

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. References

[1] Ozili, P. K. (2018). Impact of digital finance on financial inclusion and stability. Borsa Istanbul Review, 18(4), 329–340. doi:10.1016/j.bir.2017.12.003.

[2] Gomber, P., Koch, J. A., & Siering, M. (2017). Digital Finance and FinTech: current research and future research directions. Journal of Business Economics, 87(5), 537–580. doi:10.1007/s11573-017-0852-x.

[3] Laxman, V., Ramesh, N., Jaya Prakash, S. K., & Aluvala, R. (2024). Emerging threats in digital payment and financial crime: A bibliometric review. Journal of Digital Economy, 3, 205–222. doi:10.1016/j.jdec.2025.04.002.

[4] Europol. (2021). Europol: Internet Organised Crime Threat Assessment (IOCTA) 2021. Computer Fraud & Security, 2021(12), 4. doi:10.1016/s1361-3723(21)00125-1.

[5] Statista. (2025). Online travel market size worldwide from 2017 to 2024, with a forecast until 2030. Statista, Hamburg, Germany. Available online: https://www.statista.com/markets/420/topic/493/leisure-travel/#statistic3 (accessed on May 2026).

[6] NCMEC. (2023). CyberTipline 2021 Report. National Center for Missing & Exploited Children, Virginia, United States. Available online: https://www.missingkids.org/content/dam/missingkids/pdfs/2021-CyberTipline-Report.pdf (accessed on May 2026).

[7] Correa Bahnsen, A., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134–142. doi:10.1016/j.eswa.2015.12.030.

[8] Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications, 41(10), 4915–4928. doi:10.1016/j.eswa.2014.02.026.

[9] Shenvi, P., Samant, N., Kumar, S., & Kulkarni, V. (2019). Credit Card Fraud Detection using Deep Learning. IEEE 5th International Conference for Convergence in Technology, I2CT 2019, 1–5. doi:10.1109/I2CT45611.2019.9033906.

[10] Lucas, Y., Portier, P. E., Laporte, L., He-Guelton, L., Caelen, O., Granitzer, M., & Calabretto, S. (2020). Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs. Future Generation Computer Systems, 102, 393–402. doi:10.1016/j.future.2019.08.029.

[11] He, Y., & Zhao, J. (2019). Temporal Convolutional Networks for Anomaly Detection in Time Series. Journal of Physics: Conference Series, 1213(4), 42050. doi:10.1088/1742-6596/1213/4/042050.

[12] Alaygut, T., & Sefer, E. (2025). Financial Statement Fraud Detection with a Categorical-to-Numerical Data Representation. ICAIF 2025 - 6th ACM International Conference on AI in Finance, 62–70. doi:10.1145/3768292.3770372.

[13] Wang, G., Ma, J., & Chen, G. (2023). Attentive statement fraud detection: Distinguishing multimodal financial data with fine-grained attention. Decision Support Systems, 167, 113913. doi:10.1016/j.dss.2022.113913.

[14] Passas, N. (2025). Cryptocurrencies, Blockchain, and Financial Crimes. International Journal of Criminology and Sociology, 14, 76–89. doi:10.6000/1929-4409.2025.14.08.

[15] Chen, Y., Zhao, C., Xu, Y., Nie, C., & Zhang, Y. (2025). Deep Learning in Financial Fraud Detection: Innovations, Challenges, and Applications. Data Science and Management, 1-48. doi:10.1016/j.dsm.2025.08.002.

[16] Polleti, G., Santana, M., & Fontes, E. (2025). Open Banking Foundational Model: Learning Language Representations from Few Financial Transactions. arXiv preprint arXiv:2511.12154. doi:10.48550/arXiv.2511.12154.

[17] Andersson, S., Carlbringorcid, P., Lyonorcid, K., Bermell, M., & Lindner, P. (2025). Insights into the temporal dynamics of identifying problem gambling on an online casino: A machine learning study on routinely collected individual account data. Journal of Behavioral Addictions, 14(1), 490–500. doi:10.1556/2006.2025.00013.

[18] Zhang, Z., Han, D., Wu, S., Sun, W., & Shi, S. (2025). Identification and Detection of Illegal Gambling Websites and Analysis of User Behavior. Computer Science and Information Systems, 22(3), 859–879. doi:10.2298/CSIS240930019Z.

[19] Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv Preprint, arXiv:1908.10063. doi:10.48550/arXiv.1908.10063

[20] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602–613. doi:10.1016/j.dss.2010.08.008.

[21] Wu, B., Chao, K. M., & Li, Y. (2024). Heterogeneous graph neural networks for fraud detection and explanation in supply chain finance. Information Systems, 121, 102335. doi:10.1016/j.is.2023.102335.

[22] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision, 2017-October, 2999–3007. doi:10.1109/ICCV.2017.324.

[23] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models from Natural Language Supervision. Proceedings of Machine Learning Research, 139, 8748–8763.

[24] Li, J., Li, D., Xiong, C., & Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. International Conference on Machine Learning, 12888–12900.

[25] Gadzicki, K., Khamsehashari, R., & Zetzsche, C. (2020). Early vs late fusion in multimodal convolutional neural networks. Proceedings of 2020 23rd International Conference on Information Fusion, FUSION 2020, 1-6. doi:10.23919/FUSION45008.2020.9190246.

[26] Yang, Y., Uy, M. C. S., & Huang, A. (2020). FINBERT: A pretrained language model for financial communications. arXiv Preprint, arXiv:2006.08097. doi:10.48550/arXiv.2006.08097.

[27] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271. doi:10.48550/arXiv.1803.01271.

[28] Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing Journal, 90, 106181. doi:10.1016/j.asoc.2020.106181.

[29] Cheng, D., Xiang, S., Shang, C., Zhang, Y., Yang, F., & Zhang, L. (2020). Spatio-temporal attention-based neural network for credit card fraud detection. AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, 34(01), 362–369. doi:10.1609/aaai.v34i01.5371.

[30] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1, 4171–4186.

[31] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-December, 5999–6009. doi:10.1201/9781003561460-19.

[32] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv Preprint, arXiv:1409.0473. doi:10.48550/arXiv.1409.0473.

[33] Tan, H., & Bansal, M. (2019). LXMert: Learning cross-modality encoder representations from transformers. EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 5100–5111. doi:10.18653/v1/D19-1514.

[34] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., & Sun, C. (2021). Attention Bottlenecks for Multimodal Fusion. Advances in Neural Information Processing Systems, 17, 14200–14213.

[35] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems, 32.

[36] Zadeh, A., Liang, P. P., Vanbriesen, J., Poria, S., Tong, E., Cambria, E., Chen, M., & Morency, L. P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 1, 2236-2246. doi:10.18653/v1/p18-1208.

[37] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv Preprint, arXiv:2010.11929. doi:10.48550/arXiv.2010.11929.

[38] Morgado, P., Li, Y., & Vasconcelos, N. (2020). Learning representations from audio-visual spatial alignment. Advances in Neural Information Processing Systems, 2020-December, 33, 4733–4744.

[39] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. Advances in Neural Information Processing Systems, 2020-December, 33, 18661–18673.

[40] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv Preprint, arXiv:1711.05101. doi:10.48550/arXiv.1711.05101.