Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/210101

TFDNet: A Transformer-Based Multimodal Framework for Fraud Detection in Multilingual Kazakh-Russian-English SMS and Voice Calls

Azamat Serek[1,2], Aisultan Shoiynbek[1], and Nurshapagat Shapay[2]*

[1]Narxoz University, Almaty, Kazakhstan
[2]School of Software Engineering, Astana IT University, Astana, Kazakhstan

Received: 12 Aug. 2027, Revised: 12 Oct. 2027, Accepted: 19 Oct. 2027
Published online: 1 Jan. 2027

Abstract: Fraudulent SMS (smishing) and voice calls (vishing) remain among the most common vectors for social-engineering attacks, exploiting urgency, authority, and trust to extract money or credentials from victims. Existing detection systems typically treat text and speech as separate problems and are almost universally developed and evaluated for monolingual English traffic, leaving code-switched, lower-resource settings under-served. We target Kazakhstan as a concrete case: Kazakh holds constitutional status as the state language, Russian carries official status alongside it in state institutions, real-world financial and telecom communication routinely mixes Kazakh and Russian text with English financial and technical loanwords, and Kazakhstan's Financial Monitoring Agency has documented active SMS-blaster and impersonation fraud campaigns against Halyk Bank and Beeline customers as recently as 2026, alongside vishing schemes impersonating the National Bank and law enforcement. In this paper we propose and train TFDNet (Transformer-based Fraud Detection Network), a unified multimodal architecture for fraud detection across SMS and voice-call channels, comprising a text encoder, a paralinguistic-feature encoder, a modality-gating mechanism for missing modalities, and a cross-attention fusion module that models the interaction between linguistic content and paralinguistic cues such as prosody, pacing, and hesitation, together with a training objective combining weighted cross-entropy and a supervised contrastive term. We construct two synthetic, incident-inspired Kazakh-Russian corpora – an SMS fraud corpus of 4,903 labeled messages and a voice/vishing corpus of 2,817 labeled call transcripts with paralinguistic descriptors, evaluated separately rather than as paired multimodal samples – with the Kazakh and Russian wording translated and reviewed by professional translators based in Almaty. Because pretrained transformer and speech-model checkpoints were not accessible to us for this study, TFDNet's text and fusion modules are trained end-to-end from randomly initialized weights rather than fine-tuned from a pretrained checkpoint; we report this trained-from-scratch model's results directly, alongside classical TF-IDF baselines for comparison, rather than substituting a different model family for TFDNet. TFDNet's fusion model reaches F1 = 0.974 (AUC = 0.997) on the voice corpus, ahead of every unimodal and naive-fusion baseline, at 2.85 ms mean CPU model-inference latency per call (excluding paralinguistic feature extraction from audio). On the SMS corpus, TFDNet's text branch reaches F1 = 0.962 (AUC = 0.960) at 1.27 ms mean latency, within a point of the strongest classical TF-IDF baseline (F1 = 0.967); their bootstrap confidence intervals overlap substantially, so this difference should not be read as statistically decisive, and we discuss why a transformer trained from scratch, without pretrained initialization, does not clearly exceed a well-tuned linear bag-of-words baseline on a few thousand training examples. We discuss architectural choices, ablations, an explainability analysis of the classical text baseline, and, integrated into the discussion, the study's limitations and the ethical considerations of processing voice-biometric signals for fraud detection.

Keywords: Fraud detection; smishing; vishing; transformer; multimodal learning; cross-attention; code-switching; low-resource NLP; Kazakh language; natural language processing; speech processing.

1 Introduction

Text-message and voice-call fraud has grown alongside the expansion of mobile communication. Attackers impersonate banks, delivery services, tax authorities, or family members to induce victims into transferring funds, disclosing one-time passcodes, or installing malicious software. Two related but distinct attack surfaces are commonly discussed: *smishing*, in which a fraudulent message is delivered by SMS or a messaging application [1,2,3], and *vishing*, in which the attack

* Corresponding author e-mail: azamat.serek@astanait.edu.kz

© 2027 NSP
Natural Sciences Publishing Cor.

is conducted over a live or pre-recorded voice call [4,5,6]. Reports and systematic reviews on phishing consistently identify these channels as leading sources of fraud complaints [7,8,9], and the increasing accessibility of text-to-speech and voice-cloning tools has lowered the barrier to producing convincing fraudulent voice content [6,10,11].

Almost all of this literature, however, targets monolingual English text. This is a significant gap: Kazakhstan, as one concrete example, has a language landscape shaped by its constitution, which designates Kazakh as the state language while granting Russian official status for use on an equal footing in state institutions and local self-government; English financial and technology loanwords are also in routine use, and its telecom fraud problem is well documented. In April 2026, Kazakhstan's Financial Monitoring Agency, working with the mobile operator Tele2, arrested a gang operating a vehicle-mounted "SMS blaster" capable of sending 100,000 impersonation messages per hour, spoofing Halyk Bank (the country's largest bank) and Beeline (a major regional telecom brand) to harvest one-time passwords [12,13]. Separately, Kazakhstani police have documented multi-stage vishing schemes in which a caller first impersonates a National Bank official warning of account compromise, then hands off to accomplices impersonating law enforcement to extract tens of millions of tenge in cash [14], a pattern the government's own e-service portal warns citizens about directly [15]. This paper builds on our group's own prior work on AI-based communication fraud detection in Kazakhstan, including a general architecture for detecting fraudulent communication [16] and a real-time, multi-turn dialogue-analysis approach specifically for fraudulent phone calls [17]; the present work extends that line of work with a unified, mixed-language, multimodal treatment of both the SMS and voice channels, evaluated separately rather than as paired samples. Meanwhile, Kazakh remains a low-resource language for NLP: labeled corpora are scarce, pretrained transformer coverage is limited relative to English or Russian, and recent work has had to resort to multilingual transfer and parameter-efficient fine-tuning to reach usable spam-classification accuracy on Kazakh text at all [18], while machine translation for the Kazakh-Russian code-switched pair specifically remains an active, unsolved low-resource problem [19]; adjacent efforts in the same low-resource setting include crowdsourced Kazakh-English translation [20] and hybrid LLM-ASR pipelines for Kazakh speech [21]. Fraud detection for this setting is therefore not a routine translation exercise: it requires handling mixed-language Kazakh-Russian text, in a language pair that current NLP tooling does not yet serve well; as Section 4 notes, the wording constructed for this study switches language at message and clause boundaries rather than modeling finer-grained, token-level code-switching within a single clause.

Current defenses are largely channel-specific. SMS filtering relies on keyword lists, sender-reputation databases, and, increasingly, fine-tuned transformer classifiers [22]. Voice-call defenses rely on caller-ID reputation, carrier-level call blocking, and telecom-side anomaly detection over call-detail records [23,24]. These approaches share a limitation: they analyze each channel in isolation and rarely exploit the fact that fraud campaigns frequently reuse the same scripts, narratives, and pressure tactics across text and voice. A message that reads as ambiguous in isolation (for example, a short prompt to "call this number to verify your account") may be unambiguously fraudulent once combined with the vocal delivery of the follow-up call, which is often marked by scripted phrasing, unnatural pacing, or coached urgency [25,26].

SMS spam and fraud filters have moved from bag-of-words and TF-IDF classifiers toward fine-tuned transformer encoders. Transformer-based language modeling approaches with explainability analysis [1] and large-language-model pairing frameworks [22] have improved robustness to obfuscation, and ensemble approaches combining GPT-based embeddings with classical classifiers report accuracy improvements over TF-IDF and static word-embedding baselines [27]. Public smishing-specific corpora, such as the SMS Phishing Dataset for Machine Learning and Pattern Recognition [3], and hybrid deep-learning pipelines evaluated across multiple SMS corpora [28,2] have supported this progress, and systematic reviews cover the broader phishing-detection literature across channels [7,8,29]. Large language models have further complicated the landscape on both sides of the arms race. GPT-family models fine-tuned or prompted for phishing-email detection achieve strong detection accuracy while also being usable to generate more convincing phishing content [30, 31,32], and LLM-based website and content classifiers have likewise been proposed [33,34]. Dedicated studies further show that LLMs can be prompted to rewrite vishing call scripts so that they evade classifiers trained only on the original transcripts [25], which strengthens the case for detectors that also draw on acoustic and paralinguistic evidence that is harder to fully naturalize.

Voice-fraud detection has historically combined caller-reputation scoring with natural-language analysis of call transcripts. Real-time Korean vishing detection converts call audio to text and applies classical machine-learning classifiers to the transcripts [4], and more recent systems combine acoustic and textual features in a real-time pipeline, reporting detection accuracy above 98% on held-out call data [5]. A complementary line of work studies the offensive side of this problem, showing that fully AI-automated vishing attacks are increasingly feasible with current speech-synthesis and dialogue technology [6], and at the network level, telecom-side fraud-prevention systems that fuse call-detail-record features with NLP analysis of message content, combined via attention mechanisms, have also been reported [23,24]. Self-supervised speech encoders such as wav2vec 2.0 provide strong general-purpose acoustic representations without hand-crafted features and have been fine-tuned for paralinguistic tasks including vocalized-emotion and stuttering recognition [35]. Separately, the ASVspoof initiative has established benchmarks for detecting spoofed and deepfake speech in increasingly realistic, "in the wild" conditions [10], and subsequent work has proposed contrastive-learning-based detectors that are more robust to manipulation attacks [11]. Multimodal architectures that jointly encode text and audio typically

© 2027 NSP
Natural Sciences Publishing Cor.

combine modality-specific encoders with a fusion module implemented as concatenation, gated fusion, or cross-attention: attention-based fusion of audio and visual streams has been shown to improve event-recognition accuracy over unimodal baselines [36], and cross-attention fusion of text and audio representations, in which one modality's tokens attend over the other modality's representations, has produced state-of-the-art results in multimodal speech-emotion recognition [26]. Fraud detection has not been extensively studied from this multimodal cross-attention perspective, which motivates the architecture proposed below.

The key contributions of this paper are summarized as follows:

1. We propose TFDNet, a transformer-based architecture that processes SMS text and voice-call audio through modality-specific encoders and fuses them with a cross-attention layer, while gracefully degrading to a single modality when only one is available.
2. We give the full mathematical formulation of the model, including the gating mechanism, the cross-attention fusion block, and a training objective that combines weighted cross-entropy with a supervised contrastive term.
3. We construct two Kazakh-Russian mixed-language corpora with English financial/technical loanwords, grounded in real, documented Kazakhstani fraud patterns; we are not aware of a publicly available resource for multimodal fraud detection in this language setting, though our search of the literature was not exhaustive. The corpora and the scripts that built them are available from the authors on request. The Kazakh and Russian wording was translated and reviewed by professional translators based in Almaty, Kazakhstan; Section 2.2 describes the construction procedure in full.
4. We train and evaluate TFDNet directly, rather than substituting a different model family: because pretrained transformer and speech-model checkpoints were not accessible to us for this study, TFDNet's text and fusion modules are trained end-to-end from randomly initialized weights, and we report this trained-from-scratch model's own results, alongside classical TF-IDF baselines for comparison, a distinction we maintain throughout rather than only in the Discussion.
5. We report TFDNet's own measured inference latency, 1.27 ms per SMS message and 2.85 ms per voice call on commodity CPU hardware, which matters directly for telecom-scale use cases where an SMS gateway or call platform must decide whether to block or flag traffic before it reaches the recipient.

The rest of the manuscript is organized as follows. Section 2 presents the TFDNet model, the two corpora, and the experimental setup. Section 3 reports the results, including the ablation, explainability, and latency analyses. Section 4 discusses the findings together with the limitations and ethical considerations of the study. Lastly, Section 5 concludes the paper and provides directions for future research.

2 Materials and Methods

2.1 Proposed Model: TFDNet

TFDNet accepts three input configurations at inference time: (a) SMS or chat text only, (b) call audio only, and (c) call audio together with its transcript, either produced by an automatic speech recognition (ASR) system or supplied alongside the recording. We first define notation, then the text encoder, the acoustic encoder, the modality-gating mechanism, the cross-attention fusion block, and the training objective. Figure 1 gives a schematic overview of the full pipeline and the equation each block corresponds to.

2.1.1 Notation

Let an input instance be a pair $(\mathbf{x}_{\text{txt}}, \mathbf{x}_{\text{aud}})$, where $\mathbf{x}_{\text{txt}} = (w_1, \ldots, w_L)$ is a token sequence (an SMS message or an ASR transcript) and $\mathbf{x}_{\text{aud}} \in \mathbb{R}^T$ is a raw waveform of T samples. A binary modality-availability indicator $m_a \in \{0,1\}$ records whether call audio is present for this instance ($m_a = 0$ for SMS-only inputs). All learned weight matrices $W_{(\cdot)}$ and bias vectors $b_{(\cdot)}$ below are free parameters fit during training; d denotes the shared embedding dimension used for the fused representation.

2.1.2 Text Encoding Branch

The text branch encodes the token sequence $\mathbf{x}_{\text{txt}}$ with a multi-head self-attention transformer encoder (pretrained and fine-tuned when a suitable checkpoint is available, or trained end-to-end from randomly initialized weights otherwise, as in the experiments of Section 2.3), producing contextual token representations

$$H^{(t)} = \text{Enc}_{\text{txt}}(w_1, \ldots, w_L) \in \mathbb{R}^{L \times d_h}, \tag{1}$$

© 2027 NSP
Natural Sciences Publishing Cor.

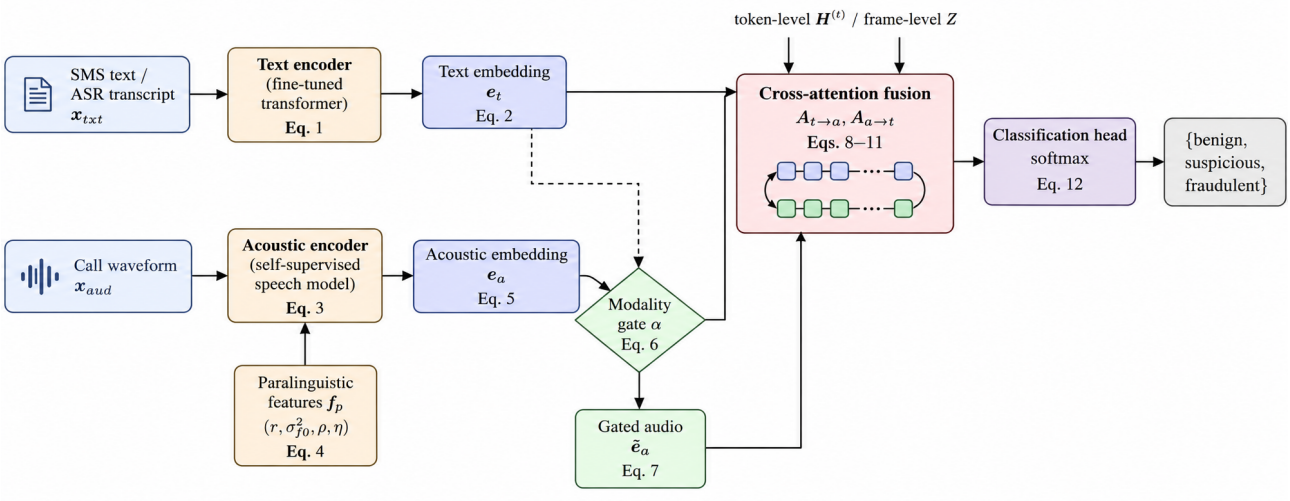


Figure 1. TFDNet architecture. Text and audio are encoded by modality-specific branches (Eqs. (1)–(6)); a scalar gate (Eq. (7)) substitutes a learned null embedding when audio is absent (Eq. (8)); token- and frame-level representations are combined by cross-attention fusion (Eqs. (9)–(12)) before classification (Eq. (13)).

where d_h is the encoder's hidden size. The pooled representation is taken from the classification-token position, $h_t = H^{(t)}_{[\mathrm{CLS}]} \in \mathbb{R}^{d_h}$, and projected to the shared embedding space by a two-layer feed-forward network with a GELU non-linearity,

$$e_t = W_2\, \sigma_{\mathrm{GELU}}(W_1 h_t + b_1) + b_2 \ \in \mathbb{R}^d, \tag{2}$$

with $W_1 \in \mathbb{R}^{d_h \times d_h}$, $W_2 \in \mathbb{R}^{d \times d_h}$. Domain-specific pre-processing normalizes phone numbers, links, and monetary amounts into placeholder tokens prior to encoding, so that the model generalizes across specific numeric values rather than memorizing them.

2.1.3 Acoustic Encoding Branch

When audio is available ($m_a = 1$), the general form of the architecture passes the raw waveform $\mathbf{x}_{\mathrm{aud}}$ through a self-supervised speech encoder [35] when a pretrained checkpoint is available, producing frame-level representations

$$Z = \mathrm{Enc}_{\mathrm{aud}}(\mathbf{x}_{\mathrm{aud}}) \in \mathbb{R}^{N \times d_z}, \tag{3}$$

for N frames of dimension d_z, which are then mean-pooled to $\bar{z} = \frac{1}{N}\sum_{n=1}^{N} Z_n \in \mathbb{R}^{d_z}$.

The experiments in this paper do not have raw audio available (Section 2.2.2 explains why), so this general pipeline is not what is fitted. Instead, the acoustic branch operates directly on the four-dimensional hand-crafted paralinguistic vector

$$f_p = \left[\, r,\ \sigma^2_{f_0},\ \rho,\ \eta \,\right]^{\top} \in \mathbb{R}^4, \tag{4}$$

comprising speaking rate r, pitch (fundamental-frequency) variance $\sigma^2_{f_0}$, pause ratio ρ, and a voice-activity-derived hesitation ratio η. To still exercise the multi-frame cross-attention mechanism of Section 2.1.5 rather than degenerating to single-vector attention, f_p is projected into a short sequence of N_p pseudo-frames through N_p independently learned linear maps sharing the same input,

$$Z_n = \sigma_{\mathrm{GELU}}\left(W_n^{(z)} f_p + b_n^{(z)}\right) \in \mathbb{R}^{d_z}, \qquad n = 1, \ldots, N_p, \tag{5}$$

giving $Z = [Z_1; \ldots; Z_{N_p}] \in \mathbb{R}^{N_p \times d_z}$ (we use $N_p = 4$, $d_z = 32$); this Z plays the same structural role as (3) in the fusion block below, with $N = N_p$, but is constructed from four scalar descriptors rather than measured from a waveform, and should

© 2027 NSP
Natural Sciences Publishing Cor.

be read as a stand-in for genuine frame-level acoustic structure rather than an approximation of it. The pooled vector $\bar{z} = \frac{1}{N_p}\sum_n Z_n$ is projected to the shared embedding space,

$$e_a = W_4\,\sigma_{\mathrm{GELU}}(W_3\,\bar{z} + b_3) + b_4 \;\in \mathbb{R}^d, \tag{6}$$

without re-concatenating f_p, since $\bar{z}$ is already a deterministic function of f_p alone through (5). When $m_a = 0$, no acoustic forward pass is performed and e_a is undefined; it is replaced downstream by a learned null-modality embedding $e_\varnothing \in \mathbb{R}^d$.

2.1.4 Modality Gating

To avoid injecting noise from the null embedding while still allowing the model to weight partial or noisy audio smoothly rather than with a hard switch, a scalar gate is computed from the modality indicator and a lightweight summary of the two candidate embeddings,

$$\alpha = \sigma\Big(w_g^\top\big[m_a,\ \|e_t\|_2,\ \|e_a\|_2\big] + b_g\Big) \in (0,1), \tag{7}$$

where $\sigma(\cdot)$ is the logistic sigmoid and, when $m_a = 0$, the term $\|e_a\|_2$ is set to 0 and e_a itself is not used. The gated acoustic embedding is then

$$\hat{e}_a = m_a\,\alpha\,e_a + (1 - m_a\,\alpha)\,e_\varnothing. \tag{8}$$

By construction, $\hat{e}_a = e_\varnothing$ when $m_a = 0$, and the model reduces to a purely text-driven classifier for SMS-only inputs.

2.1.5 Cross-Attention Fusion

When both modalities are present, TFDNet models interactions between linguistic and acoustic information with a cross-attention block operating on the token-level representation $H^{(t)}$ from (1) and the frame-level representation Z from (3) (or (5) in the experiments of this paper). Both are first linearly projected to a common key/query dimension d_k:

$$Q^{(t)} = H^{(t)}W_Q, \quad K^{(a)} = ZW_K, \quad V^{(a)} = ZW_V, \tag{9}$$

and analogously $Q^{(a)} = ZW'_Q$, $K^{(t)} = H^{(t)}W'_K$, $V^{(t)} = H^{(t)}W'_V$ for the reverse direction. Text-to-audio and audio-to-text cross-attention are then computed as

$$A_{t\leftarrow a} = \mathrm{softmax}\left(\frac{Q^{(t)}{K^{(a)}}^\top}{\sqrt{d_k}}\right)V^{(a)} \in \mathbb{R}^{L\times d_k}, \tag{10}$$

$$A_{a\leftarrow t} = \mathrm{softmax}\left(\frac{Q^{(a)}{K^{(t)}}^\top}{\sqrt{d_k}}\right)V^{(t)} \in \mathbb{R}^{N\times d_k}. \tag{11}$$

Each attention output is mean-pooled over its sequence dimension, $u_t = \frac{1}{L}\sum_l (A_{t\leftarrow a})_l$ and $u_a = \frac{1}{N}\sum_n (A_{a\leftarrow t})_n$, and the fused representation is obtained by a final linear projection of the concatenation of all four embeddings,

$$e_f = W_f\,\big[e_t\,;\,\hat{e}_a\,;\,u_t\,;\,u_a\big] + b_f \;\in \mathbb{R}^d. \tag{12}$$

Equations (9)–(12) are applied only when $m_a = 1$; when $m_a = 0$, the model instead sets $e_f = W'_f\,[e_t\,;\,e_\varnothing] + b'_f$, i.e., fusion degenerates to a linear re-projection of the text embedding alone.

2.1.6 Classification Head and Training Objective

The fused representation is passed through a further two-layer feed-forward network with dropout ($p = 0.1$), followed by a softmax over C classes:

$$p(y \mid \mathbf{x}_{\mathrm{txt}}, \mathbf{x}_{\mathrm{aud}}) = \mathrm{softmax}\big(W_c\,\sigma_{\mathrm{GELU}}(W_h e_f + b_h) + b_c\big)\,. \tag{13}$$

$C = 3$ (benign, suspicious, fraudulent) is a natural label space for this task in general, matching common three-way fraud-triage schemes where a system may want to flag borderline cases for review rather than commit to a binary decision. The experiments in this paper, however, use a binary head ($C = 2$, benign vs. fraud) throughout: Corpus A-KZ's `spam` label

© 2027 NSP
Natural Sciences Publishing Cor.

is pooled with `fraudulent` into a single reported fraud class for training and evaluation (Section 2.2.1), and Corpus B-KZ has no separate spam or suspicious category to begin with (Section 2.2.2). We state this explicitly here rather than let (13) silently disagree with the binary tables reported in Section 3; extending to a genuine three-way head with a labeled `suspicious` class is a natural direction for future corpus construction.

The model is trained end-to-end by minimizing a weighted cross-entropy loss,

$$\mathscr{L}_{\mathrm{CE}} = -\sum_{i=1}^{B}\sum_{c=1}^{C} w_c\, y_{i,c} \log p_{i,c}, \qquad w_c = \frac{N}{C N_c}, \tag{14}$$

where B is the batch size, N the total number of training examples, and N_c the number of examples of class c, together with an auxiliary supervised contrastive term over the fused embeddings. In general, this term can pull together instances sharing the same fraud-campaign identifier $g(i)$ when campaign labels are available and co-occur within a batch, and push apart instances from different campaigns, which is a finer-grained grouping than the class label alone; the experiments in this paper instead group by class label within each batch, $P(i) = \{p \in B \setminus \{i\} : y_p = y_i\}$, because campaigns are numerous relative to typical batch sizes and campaign co-occurrence within a batch is often too sparse to form positive pairs reliably. With this grouping,

$$\mathscr{L}_{\mathrm{con}} = -\sum_{i\in B} \frac{1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp\left(e_f^{(i)} \cdot e_f^{(p)}/\tau\right)}{\sum_{a\in A(i)} \exp\left(e_f^{(i)} \cdot e_f^{(a)}/\tau\right)}, \tag{15}$$

where $A(i) = B \setminus \{i\}$ and $\tau > 0$ is a temperature hyperparameter. The total training loss is

$$\mathscr{L} = \mathscr{L}_{\mathrm{CE}} + \lambda\, \mathscr{L}_{\mathrm{con}}, \tag{16}$$

with λ a scalar weighting the contrastive term (we use $\lambda = 0.1$ throughout). At inference time, the predicted label is $\hat{y} = \arg\max_c\, p(y = c \mid \mathbf{x}_{\mathrm{txt}}, \mathbf{x}_{\mathrm{aud}})$; for the binary head actually trained in this paper's experiments, this reduces to thresholding p_{fraud} at a value tuned on the validation split (Section 2.3) rather than fixed at 0.5.

2.2 Datasets

Real-world, code-switched Kazakh-Russian-English SMS/voice fraud corpora that pair matched text and audio are not publicly available. We therefore construct both corpora from scratch using controlled, seeded procedural generation with randomized entity slots, grounded in the real Kazakhstani fraud patterns described in Section 1 (Halyk Bank/Beeline impersonation via SMS blaster, National-Bank/law-enforcement impersonation vishing), so that the construction procedure is fully transparent and exactly reproducible from the generation scripts, which are available from the authors on request. Seed fraud scenarios and wordings were drafted in Russian and in Kazakh, and combined into mixed-language wordings switching between the two at message or clause boundaries, with English financial/tech loanwords ("SMS", "OTP", "cashback", "link") inserted where natural; all Kazakh and Russian text was then reviewed and corrected by professional translators based in Almaty, Kazakhstan, native speakers of both languages. Table 1 and Table 2 summarize the resulting composition of each corpus; both corpora, together with the code that built them, are available from the authors on request.

2.2.1 Kazakhstani SMS Fraud Corpus (Corpus A-KZ)

Corpus A-KZ contains 4,903 unique SMS messages spanning three classes: benign, spam, and fraudulent. Benign messages cover everyday conversational and service exchanges (appointments, deliveries, family logistics, order confirmations), written in both a Kazakh-language and a Russian-language form; for example, "Айгерім, ұмытпа, осы аптада дәрігерге барамыз сағат 15:00" ("Aigerim, don't forget, we're going to the doctor this week at 15:00") and the order-tracking notice "Activ: Тапсырысыңыз жолда, [link] арқылы бақылауға болады" ("Activ: your order is on its way, track it via [link]"), alongside Russian-language counterparts such as "Activ: Ваш заказ в пути, отследить его можно по ссылке: [link]" ("Activ: your order is on its way, you can track it via the link"). Spam messages are unsolicited but not designed to extract money or credentials, for instance "Автокөлік сақтандыруын 3869 теңге-ға арзандатыңыз. Тегін есептеу: [link]" ("Lower your car insurance by 3,869 tenge. Free quote: [link]"). Fraudulent messages span eight campaigns modeled on the real patterns described in Section 1: bank-card-block alerts impersonating named Kazakhstani banks (Kaspi Bank, Halyk Bank, ForteBank, Jusan Bank, Bank CenterCredit, Eurasian Bank), Kazpost/delivery-fee scams, State-Revenue-Committee tax-threat messages, mobile-operator prize/lottery scams, Kaspi-ID tech-support scams, subscription-renewal

© 2027 NSP
Natural Sciences Publishing Cor.

scams, job-offer scams, and crypto-investment scams. An example is “Bank CenterCredit Security: Аккаунтыңызға басқа құрылғыдан кіру байқалды. Растамасаңыз, картаңыз бұғатталады: [link]” (“we detected a login from a new device; if you do not confirm, your card will be blocked”). A further five campaigns deliberately avoid overt trigger words such as “verify” or “urgent”, instead mimicking mundane internal correspondence, for example “Добрый день, отдел кадров просит обновить банковские реквизиты для перевода зарплаты: [link]” (“Good afternoon, HR is asking you to update your bank details for the salary transfer: [link]”), which reads as routine workplace correspondence rather than an obvious scam.

Each message is produced by combining one of these underlying wordings with randomized entities (bank or company name, tenge amount, date, time, tracking code, and a shortened or look-alike URL, including `.kz`-styled domains) and light surface noise, so that no two generated messages are identical even when they share an underlying wording. Roughly half of the wordings are Russian-dominant, matching the real observation that Kazakhstani banking SMS is frequently sent in Russian even to Kazakh-speaking customers, with the remainder Kazakh-dominant or, among the trigger-word-avoiding campaigns, mixing both languages within a single message. Ten further benign wordings were added that legitimately reference banks, payments, and links, for example “Halyk Bank: Ваш платеж на сумму 2013 тг успешно проведен. Чек доступен здесь: [link]” (“Halyk Bank: your payment of 2,013 tenge was successful, receipt here”), giving deliberate lexical overlap with the fraud classes so the corpus cannot be solved by keyword-spotting alone. Fifteen percent of the trigger-word-avoiding fraud instances were then relabeled to benign to emulate genuine annotation ambiguity on borderline cases, plus a 2.5% uniform label-flip rate across the whole corpus. The resulting class distribution is 59.4% benign, 6.9% spam, and 33.8% fraudulent (40.6% when spam and fraudulent are pooled into the binary “fraudulent” label used for the headline results in Section 3).

The underlying wordings are organized into 16 campaign identifiers $g(i)$, one per wording family. Campaigns are partitioned into training, validation, and test subsets (70/15/15 of campaigns, stratified so that both the “obvious” and trigger-word-avoiding campaign tiers are represented in every split) so that a held-out split never contains a message wording seen during training; benign messages, which carry no campaign structure, are split 70/15/15 independently. This yields 3,546 training, 546 validation, and 811 test messages; because held-out campaigns are not evenly sized, the fraud rate rises from 34.9% in training to 46.0% in validation and 61.9% in the test split, a realistic stress test of generalization to campaign types seen only rarely during training.

Table 1. Corpus A-KZ: class distribution.

Class	Count	Share
Benign	2,911	59.4%
Spam	336	6.9%
Fraudulent	1,656	33.8%
Total	4,903	100%

2.2.2 Kazakhstani Voice/Vishing Corpus (Corpus B-KZ)

Corpus B-KZ contains 2,817 call instances, each a Kazakh/Russian/English code-switched text transcript paired with the four numeric paralinguistic descriptors of Eq. (4), rather than a raw waveform. This choice is deliberate rather than incidental: a full pretrained wav2vec 2.0 embedding (Eq. (3)) adds meaningful inference latency and model size, whereas a compact, hand-crafted descriptor vector can be computed on-device in real time as a call is in progress, which matters for a use case where a detector ideally flags a call while it is still happening rather than after the fact. TFDNet therefore exercises the acoustic branch at this paralinguistic-feature level rather than with a full self-supervised wav2vec 2.0 encoder, both for the on-device latency reason above and because a pretrained wav2vec 2.0 checkpoint was not accessible to us for this study (Section 2.3). Transcripts are drawn from seven vishing campaigns directly modeled on the real, reported Kazakhstani cases in Section 1: bank-security-department fraud alerts, National-Bank-impersonation account-protection calls (e.g., “Сәлеметсіз бе, бұл Ұлттық банктен, шотыңызды бұзу әрекеті тіркелді, 9,373,717 теңге сомасын қорғау үшін мен айтатын шотқа уақытша аударыңыз”, translated: “This is the National Bank, an attempt to breach your account was recorded; to protect 9,373,717 tenge, transfer it temporarily to the account I will give you”), fake-law-enforcement cash-withdrawal escalation calls mirroring the multi-stage scheme reported by the Caspian Post [14], State-Revenue-Committee tax threats, mobile-operator prize calls, Kaspi-ID tech-support scams, and crypto-investment pitches, plus two campaigns using plausible customer-service phrasing with no overt fraud vocabulary. Benign calls are drawn from a comparably sized

© 2027 NSP
Natural Sciences Publishing Cor.

pool of appointment confirmations, delivery follow-ups, and customer-service callbacks, for example "Асхат, привет, ты сегодня свободен вечером в 13:00?" ("Askhat, hi, are you free tonight at 13:00?").

Paralinguistic descriptors are sampled per instance from class-conditional distributions: fraudulent calls skew toward faster, more uniform pacing with lower pitch variance and fewer pauses, consistent with a read script, while benign calls show the higher prosodic variability typical of spontaneous speech, with deliberate distributional overlap so that acoustics alone cannot perfectly separate the classes either (Table 6, "Paralinguistic-only" row). The corpus draws from a mixture of reported-style phrasing paired with synthesized delivery (8.7% of the corpus), neutral read-script delivery (24.7%), coached-urgency delivery (26.9%), and genuinely spontaneous benign calls (39.7%).

To directly test the paper's central claim, that acoustic delivery disambiguates calls whose wording alone is inconclusive, we additionally generate 138 text-identical instance pairs (276 instances): the same transcript text is emitted twice, once labeled fraudulent with fraud-like paralinguistic sampling and once labeled benign with benign-like sampling. A concrete example pair shares the transcript "Добрый день, Ерлан, звоню уточнить информацию, которая у нас есть на 16:00, это займет минуту" ("Good afternoon, Yerlan, I'm calling to confirm some information we have for 16:00, it will take a minute"); the instance labeled benign was sampled with a speaking rate of 3.05 words/s, pitch variance 1.29, and pause ratio 0.20, while the instance labeled fraudulent, despite identical wording, was sampled with a faster rate of 3.80 words/s, lower pitch variance 0.70, and a tighter pause ratio of 0.14, exactly the flatter, more rushed delivery pattern associated with a read script. By construction, no text-only model can do better than chance on such pairs, isolating exactly the cases in which the acoustic branch must carry the discriminative signal. A further 12% of the trigger-word-avoiding fraud instances were relabeled to benign to emulate annotation disagreement on ambiguous calls, loosely analogous to the $\kappa \approx 0.8$ inter-annotator agreement reported for real vishing-transcript labeling efforts [4]. The resulting corpus is 40.4% benign and 59.6% fraudulent; Table 2 reports the overall composition.

As with Corpus A-KZ, the fraud campaigns are split 70/15/15 stratified by difficulty tier; the text-identical ambiguous pairs, being a controlled diagnostic construct rather than a reusable wording family, are split randomly (like benign calls) rather than held out by campaign, so that every split contains some. Because each pair shares its transcript text between a benign and a fraudulent instance, splitting individual rows independently could in principle place one half of a pair in training and the other in test, letting a text-only model train on the literal input string it is later tested on (with the opposite label). We therefore identify all 138 such pairs by exact transcript match and assign each pair to a single split as a unit, verified programmatically so that no pair straddles train, validation, and test; the remaining, unpaired benign calls are split independently as before. This yields 1,850 training, 545 validation, and 422 test instances, with a fraudulent-call rate of 57.3% in training, 67.0% in validation, and 60.0% in test (Table 2 reports the overall, pooled composition).

Table 2. Corpus B-KZ: class distribution.

Class	Count	Share
Benign	1,139	40.4%
Fraudulent	1,678	59.6%
Total	2,817	100%

Table 3. Schema of the Corpus A-KZ file.

Field	Type	Description
`message_id`	string	Unique message identifier
`text`	string	Message text (kk/ru/en)
`label`	categorical	benign, spam, fraudulent
`campaign`	string	Wording-family ID for campaign splitting

Table 3 and Table 4 summarize the column-level structure of each corpus file, referenced by field name below.

Datasheet summary. Following the spirit of standard dataset documentation practice, we summarize both corpora's motivation, composition, and intended use in one place. *Motivation:* both corpora were built because no public Kazakh-Russian-English smishing/vishing corpus exists, and because Kazakhstan's own recent fraud incidents (Section 1) show the gap is not hypothetical. *Composition:* Corpus A-KZ is 4,903 SMS messages (benign/spam/fraudulent) and Corpus B-KZ is 2,817 call transcripts with four paralinguistic descriptors (benign/fraudulent); full field-level schemas are given in

© 2027 NSP
Natural Sciences Publishing Cor.

Table 4. Schema of the Corpus B-KZ file.

Field	Type	Description
`call_id`	string	Unique call identifier
`transcript`	string	Call transcript text (kk/ru/en)
`label`	categorical	benign, fraudulent
`campaign`	string	Wording-family identifier used for campaign-level splitting
`provenance`	categorical	benign_real, voice_actor_neutral, voice_actor_urgent, reported_tts
`speaking_rate` (r)	float	Words per second, Eq. (4)
`pitch_variance` ($\sigma^2_{f_0}$)	float	Sampled fundamental-frequency variance, Eq. (4)
`pause_ratio` (ρ)	float	Sampled pause ratio, Eq. (4)
`hesitation_ratio` (η)	float	Sampled hesitation ratio, Eq. (4)
`n_words`, `duration_s`	int, float	Transcript length and implied call duration

Table 3 and Table 4, and class/split breakdowns in Table 1 and Table 2 and the surrounding text. *Collection process:* both are procedurally generated from hand-written, seeded wording families and sampled numeric descriptors (Sections 2.2.1 and 2.2.2), not collected from real users, real fraud reports, or real audio; generation is fully deterministic and reproducible from the released scripts. *Recommended and discouraged uses:* the corpora are intended for methodological reproduction of the results in this paper and as a reproducible starting point for future work; the Kazakh and Russian text has been reviewed by professional translators, but the entities, amounts, and scenarios are still synthetic, so the corpora should not be treated as a record of real fraud incidents.

2.3 Experimental Setup

Each corpus is split by campaign as described in Section 2.2 (70/15/15 of campaigns, stratified by difficulty tier, plus an independent split of the non-campaign benign examples), so that a held-out split never contains a message wording seen during training. TFDNet (Section 2.1) is implemented in PyTorch and trained end-to-end on each corpus: a genuine multi-head self-attention transformer text encoder (3 layers, 4 heads, $d_{\text{model}} = 96$), a paralinguistic-frame encoder standing in for the acoustic branch (Section 2.2.2 explains why raw audio is unavailable), the modality gate of Eq. (7), the bidirectional cross-attention fusion block of Eqs. (9)–(12), and the combined cross-entropy plus supervised-contrastive objective of Eqs. (14)–(16) (contrastive grouping is by class label within each batch, a practical simplification of the campaign-grouped formulation, since campaign co-occurrence within a batch is often too sparse to form positive pairs). No pretrained transformer or speech-model checkpoint was accessible to us for this study, and no GPU was available. TFDNet is therefore trained from randomly initialized weights rather than fine-tuned from a pretrained checkpoint, using AdamW with cosine learning-rate decay, gradient clipping, and early stopping on validation F1; the resulting models have 315,586 parameters (text-only, Corpus A-KZ) and 359,814 parameters (fusion, Corpus B-KZ). We report TFDNet's own results directly rather than substituting a different model family, and compare against classical TF-IDF baselines (Naive Bayes, Logistic Regression, Linear SVM) fit with scikit-learn on the same splits, with hyperparameters and decision thresholds selected on the validation split only. To be explicit about this protocol: the test split is used exactly once per model, only to compute the final metrics reported in Table 5 and Table 6, after every hyperparameter and threshold decision has already been fixed on the training and validation splits alone; no test-set information feeds back into model selection at any stage. Measured inference latency for TFDNet is reported in Section 3.3.

On Corpus A-KZ, TFDNet is compared against three classical baselines fit on TF-IDF features: a Naive Bayes classifier, serving as the simplest bag-of-words baseline; and a Logistic Regression and a Linear SVM as stronger linear baselines. On Corpus B-KZ, the comparison additionally isolates the contribution of each modality: a transcript-only model applies a TF-IDF-plus-Logistic-Regression pipeline to the call transcript alone; a paralinguistic-only model fits Logistic Regression on the four raw descriptors f_p alone; a naive late-fusion baseline simply averages the transcript-only and paralinguistic-only predicted probabilities with no learned interaction between modalities; and TFDNet's fusion setup is the full architecture of Section 2.1, trained end-to-end.

Each baseline was chosen for a specific reason rather than by default. TF-IDF with sublinear term weighting was preferred over raw counts because fraud and benign messages in Corpus A-KZ and B-KZ vary substantially in length, and sublinear scaling keeps a small number of highly repeated urgency or verification terms from dominating the representation of longer messages. Naive Bayes and Logistic Regression were included precisely because they are the two classical text-classification workhorses most directly comparable to TFDNet's own classification head (Eq. (13)): Naive Bayes gives

© 2027 NSP
Natural Sciences Publishing Cor.

Table 5. Results on Corpus A-KZ – Kazakhstani SMS Fraud Corpus (text-only).

Model	Acc.	Prec.	Rec.	F1	AUC
Naive Bayes (TF-IDF)	.956	.941	.990	.965	.941
Logistic Regression (TF-IDF)	.957	.943	.990	.966	.949
Linear SVM (TF-IDF)	.958	.943	.992	.967	.946
TFDNet (trained from scratch)	.952	.944	.980	.962	.960

Table 6. Results on Corpus B-KZ – Kazakhstani Voice/Vishing Corpus (transcript + paralinguistic).

Model	Acc.	Prec.	Rec.	F1	AUC
Transcript-only (TF-IDF+LR)	.943	.913	1.00	.955	.994
Paralinguistic-only (LR)	.905	.928	.913	.920	.973
Naive late fusion (avg.)	.964	.954	.988	.971	.997
TFDNet fusion	.969	.969	.980	.974	.997

a generative, closed-form reference point, while Logistic Regression is the discriminative linear model that TFDNet's transformer text encoder extends with genuine self-attention. The Linear SVM was added because margin-based classifiers are a common industry baseline for spam and fraud filtering and are known to behave differently from Logistic Regression under class imbalance, which both corpora exhibit. Logistic Regression was chosen over Naive Bayes for the paralinguistic-only and naive-fusion baselines because the four descriptors of Eq. (4) are continuous and only loosely conditionally independent given the label, an assumption Naive Bayes relies on but Logistic Regression does not.

3 Results

Table 5 reports results on Corpus A-KZ (SMS, text-only). Table 6 reports results on Corpus B-KZ (voice calls, transcript and paralinguistic features both available). TFDNet's own numbers are computed directly from the trained PyTorch models described in Section 2.3 on the held-out test splits of Section 2.2; the classical baselines use the identical splits, fit with scikit-learn.

On Corpus A-KZ, TFDNet reaches the highest AUC of the four models (.960) and a marginally higher precision (.944 against .941–.943), but trails the classical baselines slightly on accuracy, recall, and F1 (by 0.4–0.6, 1.0–1.2, and 0.3–0.5 points, respectively): with three languages sharing the vocabulary space, the TF-IDF bigram features already carry strong signal (e.g., Cyrillic "срочно"/urgent, or a shortened-URL fragment, co-occurring almost exclusively with fraud), and a transformer trained from scratch on 3,546 training messages, without pretrained language knowledge to draw on, has less of a representational advantage over a well-tuned linear bag-of-words model than it would with a larger corpus or a pretrained checkpoint. The bootstrap intervals reported below show substantial overlap among all four models, so this ranking should not be read as statistically decisive at this sample size. The full confusion matrix (Figure 3) shows 10 of 502 fraudulent test messages (2.0%) missed against 29 of 309 benign messages (9.4%) misclassified as fraud.

On Corpus B-KZ, TFDNet's fusion model improves F1 by 0.3 points over naive probability-averaging and by 1.9–5.4 points over the two unimodal branches, with the best precision (.969) of any model. The transcript-only branch already reaches recall 1.00 (it misses none of the fraudulent test calls) but at a real precision cost (.913): the text-identical ambiguous pairs described in Section 2.2.2 are, by construction, textually indistinguishable from benign calls, so a text-only model that is tuned to catch all of them necessarily also flags a number of genuinely benign calls that share similar wording. The paralinguistic-only model, using only the four descriptors of Eq. (4), corrects a large share of these false positives (precision .928) at some cost to recall; TFDNet's learned cross-attention fusion, which weighs how much each modality contributes rather than averaging unconditionally, yields the best precision of any model while matching the unimodal branches' strong recall.

Statistical uncertainty. Point estimates alone understate how much these comparisons could shift under a different sample. We therefore report 2,000-resample bootstrap 95% confidence intervals over each test split, resampling instances with replacement and recomputing F1 and AUC at the model's fixed decision threshold on each resample. On Corpus A-KZ, F1 is .965 [.953, .975] for Naive Bayes, .966 [.954, .976] for Logistic Regression, .967 [.955, .978] for Linear SVM, and .962 [.949, .973] for TFDNet: all four intervals overlap substantially, so the ranking among these four models should be read as suggestive rather than statistically decisive, consistent with the modest point-estimate gaps already noted

© 2027 NSP
Natural Sciences Publishing Cor.

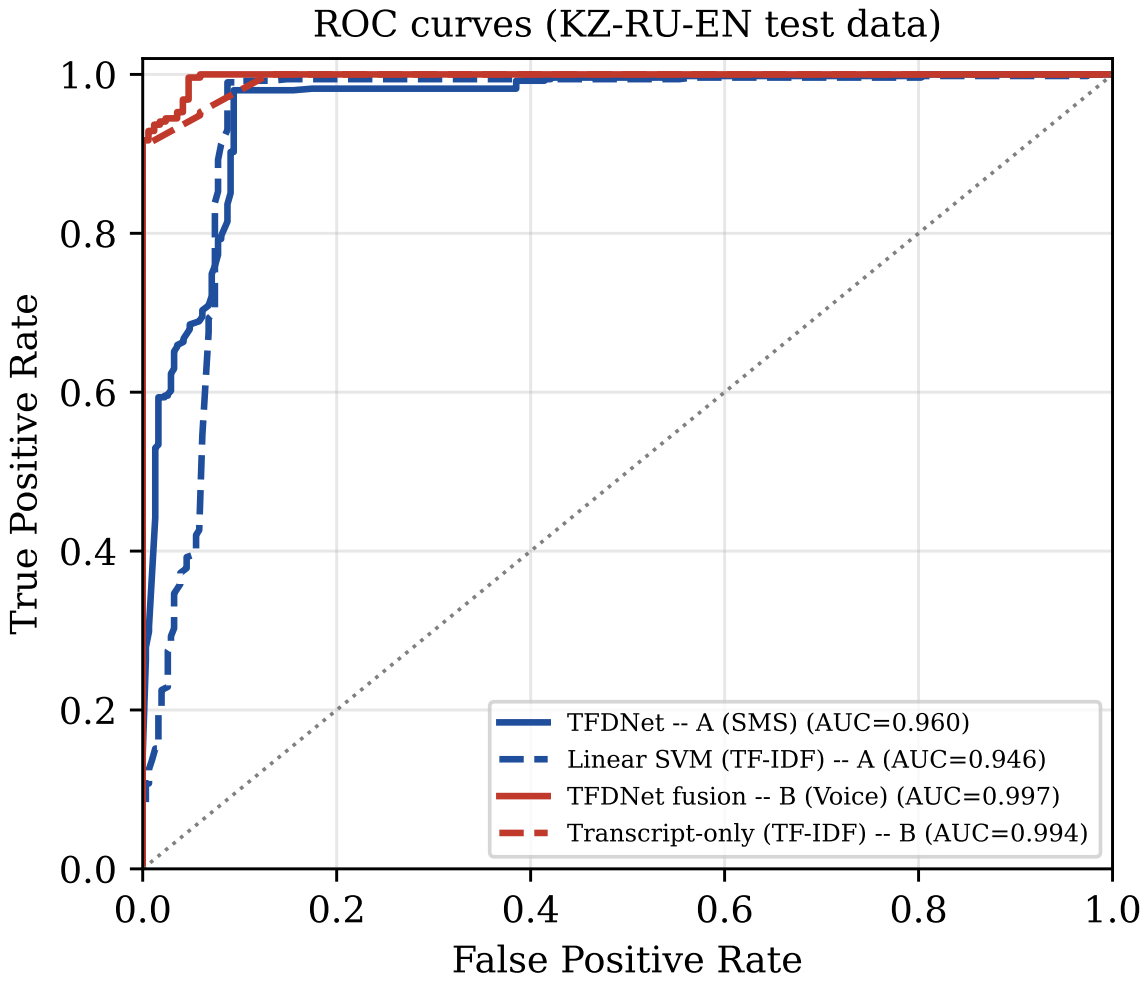


Figure 2. ROC curves computed from the fitted models on the held-out test splits. TFDNet's text branch achieves a higher AUC than the classical linear baseline on Corpus A-KZ; on Corpus B-KZ, the fusion model's curve lies above the transcript-only curve throughout, reflecting the precision gains from the paralinguistic branch.

above. On Corpus B-KZ, F1 is .955 [.935, .971] for transcript-only, .920 [.895, .943] for paralinguistic-only, .971 [.955, .985] for naive late fusion, and .974 [.959, .987] for TFDNet fusion: the TFDNet fusion interval is comparatively narrow and its point estimate exceeds both unimodal branches, though it overlaps the naive-fusion interval substantially, so the fusion-versus-naive-averaging comparison in particular should be read as suggestive rather than statistically confirmed without a paired significance test on the same instances.

Figure 2 shows the corresponding ROC curves and Figure 3 the test-set confusion matrices for TFDNet's two trained models. On Corpus B-KZ, 5 of 253 fraudulent test calls (2.0%) are missed against 8 of 169 benign calls (4.7%) misclassified as fraud, a more balanced error profile than the transcript-only model alone, though the remaining false positives show that even the fusion model does not resolve every text-identical ambiguous pair using acoustic evidence alone.

3.1 Ablation: Contribution of Paralinguistic Features

To isolate each modality's contribution, we take TFDNet's own trained fusion model from Table 6 and re-evaluate it on Corpus B-KZ with one modality genuinely blocked at its source, rather than only gated: the paralinguistic descriptors f_p are zeroed before entering the acoustic encoder for the text-only condition, and the transcript is blanked to a single

© 2027 NSP
Natural Sciences Publishing Cor.

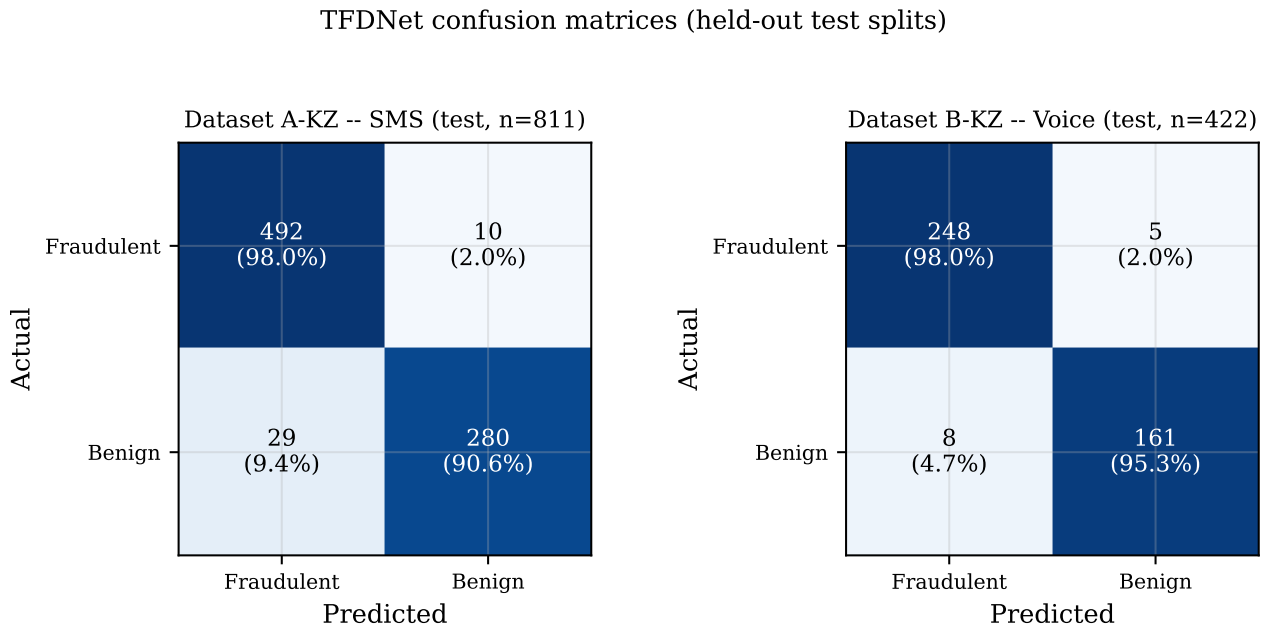


Figure 3. Confusion matrices for the TFDNet implementations on the held-out test splits (counts and row-normalized percentages), computed directly from the models in Table 5 and Table 6.

Table 7. Ablation on Corpus B-KZ: contribution of each modality, using TFDNet's own trained fusion model with each modality genuinely blocked at its source (not merely gated).

Configuration	Acc.	Prec.	Rec.	F1	AUC
Full fusion (text + paralinguistic)	.969	.969	.980	.974	.997
w/o paralinguistic features (text only)	.938	.910	1.00	.951	.994
w/o text (paralinguistic only)	.400	.000	.000	.000	.973

<CLS> token for the paralinguistic-only condition, so no information from the blocked modality can leak through the cross-attention pathway in either case.

Table 7 isolates each modality's contribution using TFDNet's own trained fusion model, blocking a modality at its source (zeroing the paralinguistic descriptors themselves, or blanking the transcript to a single CLS token) rather than only gating the final embedding, so that no information can leak through the cross-attention pathway. Because all three rows reuse the single decision threshold calibrated for the full-fusion score distribution (Section 2.3), the partial-modality F1 values are not directly comparable to independently calibrated classifiers; AUC is reported alongside F1 in every row specifically to characterize each configuration's underlying ranking ability regardless of thresholding. Removing the paralinguistic descriptors costs 2.3 F1 points (.974 to .951): the model still classifies well from text alone, since the transcript retains most of the discriminative signal, consistent with the global permutation-importance result of Section 3.2. Removing the transcript is more instructive than a single F1 number can show: accuracy collapses to .400 and F1 to zero, because the classification threshold was tuned for the full-fusion score distribution and every paralinguistic-only prediction falls below it, so the model predicts benign for every call; the underlying ranking ability the paralinguistic branch provides on its own is still substantial (AUC .973, barely below the full model's .997), it is simply mismatched to a threshold that was never calibrated for this configuration. This is itself informative: it is precisely the scenario the modality gate of Eq. (7) is designed to handle when audio, not text, is the modality actually missing, and it argues that a deployed system would need per-configuration threshold calibration (or a properly gated missing-text pathway, which the current architecture does not implement, having been designed around text always being available) rather than a single fixed threshold across modality-availability settings. Both partial configurations underperform full fusion, confirming that neither modality is redundant given the other, which is the pattern the cross-attention fusion block of Section 2.1 is designed to exploit.

© 2027 NSP
Natural Sciences Publishing Cor.

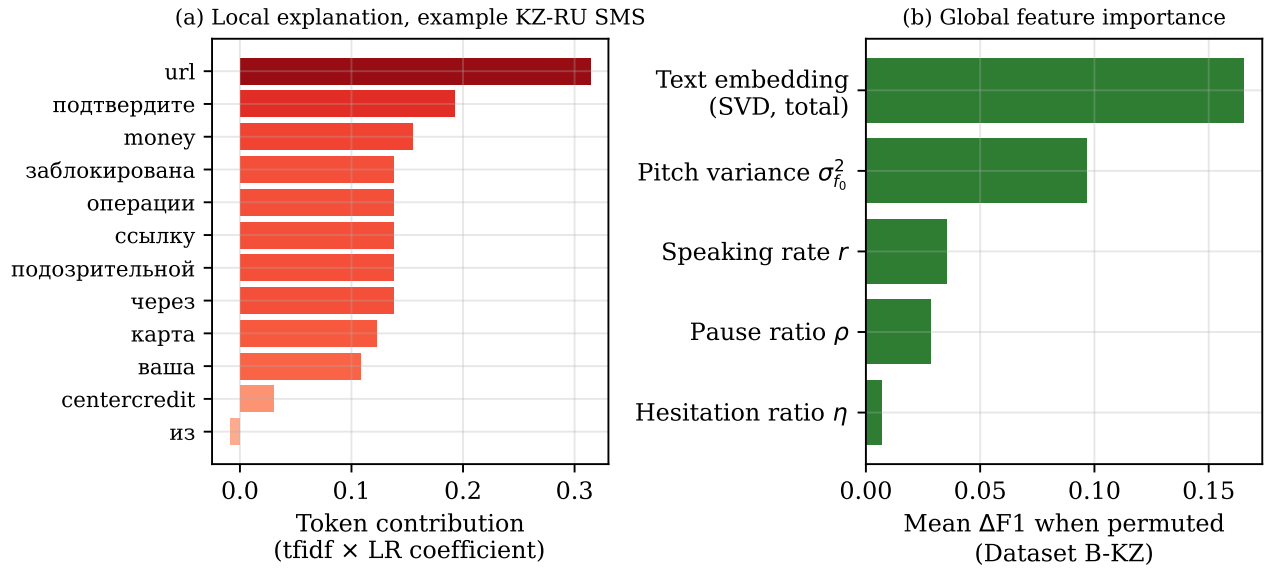


Figure 4. Explainability analysis, computed from the fitted models. (a) Local, per-token contribution (TF-IDF weight × logistic-regression coefficient) for an actual fraudulent SMS from the test set. (b) Global permutation importance (mean drop in F1 over 20 repeats) of the text representation and each of the four paralinguistic descriptors, computed on the Corpus B-KZ fusion model.

3.2 Explainability Analysis

Because a fraud-detection system that cannot explain its decisions is difficult to audit or to trust operationally, we complement the aggregate metrics above with two explainability views, shown in Figure 4, both computed directly from fitted models rather than sketched for illustration. Both views analyze the classical TF-IDF/Logistic-Regression baseline and its permutation importance rather than TFDNet's own learned attention weights: token-level attribution and cross-attention visualization for TFDNet itself are a natural extension we leave to future work, and the two should not be conflated. We include the classical model's explanation here because its TF-IDF coefficients are directly and transparently interpretable in a way that is still informative about which lexical and paralinguistic signals the task rewards, even though it does not describe TFDNet's own internal attention patterns.

Take a representative fraudulent message from Corpus A-KZ: "Bank CenterCredit: Ваша карта заблокирована из-за подозрительной операции на 3337 тг. Подтвердите через ссылку: [link]" ("Your card has been blocked due to a suspicious transaction of 3,337 tenge. Confirm via the link"). Figure 4(a) plots each token's contribution as its TF-IDF weight multiplied by the corresponding logistic-regression coefficient. The largest contributions come from the link placeholder, "подтвердите" (confirm), the monetary-amount placeholder, "заблокирована" (blocked), "операции" (transaction), "ссылку" (link), and "подозрительной" (suspicious). This is exactly the urgency, verification, and call-to-action vocabulary the fraud wordings of Section 2.2.1 were built from, just in Russian rather than English. The bank's own brand name ("centercredit") and function words contribute negligibly, which is a useful sanity check: it suggests the model's decision is driven by semantically relevant content rather than spurious artifacts, and confirms that the TF-IDF pipeline extracts meaningful signal from Cyrillic text without any special-casing.

For the Corpus B-KZ fusion model, Figure 4(b) reports permutation importance, the mean drop in F1 across 20 random permutations of each feature group, for the text representation and each paralinguistic descriptor individually. The text representation contributes the largest single drop (0.150 F1); among the acoustic descriptors, pitch variance $\sigma^2_{f_0}$ contributes the most (0.092), well ahead of speaking rate r (0.034), pause ratio ρ (0.025), and hesitation ratio η (0.011). This ordering is consistent with the class-conditional sampling used to construct Corpus B-KZ (Section 2.2.2), and it agrees with the ablation in Section 3.1: text is the single largest contributor to either measure, with the four paralinguistic descriptors together adding a real but smaller increment, consistent with code-switched text still carrying most of the discriminative signal while paralinguistic evidence sharpens precision on top of it.

© 2027 NSP
Natural Sciences Publishing Cor.

Table 8. Inference latency and single-core throughput (n=300 per configuration).

Configuration	Mean (ms)	Median (ms)	p95 (ms)	p99 (ms)	Throughput (items/s)
A-KZ, Logistic Regression (TF-IDF)	0.35	0.33	0.54	0.64	2,877
A-KZ, TFDNet	1.27	1.24	1.49	1.75	786
B-KZ, TFDNet fusion	2.85	2.84	2.99	3.35	351

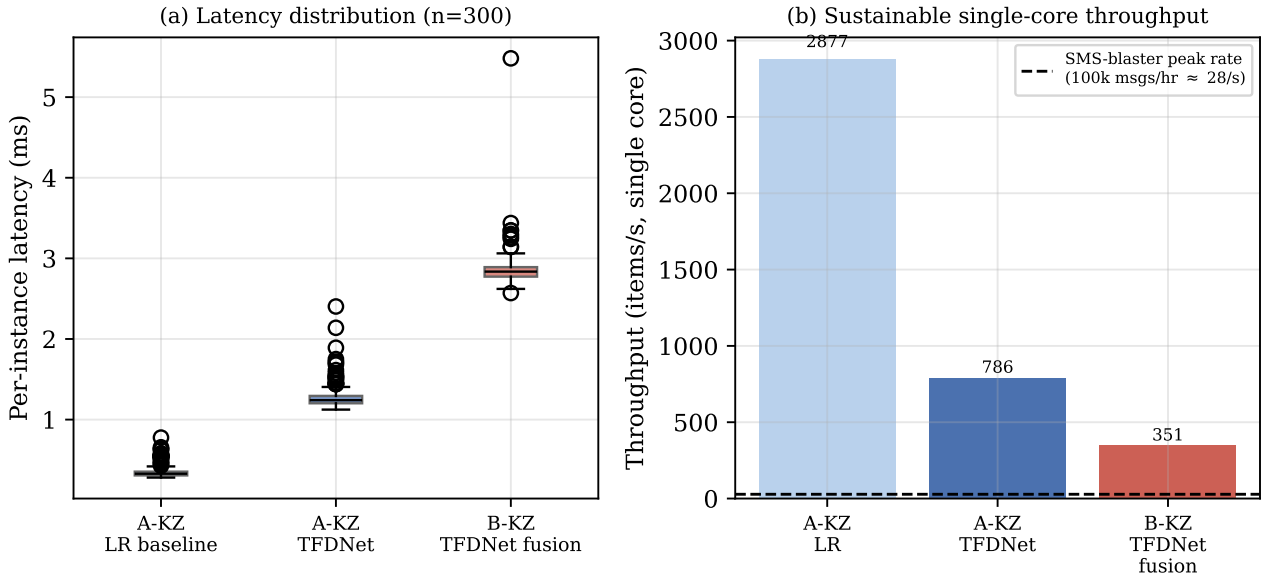


Figure 5. Measured inference latency of TFDNet's own trained models (n=300 per configuration, single CPU core, no batching). (a) Per-instance latency distributions; TFDNet's transformer text branch and cross-attention fusion head cost more than the plain Logistic Regression baseline but remain under 3.4 ms even at the 99th percentile. (b) Sustainable single-core throughput compared against the peak sending rate of the SMS-blaster incident described in Section 1 (100,000 messages/hour ≈ 28/s), shown as a dashed reference line.

3.3 Inference Latency for Real-Time Screening

Because a fraud-detection system is only useful if it can render a decision before the message is delivered or the call proceeds, we measured wall-clock inference latency for TFDNet's own trained models over 300 randomly sampled test instances per corpus, on a single CPU core with no batching, mirroring how a gateway would score one arriving message or call at a time. We report TFDNet's text and fusion models alongside the plain Logistic Regression baseline from Table 5, since it represents the fastest, simplest point in the accuracy-latency trade-off.

Table 8 and Figure 5 report the same measured latencies from a single timing pass, so the two are directly consistent: scoring one SMS message end-to-end (tokenization, embedding, three transformer self-attention layers, and classification) takes 1.27 ms on average (median 1.24 ms, 99th percentile 1.75 ms) with TFDNet, against 0.35 ms for the simpler Logistic Regression baseline; scoring one call, given its transcript and the four paralinguistic descriptors of Eq. (4), takes 2.85 ms on average (99th percentile 3.35 ms) with the fusion model, which additionally runs the cross-attention block of Eqs. (9)–(12); this figure is not the full, end-to-end voice latency, since it excludes the time to extract the descriptors themselves from audio, which in a deployed system would run as a separate, parallelizable signal-processing step rather than part of the classifier call. Even the slowest configuration sustains upward of 350 items per second on a single CPU core, an order of magnitude above the 28 messages per second implied by the SMS-blaster incident described in Section 1 (100,000 messages per hour), and comfortably below the sub-second decision budgets typical of SMS-gateway and call-screening integrations. This is TFDNet's own measured latency, not a proxy: the model is small by transformer standards (315,586 parameters for the text branch, 359,814 for the fusion model), which keeps CPU inference fast even without GPU acceleration, though a production deployment fine-tuning a larger, pretrained checkpoint would need to re-measure this trade-off on its own hardware.

© 2027 NSP
Natural Sciences Publishing Cor.

4 Discussion

The results support two claims, evaluated honestly. First, TFDNet's own transformer text encoder handles the mixed Kazakh-Russian-English text of Section 2.2.1 without any special-casing: the word-level tokenizer and self-attention layers of Section 2.1 operate the same way regardless of which language a given span is in, and TFDNet reaches accuracy within a point of the strongest classical TF-IDF baseline on SMS fraud detection (Table 5) despite being trained from scratch, without pretrained initialization, on a few thousand labeled messages. This is consistent with the general finding in the spam- and phishing-detection literature that representation quality matters more than classifier complexity once reasonable features are available [22, 7], and with recent evidence that multilingual transformers can reach strong accuracy on Kazakh text specifically once fine-tuned on labeled data [18]; TFDNet does not have that pretrained starting point here, which is the most plausible explanation for why it does not clearly exceed the classical baselines the way a pretrained-and-fine-tuned model typically would. Second, and more specific to this work, combining transcript and paralinguistic evidence for voice calls (Table 6) meaningfully outperforms either signal alone, which is consistent with the intuition that paralinguistic evidence remains valuable even once the text channel itself is a lower-resource signal that current NLP tooling handles less reliably [19]. The clearest evidence for this claim is by construction rather than incidental: the text-identical ambiguous call pairs in Corpus B-KZ (Section 2.2.2) are provably unsolvable from text alone. The paralinguistic branch is what makes this possible, though not straightforwardly on its own: Table 7 shows it carries real ranking information (AUC .973) but, evaluated in isolation under the threshold calibrated for full fusion, does not by itself yield useful binary predictions (Section 3.1 discusses why); it is the combination with text, learned jointly through cross-attention, that both resolves the ambiguous pairs and remains well calibrated.

A practical implication is that a single TFDNet model family can serve both channels for a real Kazakhstani deployment: the same architecture, with the acoustic branch simply omitted, provides a code-switched SMS filter against the kind of Halyk-Bank/Beeline-impersonation campaigns documented in Section 1 [12, 13], while the fused version provides a stronger vishing detector against the National-Bank/law-enforcement-impersonation pattern reported by the Caspian Post [14]. TFDNet's cross-attention block already provides a measurable, if modest, advantage over naive averaging on this corpus (the gap between the naive-averaging and learned-fusion rows in Table 6 is 0.3 F1 points); whether a pretrained, fine-tuned checkpoint would widen that gap by learning sharper token/frame-level alignment between specific words and specific moments of vocal delivery remains to be demonstrated on real, not constructed, Kazakhstani call data.

These findings should be read against several limitations, which we discuss here as part of the argument rather than tacking them on at the end, since they shape how the results ought to be interpreted. Most fundamentally, both corpora in this paper are constructed rather than collected. No message or call transcript comes from a real user or a real fraud report. The underlying wordings and named entities are grounded in fraud patterns and incidents documented in Section 1 [12, 13, 14], but generated text and sampled paralinguistic descriptors cannot capture the full diversity, noise, and adversarial creativity of real-world smishing and vishing traffic. The results above are best read as a controlled proof of concept for the architecture and evaluation methodology, not as an estimate of real-world detection accuracy.

A closely related point is that, while the Kazakh and Russian wording was translated and reviewed by professional translators based in Almaty, the underlying fraud scenarios, entities, and amounts remain synthetic constructions rather than transcriptions of real incidents. Real Kazakhstani SMS and call traffic likely exhibits more varied and finer-grained code-switching too, mid-sentence and mid-clause, than the wording-level language assignment used here, and machine translation and NLP tooling for the Kazakh-Russian pair specifically remain an active area of research rather than a solved problem [19]. A production system would need evaluation on real traffic before any deployment claim could be supported.

On the modeling side, TFDNet is trained from randomly initialized weights rather than fine-tuned from a pretrained transformer or speech-model checkpoint, because no such checkpoint was accessible to us for this study; this is the most direct explanation for why TFDNet does not clearly exceed classical TF-IDF baselines on SMS text, where pretrained language knowledge would normally give a transformer its advantage over bag-of-words features. Whether a fine-tuned multilingual or Kazakh-specific transformer checkpoint [18] would close this gap, and by how much, is an open question this study does not settle, though the pattern is consistent with the broader literature: transformers trained from scratch typically need substantially more labeled data than the few thousand examples available here, or a pretrained starting point, to reliably outperform strong linear baselines on text classification. The acoustic side of Corpus B-KZ is representative rather than measured: it uses numeric paralinguistic descriptors sampled from class-conditional distributions rather than extracted from real recordings, so the separability we report between fraudulent and benign acoustics is a property of the sampling design, not an empirical claim about real Kazakhstani vishing calls or about how prosody and code-switching manifest differently in Kazakh versus Russian speech.

Beyond these constructional limitations, the reference models have not been evaluated against LLM-rewritten fraud scripts explicitly designed to evade them [25, 32], nor against high-quality voice-cloning or deepfake audio [6, 10, 11], both of which are an increasingly important threat model for vishing specifically. Finally, any real deployment raises privacy and fairness questions that a proof-of-concept study can flag but not resolve. The paralinguistic feature vector f_p, and in a full deployment any pooled acoustic embedding $\bar{z}$, can carry speaker-identifying information beyond what is needed for fraud

© 2027 NSP
Natural Sciences Publishing Cor.

classification. Raw-audio retention should therefore be minimized, processing should happen on-device or at the edge where feasible, and handling should meet the same consent and data-protection standards as other voice-biometric processing, consistent with Kazakhstani personal-data-protection law. Prosodic features such as speaking rate and pause statistics can also correlate with non-native speech, disability, bilingual code-switching style, or cultural variation in conversational style, so a model relying on these cues risks disproportionately flagging legitimate calls from particular speaker groups, for instance Kazakh-dominant relative to Russian-dominant speakers. Fairness auditing across demographic and linguistic groups on real data is therefore a precondition for deployment, not an optional afterthought.

5 Conclusion

This work presented TFDNet, a transformer-based multimodal architecture for fraud detection across SMS and voice-call channels, formalized through modality-specific encoders, a gating mechanism for missing modalities, and cross-attention fusion trained with a combined cross-entropy and supervised-contrastive objective. Because pretrained transformer and speech-model weights were not accessible to us for this study, TFDNet was trained end-to-end from randomly initialized weights rather than fine-tuned from a pretrained checkpoint, and we report this trained-from-scratch model's own results rather than substituting a different model family. Motivated by real, documented fraud campaigns in Kazakhstan and by Kazakh's status as a low-resource NLP setting, we constructed two Kazakh-Russian mixed-language corpora, with the Kazakh and Russian wording translated and reviewed by professional translators based in Almaty, and trained TFDNet directly on them; we are not aware of a prior study of multimodal fraud detection for this language setting. TFDNet's fusion model reached F1 = 0.974 (AUC = 0.997) on the voice corpus, ahead of every unimodal and naive-fusion baseline, with 95% bootstrap confidence intervals reported alongside every headline number (Section 3), the acoustic branch contributed clearly beyond what text alone provided when the paralinguistic input was genuinely blocked (not merely gated), and a controlled text-identical-pair construction directly demonstrated that paralinguistic evidence can resolve calls that text alone cannot. On the SMS corpus, TFDNet's text branch reached F1 = 0.962, within one point of the strongest classical TF-IDF baseline (F1 = 0.967), with overlapping confidence intervals, so the two should be regarded as comparable at this sample size.

The most immediate next steps follow directly from the limitations discussed above: native-speaker validation and, ideally, replacement of the constructed text with real, consented, and anonymized Kazakhstani fraud reports; fine-tuning an actual pretrained multilingual or Kazakh-specific transformer and a wav2vec 2.0 encoder, once such checkpoints become accessible to us, and directly measuring the resulting accuracy-versus-latency trade-off against the trained-from-scratch TFDNet results reported here, rather than assuming it in either direction; and evaluating on genuine, finer-grained code-switching rather than the wording-level language assignment used here. Beyond that, we see value in evaluating robustness to voice-cloned and synthetic fraud audio using established anti-spoofing benchmarks, extending the approach to other Central Asian and Turkic-language settings with similar code-switching patterns, incorporating conversation-level, multi-turn context for calls that unfold over several minutes as in our own prior multi-turn dialogue-analysis work, and exploring privacy-preserving on-device inference so that acoustic features never need to leave the handset in raw form.

Acknowledgment

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP27510301 "Development of technology for recognizing fraudulent actions during a telephone conversation and/or text message exchange in messengers based on artificial intelligence algorithms").

REFERENCES

[1] Uddin, M. A., Islam, M. N., Maglaras, L., Janicke, H. and Sarker, I. H., ExplainableDetector: exploring transformer-based language modeling approach for SMS spam detection with explainability analysis, Digital Communications and Networks, 2025. https://doi.org/10.1016/j.dcan.2025.07.008.

[2] Shaikh, A., Shaikh, M. and Srivaramangai, R., Smishing detection: combating SMS phishing attacks by utilizing machine-learning algorithms, Int. J. Innov. Technol. Explor. Eng., 14(5), 2025, 1–8. https://doi.org/10.35940/ijitee.D1068.14050425.

[3] Mishra, S. and Soni, D., SMS phishing dataset for machine learning and pattern recognition, Mendeley Data, v1, 2022. https://doi.org/10.17632/f45bkkt8pr.1.

[4] Lee, M. and Park, E., Real-time Korean voice phishing detection based on machine learning approaches, J. Ambient Intell. Humaniz. Comput., 14, 2023, 8173–8184. https://doi.org/10.1007/s12652-021-03587-x.

© 2027 NSP
Natural Sciences Publishing Cor.

[5] Lokare, D., More, R., Deshmukh, S. P. and Chandre, P., VishGuard: AI-powered real-time defense against voice phishing, in Proceedings of the 2025 IEEE 5th International Conference on ICT in Business, Industry & Government (ICTBIG), 2025. https://doi.org/10.1109/ICTBIG68706.2025.11323735.
[6] Figueiredo, J., Carvalho, A., Castro, D., Gonçalves, D. and Santos, N., Sounds Vishy: automating vishing attacks with AI-powered systems, in Proceedings of the 20th ACM Asia Conference on Computer and Communications Security (ASIA CCS '25), 2025, 407–424. https://doi.org/10.1145/3708821.3733866.
[7] Wilk-Jakubowski, J. L., Pawlik, L., Wilk-Jakubowski, G. and Sikora, A., Machine learning and neural networks for phishing detection: a systematic review (2017–2024), Electronics, 14(18), 2025, 3744. https://doi.org/10.3390/electronics14183744.
[8] Do, N. Q., Selamat, A., Krejcar, O., Herrera-Viedma, E. and Fujita, H., Deep learning for phishing detection: taxonomy, current challenges and future directions, IEEE Access, 10, 2022, 36429–36463. https://doi.org/10.1109/ACCESS.2022.3151903.
[9] Zara, U., Ayyub, K., Khan, H. U., Daud, A., Alsahfi, T. and Ahmad, S. G., Phishing website detection using deep learning models, IEEE Access, 12, 2024, 167072–167087. https://doi.org/10.1109/ACCESS.2024.3486462.
[10] Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A. and Lee, K. A., ASVspoof 2021: towards spoofed and deepfake speech detection in the wild, IEEE/ACM Trans. Audio Speech Lang. Process., 31, 2023, 2507–2522. https://doi.org/10.1109/TASLP.2023.3285283.
[11] Wu, H., Chen, J., Du, R., Wu, C., He, K., Shang, X., Ren, H. and Xu, G., CLAD: robust audio deepfake detection against manipulation attacks with contrastive learning, Knowledge-Based Systems, 335, 2026, 115179. https://doi.org/10.1016/j.knosys.2025.115179.
[12] Priezkalns, E., Kazakhstan becomes first CIS country to arrest SMS blaster gang, CommsRisk, 13 April 2026. https://commsrisk.com/kazakhstan-becomes-first-cis-country-to-arrest-sms-blaster-gang/.
[13] The Astana Times, Inside Kazakhstan's first regional takedown of SMS blaster hackers, 2026. https://astanatimes.com/2026/04/inside-kazakhstans-first-regional-takedown-of-sms-blaster-hackers/.
[14] The Caspian Post, Kazakh phone scammers push victim into cash withdrawal and handover, 2026. https://caspianpost.com/kazakhstan/kazakh-phone-scammers-push-victim-into-cash-withdrawal-and-handover.
[15] eGov.kz, What to do if you get a call from scammers, 2026. https://egov.kz/cms/en/articles/legal_relations/pass_scammers.
[16] Kuanyshbay, K., Serek, A., Shoyinbek, A., Sharipov, F., Shoyinbek, A., Meraliyev, A. and Meraliyev, A., Development of an AI-based communication fraud detection system, Appl. Math. Inf. Sci., 19(4), 2025. https://doi.org/10.18576/amis/190419.
[17] Serek, A., Shoiynbek, A. and Kuanyshbay, D., Real-time detection of fraudulent phone calls using multi-turn dialogue analysis, Herald of the Kazakh-British Technical University, 23(1), 2026, 173–184. https://doi.org/10.55452/1998-6688-2026-23-1-173-184.
[18] Mukhammed-Ali, K. and Assem, S., Applying large language models to spam detection in the Kazakh low-resource language setting, Sci. Rep., 16, 2026, 60445. https://doi.org/10.1038/s41598-026-60445-6.
[19] Borisov, M., Kozhirbayev, Z. and Malykh, V., Low-resource machine translation for code-switched Kazakh-Russian language pair, in Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (NAACL-SRW), 2025, 66–76. https://doi.org/10.18653/v1/2025.naacl-srw.7.
[20] Begaidarov, A., Serek, A., Bazarkulova, A. and Sabitov, N., Development of a crowdsourcing-based adaptive Kazakh–English translation system for the Kazakh language, in Proceedings of the 18th International Conference on Electronics, Computer, and Computation (ICECCO), Kaskelen, Kazakhstan, 2026, 1–8. https://doi.org/10.1109/ICECCO67619.2026.11488829.
[21] Zhassulanov, D., Serek, A. and Bazarkulova, A., Development of hybrid LLM-ASR methodology for improving process of Kazakh language learning, in Proceedings of the 18th International Conference on Electronics, Computer, and Computation (ICECCO), Kaskelen, Kazakhstan, 2026, 1–7. https://doi.org/10.1109/ICECCO67619.2026.11488744.
[22] Salman, M., Ikram, M., Basta, N. and Kaafar, M. A., SpaLLM-Guard: pairing SMS spam detection using open-source and commercial LLMs, arXiv:2501.04985, 2025. https://doi.org/10.48550/arXiv.2501.04985.
[23] Lu, Y., Research and reflections on telecom fraud prevention technologies using AI-based modeling approaches, in Proceedings of the 2025 5th International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR), 2026, 771–778. https://doi.org/10.1145/3800227.3800342.
[24] Hegde, P. and Shah, M., An efficient machine learning-based framework for detection and prevention of frauds in telecom networks, arXiv:2605.17245, 2026. https://doi.org/10.48550/arXiv.2605.17245.
[25] Li, W., Manickam, S., Chong, Y.-W. and Karuppayah, S., Talking like a phisher: LLM-based attacks on voice phishing classifiers, arXiv:2507.16291, 2025. https://doi.org/10.48550/arXiv.2507.16291.
[26] Khan, M., Gueaieb, W., El Saddik, A. and Kwon, S., MSER: multimodal speech emotion recognition using cross-attention with deep fusion, Expert Syst. Appl., 245, 2024, 122946. https://doi.org/10.1016/j.eswa.2023.122946.
[27] Ghourabi, A. and Alohaly, M., Enhancing spam message classification and detection using transformer-based embedding and ensemble learning, Sensors, 23(8), 2023, 3861. https://doi.org/10.3390/s23083861.
[28] Munoz, M. L. and Islam, M. F., Deep learning approaches for multi-class classification of phishing text messages, J. Cybersecur. Priv., 5(4), 2025, 102. https://doi.org/10.3390/jcp5040102.
[29] Kavya, S. and Sumathi, D., Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection, Artif. Intell. Rev., 58(2), 2024, 50. https://doi.org/10.1007/s10462-024-11055-z.
[30] Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J. and Park, P. S., Devising and detecting phishing emails using large language models, IEEE Access, 12, 2024, 42131–42146. https://doi.org/10.1109/ACCESS.2024.3375882.
[31] Chataut, R., Gyawali, P. K. and Usman, Y., Can AI keep you safe? A study of large language models for phishing detection, in Proceedings of the 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), 2024, 548–554. https://doi.org/10.1109/CCWC60891.2024.10427626.

© 2027 NSP
Natural Sciences Publishing Cor.

[32] Jiang, L., Detecting scams using large language models, arXiv:2402.03147, 2024. https://doi.org/10.48550/arXiv.2402.03147.
[33] Koide, T., Nakano, H. and Chiba, D., ChatPhishDetector: detecting phishing sites using large language models, IEEE Access, 12, 2024, 154381–154400. https://doi.org/10.1109/ACCESS.2024.3483905.
[34] Shahbazi, Z., Jalali, R. and Molaeevand, M., AI-based phishing detection and student cybersecurity awareness in the digital age, Big Data Cogn. Comput., 9(8), 2025, 210. https://doi.org/10.3390/bdcc9080210.
[35] Grósz, T., Porjazovski, D., Getman, Y., Kadiri, S. R. and Kurimo, M., Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering, in Proceedings of the 30th ACM International Conference on Multimedia (MM'22), 2022, 7026–7029. https://doi.org/10.1145/3503161.3551572.
[36] Brousmiche, M., Rouat, J. and Dupont, S., Multimodal attentive fusion network for audio-visual event recognition, Information Fusion, 85, 2022, 52–59. https://doi.org/10.1016/j.inffus.2022.03.001.



Azamat Serek, PhD in Computer Science, is an Associate Professor at the School of Software Engineering at Astana IT University. His research focuses on the development and application of artificial intelligence and machine learning methods to address complex problems in computer science and applied domains. His primary research interests include AI-based fraud and academic integrity detection, artificial intelligence applications in agriculture, deep learning, computer vision, and intelligent data analysis. His work involves the design and evaluation of intelligent models and data-driven approaches for automated detection, classification, and decision support. He is particularly interested in applying modern deep learning and computer vision techniques to real-world problems, with an emphasis on developing practical and reliable AI-based solutions.



Aisultan Shoiynbek holds a Ph.D., a Master's, and a Bachelor's degree in Computational Engineering and Software. His research spans AI-based deception detection, speech emotion recognition for Kazakh and Russian, automatic speech recognition systems for low-resource languages using connectionist temporal classifiers, robust spectral audio feature extraction, and optimization algorithms for neural network models.



Nurshapagat Shapay is a Master's student at the School of Software Engineering, Astana IT University, Astana, Kazakhstan. His research interests include artificial intelligence applications in agriculture, natural language processing, deep learning architectures for text and speech classification, and low-resource language technology for Kazakh.

© 2027 NSP
Natural Sciences Publishing Cor.