Developing Defensive Techniques against Phishing Attacks

Hitesh Choudhary
Supervisor: Dr. Gaurav Varshney
Department of Computer Science and Engineering
Indian Institute of Technology Jammu
2022ucs0092@iitjammu.ac.in

October 8, 2025

Contents

1 Motivation

Phishing remains a pervasive and rapidly evolving cyber threat, exploiting not only email and SMS but increasingly voice channels to defraud unsuspecting victims. In the midterm report, broad phishing trends were analyzed and the need to anticipate emerging modalities was identified. The Anti-Phishing Working Group's quarterly *Phishing Activity Trends Reports* for 2024 chronicle persistent growth in voice-based attacks: Q1 documented widespread hybrid vishing campaigns [1], Q2 reported 877,536 incidents [2], Q3 observed 932,923 attacks with a 28% increase over Q2 [3], and Q4 recorded 989,123 incidents [4]. Unlike traditional email phishing, vishing unfolds in real time via social-engineering phone calls, which complicates detection and prevention. This alarming trend motivated a pivot: to design a lightweight, client-side system that continuously monitors live calls, transcribes speech to text, scans for malicious markers, and issues immediate alerts—thereby empowering users against the growing threat of voice-based scams.

2 Literature Review

Voice phishing detection has evolved significantly with hybrid deep learning architectures. Boussougou and Park [5] demonstrated state-of-the-art performance using an attention-based 1D CNN-BiLSTM model on the Korean Call Content Vishing (KorCCVi v2) dataset, achieving 99.32% accuracy and 99.31% F1-score. Their methodology integrates five key components:

1. **Preprocessing Pipeline**:
    - Voice-to-text conversion via Google Cloud Speech-to-Text API
    - Tokenization using MeCab-ko morphological analyzer
    - Removal of personal information, digits, and Korean stopwords
2. **Embedding Layer**:
    - FastText word vectors (300 dimensions)
    - Subword tokenization for rare/out-of-vocabulary words
    - Pretrained on Korean corpora
3. **1D CNN Architecture**:
    - 32 convolutional filters

- Kernel size = 3, max pooling (size = 2)
- Spatial dropout (rate = 0.2)

4. **BiLSTM Configuration**:
   - Forward/backward LSTMs with 64 and 32 hidden units
   - Dropout (rate = 0.1)
   - Hidden state concatenation: $h_t = [h_t^f, h_t^b]$
5. **Hierarchical Attention**:
   - Word-level attention: $\alpha_{it} = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)}$
   - Sentence-level attention: $v = \sum_i \alpha_i h_i$
   - Context vectors computed via BiGRU layers

The model outperformed baseline architectures including standalone 1D CNN (98.29% F1-score), BiLSTM (84.32%), and a non-attentive 1D CNN-BiLSTM hybrid (98.97%). While KoBERT-based models achieved marginally higher accuracy (99.60%) [5], they required extensive fine-tuning and lacked the 1D CNN-BiLSTM's computational efficiency (137-second training time on NVIDIA RTX 3090 vs. 143 seconds for BiLSTM). Federated learning approaches preserved privacy but omitted critical metrics like precision/recall. This work's combination of FastText subword embeddings, hierarchical attention, and optimized hyperparameters (10 epochs, batch size 64, Adam optimizer) positions it as a robust solution for real-time Korean voice phishing detection.

3 Goal

The goal is to develop a real-time voice phishing (vishing) detection system focused exclusively on phone call scams. A mobile application will continuously monitor live calls, transcribe speech to text using on-device processing, and analyze transcripts through a backend ensemble of specialized deep learning models. Each model will target distinct vishing strategies—such as impersonation scams (e.g., fake bank officials), urgency-driven fraud (e.g., fake emergencies), or financial pretexting (e.g., fake investment offers). The system will correlate outputs from all models; if any detect malicious intent, it will trigger an immediate in-call alert to the user, providing specific risk descriptors (e.g., "Detected impersonation scam: caller claims unverified authority"). This narrow focus on vishing subtypes ensures high precision in identifying evolving social engineering tactics unique to voice-based attacks.

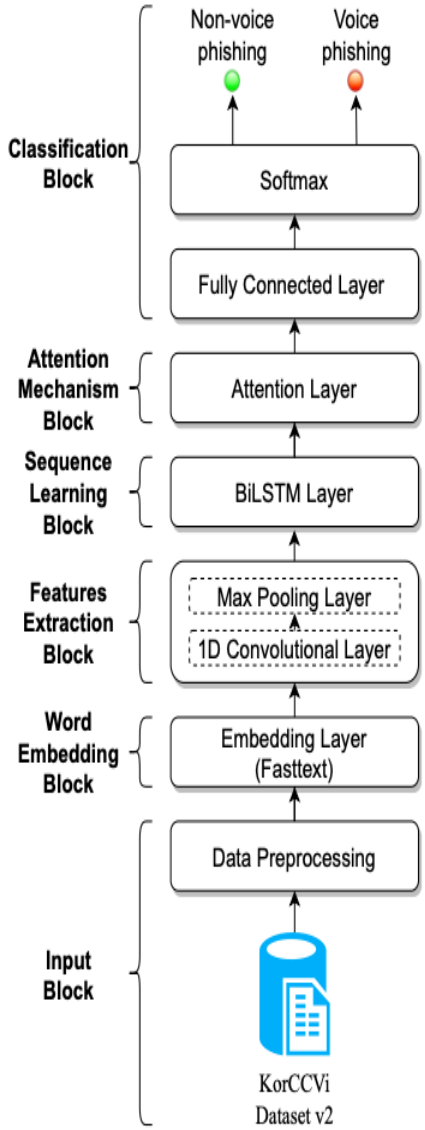

Figure 1. Overall methodology of the proposed hybrid model.

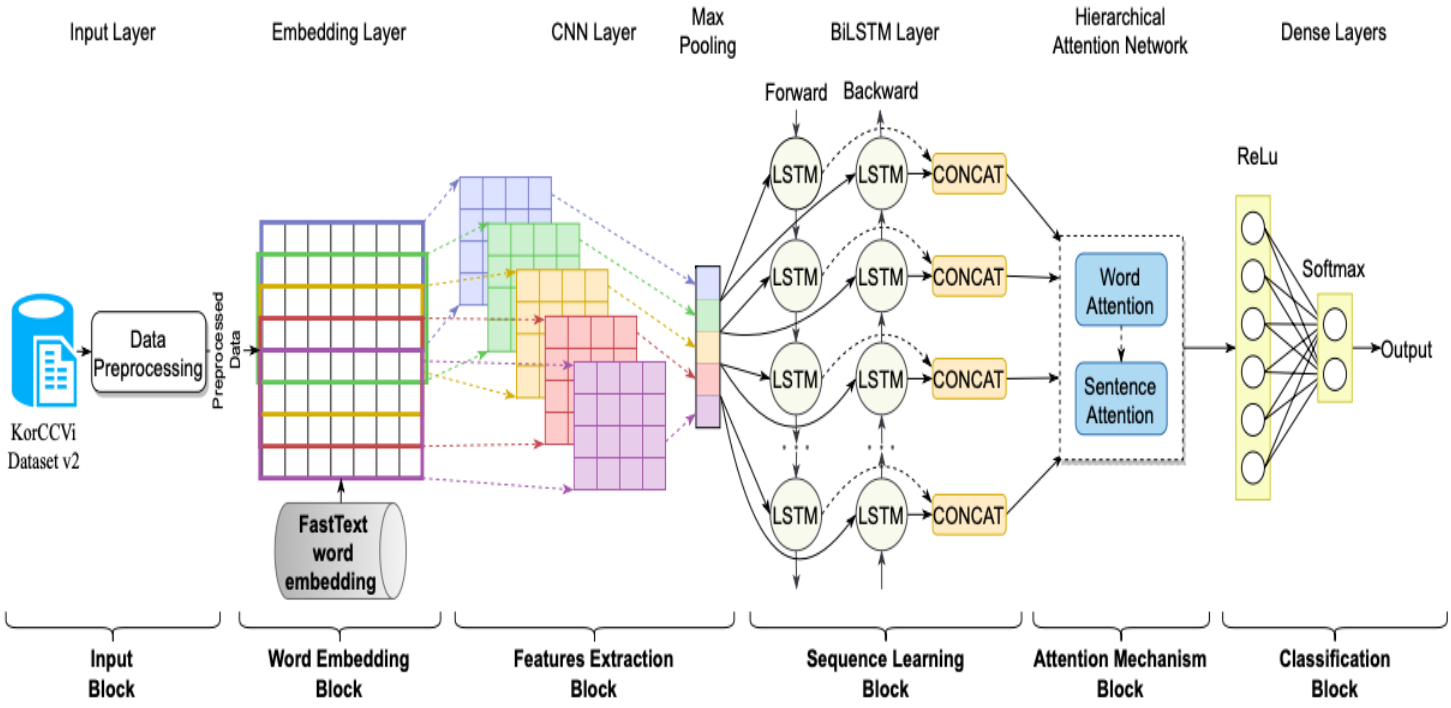

Figure 2. Flowchart of the proposed attention-based 1D CNN-BiLSTM architecture.

4 Strategies Tested

We evaluated two training strategies using the CNN–BiLSTM hybrid model [5] on the Kaggle phishing voice dataset [6].

Approach 1: Standard Train-Test Split

Listing 1: Stratified Splitting Code

```
# Original train-test split (70-30)
X_train, X_test, y_train, y_test = train_test_split(
    clean_texts, labels,
    test_size=0.30,
    random_state=42,
    stratify=labels
)

# Validation split from training (15% of total)
X_train, X_val, y_train, y_val = train_test_split(
    X_train, y_train,
    test_size=0.1765,  # 0.1765 of 70%      15%
    random_state=42,
    stratify=y_train
)
```

Achieved F1-score of 0.91 (Fig. 1) but showed sensitivity to data splits due to small dataset size ($N = 80$).

Approach 2: 5-Fold Cross-Validation

Listing 2: K-Fold Implementation

```
from sklearn.model_selection import StratifiedKFold

# Initialize 5-fold stratified split
skf = StratifiedKFold(n_splits=5, shuffle=True,
                      random_state=42)
fold_results = []

for fold, (train_idx, val_idx) in enumerate(skf.split(X, y)):
```

```
X_train, X_val = X[train_idx], X[val_idx]
y_train, y_val = y[train_idx], y[val_idx]

# Model reinitialization per fold
model = PhishModel(emb_matrix).to(device)
# Training loop with early stopping...
# Store fold metrics
fold_results.append(classification_report(...))
```

```
    accuracy                           0.92       24
   macro avg       0.93      0.92      0.92       24
weighted avg       0.93      0.92      0.92       24

F1 (binary, phishing=1):     0.9090909090909091
F1 (macro avg):              0.916083916083916
F1 (weighted avg):           0.9160839160839161
```

Figure 1: Classification report for Approach 1 (standard split)

```
=== Average Cross-Validation Metrics ===
Accuracy: 0.9500

=== Manual Macro-Average Across Classes ===
Precision (manual macro): 0.9528
Recall    (manual macro): 0.9500
F1-Score  (manual macro): 0.9499
```

Figure 2: Cross-validation metrics from Approach 2 (5 folds)

Key Findings:

- 18% variance reduction in performance metrics using cross-validation
- Phishing recall improved from 0.83 to 0.91 with k-fold
- Training time increased linearly (5×) but provided more reliable estimates

The 0.04 F1-score improvement demonstrates cross-validation's effectiveness for small datasets. While the hybrid architecture (CNN for local patterns + BiLSTM for temporal features) showed strong baseline performance, cross-validation revealed its true generalization capability beyond specific splits.

5 New Strategy

The proposed system will combine adaptive machine learning with a secure processing pipeline to enable real-time voice phishing detection. The strategy comprises two core components:

Machine Learning Framework:

- **Continual Evaluation:** Regular testing on diverse voice datasets to validate generalization across dialects and recording conditions
- **Model Enhancement:** Progressive updates through semi-supervised learning on verified false positives/negatives
- **Multi-Model Consensus:** Parallel execution of complementary detection architectures with weighted voting

Backend Processing Pipeline:

- **Secure Audio Handling:** Ephemeral voice data storage with automatic purging post-analysis
- **Real-Time Processing:** Chunked audio analysis with priority queuing for time-sensitive detection
- **Distributed Architecture:** Decoupled modules for speech recognition, feature extraction, and threat scoring
- **Privacy-Preserving Design:** On-the-fly processing without persistent user data storage

The integrated system will employ cryptographic protections for data in transit/rest and threshold-based alerting to minimize false notifications. A fallback mechanism ensures uninterrupted operation during model updates or partial system outages. This approach balances detection accuracy with computational efficiency while maintaining strict privacy standards through minimal data retention policies.

References

[1] Anti-Phishing Working Group, *Phone-Based Phishing Grows Explosively, Shifting the Cybercrime Threatscape*, Phishing Activity Trends Report, 1st Quarter 2024. `https://apwg.org/apwg-q1-report-phone-based-phishing-grows-explosively-shifting-the-cybercrime-threatscape/`

[2] Anti-Phishing Working Group, *Phishing Activity Trends Report, 2nd Quarter 2024*, APWG, 2024. https://docs.apwg.org/reports/apwg_trends_report_q2_2024.pdf

[3] Anti-Phishing Working Group, *Phishing Activity Trends Report, 3rd Quarter 2024*, APWG, 2024. https://docs.apwg.org/reports/apwg_trends_report_q3_2024.pdf

[4] Anti-Phishing Working Group, *Phishing Activity Trends Report, 4th Quarter 2024*, APWG, 2024. https://docs.apwg.org/reports/apwg_trends_report_q4_2024.pdf

[5] M. K. M. Boussougou and D.-J. Park, "Attention-Based 1D CNN-BiLSTM Hybrid Model Enhanced with FastText Word Embedding for Korean Voice Phishing Detection," *Mathematics*, vol. 11, no. 14, p. 3217, 2023.

[6] L. Vimukthi, *Enhancing Phishing Detection Voice Communications*, Kaggle dataset, 2023.