*mathematics*

MDPI

*Article*

# Attention-Based 1D CNN-BiLSTM Hybrid Model Enhanced with FastText Word Embedding for Korean Voice Phishing Detection †

**Milandu Keith Moussavou Boussougou [1] and Dong-Joo Park [2,*]**

[1] Department of Computer Science and Engineering, Soongsil University, Seoul 06978, Republic of Korea; mbmk92@soongsil.ac.kr
[2] School of Computer Science and Engineering, Soongsil University, Seoul 06978, Republic of Korea
* Correspondence: djpark@ssu.ac.kr
† Korea Computer Science Conference 2022 (KCC 2022), which was honored with the Best Paper Award.

**Abstract:** In the increasingly complex domain of Korean voice phishing attacks, advanced detection techniques are paramount. Traditional methods have achieved some degree of success. However, they often fail to detect sophisticated voice phishing attacks, highlighting an urgent need for enhanced approaches to improve detection performance. Addressing this, we have designed and implemented a novel artificial neural network (ANN) architecture that successfully combines data-centric and model-centric AI methodologies for detecting Korean voice phishing attacks. This paper presents our unique hybrid architecture, consisting of a 1-dimensional Convolutional Neural Network (1D CNN), a Bidirectional Long Short-Term Memory (BiLSTM), and Hierarchical Attention Networks (HANs). Our evaluations using the real-world KorCCVi v2 dataset demonstrate that the proposed architecture effectively leverages the strengths of CNN and BiLSTM to extract and learn contextually rich features from word embedding vectors. Additionally, implementing word and sentence attention mechanisms from HANs enhances the model's focus on crucial features, considerably improving detection performance. Achieving an accuracy score of 99.32% and an F1 score of 99.31%, our model surpasses all baseline models we trained, outperforms several existing solutions, and maintains comparable performance to others. The findings of this study underscore the potential of hybrid neural network architectures in improving voice phishing detection in the Korean language and pave the way for future research. This could involve refining and expanding upon this model to tackle increasingly sophisticated voice phishing strategies effectively or utilizing larger datasets.

**Keywords:** voice phishing; phishing; artificial intelligence; natural language processing; deep learning; attention mechanism; text classification; data-centric AI; model-centric AI

**MSC:** 68T50; 68T07; 68T05



**Citation:** Moussavou Boussougou, M.K.; Park, D.-J. Attention-Based 1D CNN-BiLSTM Hybrid Model Enhanced with FastText Word Embedding for Korean Voice Phishing Detection. *Mathematics* **2023**, *11*, 3217. https://doi.org/10.3390/math11143217

Academic Editors: Heui Seok Lim, Sanghyuk Lee, Yeongwook Yang and Imatitikua Aiyanyo

Received: 13 May 2023
Revised: 1 July 2023
Accepted: 13 July 2023
Published: 21 July 2023



## 1. Introduction

Voice Phishing, also called vishing or phone call scam, is a prominent cybercrime categorized as a phishing attack affecting thousands of people worldwide daily. During such attacks, the perpetrator impersonates a trustworthy entity and aims to gather personal and sensitive data (e.g., account credentials, bank account information, credit/debit card numbers, etc.) from their victims through persuasive phone calls using different social engineering tactics. In successful cases, victims can even be led to perform electronic money transfers to the fraudster's bank account.

The rapid development of new technologies and digitalization has enabled these attackers to adapt their stratagems and develop more sophisticated voice phishing attacks. These sophisticated attacks can now leverage artificial intelligence (AI) techniques, such as deepfake (voice clone) [1], to mimic a trusted individual's voice, making detection difficult

for targets and almost impossible to trace the attackers, often in a different country [2,3]. Moreover, in Korea, where this current study is focused, an example of sophisticated voice phishing tactics we observed is caller ID spoofing. Perpetrators utilize caller ID spoofing techniques during their calls, causing the victims' smartphone screens to display the name and phone number of someone they may trust, such as a family member, friend, or acquaintance. This tactic gives more credibility to the attacker and significantly raises their chances of successfully defrauding their victims. To reduce the number of people damaged by this cybercrime, entities impersonated by these attackers, such as banks and police departments, are stepping up their efforts. They are launching extensive campaigns, providing preventive education, and issuing media alerts to increase public awareness of this issue.

The Korean National Policy Agency reported that over the past five years, there have been 156,249 instances of voice phishing in the nation. These fraudulent activities have led to financial losses surpassing 3 trillion won (USD 2.2 billion) [4]. This impact has prompted academics and professionals to conduct more research to mitigate it, approaching it as a type of phone spam detection problem. Numerous studies have proposed voice phishing detection approaches that utilize machine learning (ML) and deep learning (DL) algorithms [5–9]. Some mobile applications leveraging ML and DL have also been publicly deployed to address this issue.

Indeed, ML [10,11] and DL [12,13] have revolutionized many areas of study by providing innovative solutions to complex problems, and their importance cannot be overstated. ML algorithms learn from and make data-based decisions, enabling pattern recognition, prediction, and decision-making, which are vital for numerous applications. On the other hand, DL, a subset of ML, leverages artificial neural networks (ANN) to extract intricate structures from high-dimensional data, making it especially suitable for image and speech recognition tasks. As in our study, ML and DL form the backbone of classification models [14–17], vital in tasks such as spam detection, sentiment analysis, and voice phishing detection. Classification models apply these learned patterns to new data, classifying them into specific categories. The advent of ML and DL and their application through classification models have been essential in making significant progress in cybersecurity, including detecting voice phishing attacks by enabling the identification of malicious activities with increased precision and reliability.

The various existing detection techniques and preventive measures are significant steps toward mitigating voice phishing crime in Korea. However, they still need to be improved, as we observed significantly fewer studies on Korean voice phishing detection, despite voice phishing continuing to be a major concern in the country. This pernicious cybercrime continues to cause significant financial and social loss, highlighting the need for more efficient detection systems. One of the main challenges that hinder progress in addressing this issue is the limited availability and small sizes of real-world Korean voice phishing datasets. To this end, our study proposes a hybrid ANN to enhance the detection of Korean voice phishing attacks. We explore for the first time the performances of an attention-based 1D CNN-BiLSTM model for detecting voice phishing in Korea using the updated version of the Korean Call Content Vishing (KorCCVi) dataset [7]. This novel approach combines data-centric and model-centric AI methodologies, offering a complementary solution to the limited dataset problem. The significance of this approach arises from the ongoing debate on whether AI research should focus on a model-centric or data-centric approach. While a data-centric approach focuses on data collection and preparation, a model-centric approach emphasizes the model's architecture and optimization techniques [18]. However, recent work proposed a combination of both [19], suggesting a more robust and compact solution, especially in the face of limited data availability, such as in the case of Korea.

This paper extends our prior work, initially presented at the Korea Computer Congress (KCC) 2022 [20]. Noting the successful application of various DL techniques in phishing attack detection, the proposed attention-based 1D CNN-BiLSTM hybrid architecture, which leverages the strengths of the combined DL techniques, has the potential to improve

existing Korean voice phishing detection methods significantly. With this in mind, our study aims to address the challenge of feature extraction from small voice phishing datasets and build an efficient DL model that may outperform existing works. Our contribution is summarized as follows:

- For the first time, we investigate the performance of a novel hybrid ANN that combines a 1-dimensional Convolutional Neural Network (1D CNN), a Bidirectional Long Short-term Memory (BiLSTM) architecture, and a Hierarchical Attention Network (HAN) for detecting Korean voice phishing attacks.
- We demonstrate the effectiveness of a complementary approach combining data-centric and model-centric AI methodologies to address the challenge of limited dataset availability in the context of voice phishing detection in Korea.
- We train and evaluate the prediction performance of our proposed hybrid detection model and other baseline models using the updated version of the KorCCVi dataset.
- We compare our proposed method with several existing methods for detecting Korean voice phishing.

The remainder of this study is organized as follows: Section 2 presents related work, including prior studies on voice phishing detection and a summary of relevant works leveraging hybrid ANN architectures in natural language processing (NLP) tasks. Section 3 details the methodology of the proposed attention-based 1D CNN-BiLSTM model for Korean voice phishing detection. Section 4 discusses the implementation and experimental results, including model performance evaluation and analysis of the results. Finally, Section 6 presents the conclusion and future works.

All experimental codes and the dataset used in this study are available at: https://github.com/selfcontrol7/Korean_Voice_Phishing_Detection/tree/main/Attention (accessed on 10 May 2023).

## 2. Related Work

Examining the literature, it is clear that various methodologies have been employed in detecting different types of phishing attacks, among which AI stands out as the highly utilized approach [21–23]. Alongside AI, there are other strategies such as rule-based and reputation-based systems, content analysis, anomaly detection methods, blacklisting or whitelisting, user awareness, and anti-phishing toolbars and browser extensions that cybersecurity researchers have adopted for phishing detection [24,25]. However, we noticed that most of these studies predominantly focus on websites, Uniform Resource Locator (URL), and email phishing detection, leaving voice phishing detection relatively underexplored, as reported in this study on detailed analysis of mobile phishing [23].

Nonetheless, regarding voice phishing and phone spam detection, a few AI-enabled approaches using NLP, ML algorithms, DL algorithms, and hybrid networks have been widely explored as suitable detection solutions. These methodologies are now considered the standard due to their adaptive and sophisticated nature. Grouped by the strategy used, Table 1 summarizes the existing few studies on voice phishing detection using the abovementioned methodologies. Our work aims to address this research gap by providing a comprehensive study on Korean voice phishing detection and a novel robust detection solution.

**Table 1.** Summary of the existing studies on voice phishing detection.

| Ref. | Year | Phishing Types | Strategies | Methods | Datasets | Advantages | Limitations |
|---|---|---|---|---|---|---|---|
| [26] | 2014 | Voice phishing, Spoofing | Caller ID verification | Trace back incoming call to its originating gateway | N/A | Real-time Caller ID Spoofing Detection, Minimal Call Setup Time Impact, Device Compatibility | End-User Response Dependence, Stakeholder Participation Dependence, Lack of Real-World Testing |
| [5] | 2020 | Voice phishing | Blacklisting and whitelisting | ML | Global phone book | Caller Identification, Mobile Phishing Attack Prevention, Comprehensive Solution Against Mobile Phishing | Inefficiency Handling New Phone Numbers |
| [27] | 2022 | Voice phishing, Deepfake | User authentication and Deep voice detection | AutoEncoder | ASVspoof 2019 | Synthetic Voice Detection, Deepfake Voice Identification, Sender Identity Validation | Generalizability on Non-English Datasets, Computational Demand, Lack of Comparative Analysis with Existing Methods |
| [28] | 2021 | Voice phishing | Conversation semantic content analysis | K-Means | Human hand made scams conversations and CallHome dataset | Scam Signatures Introduction, Novel Concept for Scam Calls Detection | Unrepresentative Nature of Human-Generated Telephone Scam Conversation Dataset |
| [29] | 2019 | Voice phishing (telecommunication finance fraud) | Hybrid (Blacklisting, Rule, CNN) | Filtering, CNN | Financial transaction data | Rule Models and AI Algorithms Integration, Synergistic Combination of Techniques | Lack of Comparative Analysis with Existing Methods, Lack of Real-World Testing |
| [30] | 2018 | Voice phishing (telecommunication fraud) | Call content analysis | ML, NLP, rules | Fraudulent call description crawled on social medias | Superior Performance over Blacklisting Strategies, Development of Android Application | Small Dataset Size, Lack of Real-world Samples in Dataset, Inferior Performance of Local-based Speech Recognition vs. Cloud-based |
| [31] | 2018 | Voice phishing | Call content analysis | TF-IDF [1], Label propagation community (LPA) | Call texts | Fraud Calls Detection in Community Network, Non-reliance on Passive Interception on Smartphone Terminals | Potential for Missed Fraudulent Calls, Analysis Indicating Vulnerabilities within Isolated Communities |
| [8] | 2021 | Voice phishing | Call content analysis | Latent semantic analysis (LSA), K-means | FSS [2] | Comprehensive Comparison of Speech Recognition Techniques, Detailed Examination of Embedding Techniques | Small Dataset Size, Resulting Low Performance, Lack of Experiment Details, Issues with Replicability |

**Table 1.** *Cont.*

| Ref. | Year | Phishing Types | Strategies | Methods | Datasets | Advantages | Limitations |
|---|---|---|---|---|---|---|---|
| [7] | 2021 | Voice phishing | Call content analysis | NLP, Random Forest, XGBoost, LGBM, and CatBoost, Linear SVC, RNN, BiLSTM, GRU | KorCCVi v1 | Introduction of KorCCVi Dataset, Inclusion of Real-world Voice Phishing Data, Real-Time Voice Phishing Detection Efficiency | Small Dataset Size, Generalizability on Non-Korean Datasets, Lack of Real-World Testing |
| [9] | 2021 | Voice phishing | Call content analysis | NLP, CatBoost, Gradient XGBoost, LGBM, Linear SVC | KorCCVi v1 | Rapid Training Time, Swift Inference Time, Efficiency of ML Model | Small Dataset Size, Lack of Deep Learning Architectures Explored |
| [6] | 2021 | Voice phishing | Call content analysis | SVM, Logistic Regression, Decision Tree, Random Forest, XGB | FSS + NIKL [3] | Real-Time Detection Capability, Comparative Evaluation of Two Korean Morpheme Analyzers | Imbalanced Dataset, Lack of DL Architectures Explored, Lack of Comparative Analysis with Existing Methods |
| [32] | 2021 | Voice phishing | Call content and sentiment analysis | CNN, BiLSTM | - | Inclusion of Sentiment Analysis for Enhanced Detection, Reliable and Efficient Model Implementation | Insufficient Dataset Details, Requirement of Domain-Independent Sentiment Lexicon |
| [33] | 2021 | Voice phishing | Call content analysis | Naive bayes, CNN | Mix of conversational transcripts and human made fraud calls transcripts | Use of Oversampling Method for Dataset Skewness, High Performance of Intent Analysis Models | Highly Imbalanced Dataset, Lack of Real-world Samples in Dataset, Need for Additional Algorithm Testing |
| [34] | 2022 | Voice phishing | Call content analysis | KoBERT | KorCCVi v1 | High Accuracy Achieved with KoBERT-based Model, Extensive Comparison with ML and DL Algorithms | Small Dataset Size, Imbalanced Dataset, Lack of Hyperparameter Optimization |
| [35] | 2023 | Voice phishing | Call content analysis | KoBERT | FSS + AI Hub | Impressive Accuracy of KoBERT-based Model, Provision of Educational Content for Potential Victims, Risk Evaluation API Service | Lack of Dataset Details, Overemphasis on Model Accuracy Metric, Overfitting Issue Beyond 10 Epochs |
| [36] | 2023 | Voice phishing | Call content analysis | Federated Learning | KorCCVi v2 | User Data Privacy Preservation, Communication Efficiency, Client Grouping Based on Characteristics, Personalized Data Requirement Recommendations | Overemphasis on Model Accuracy Metric, Lack of Comparative Analysis with Existing Methods |

[1] Term Frequency-Inverse Document Frequency (TF-IDF); [2] The Financial Supervisory Service of Korea (FSS); [3] The National Institute of the Korean Language (NIKL).

### *2.1. Voice Phishing Detection*

Various approaches have been explored to detect voice phishing in Korea. Song et al. [26] conducted a pioneer detection study that proposed a novel approach to combat voice phishing by identifying deceptive manipulation of caller IDs in real time. The proposed solution, iVisher, utilizes Session Initiation Protocol-based (SIP) Voice-over-Internet Protocol (VoIP) to detect and authenticate caller IDs, mitigating the risk of spoofing. This system enhances the security of telephone communications by tracking the incoming call back to its originating gateway and verifying the consistency of the displayed name with the actual caller ID. However, the authors noted that the effectiveness of iVisher mainly relies on user responsiveness and the cooperation of organizations.

Similarly, Kang et al. [27] recently introduced an innovative solution known as Deep-Detection, using an autoencoder for two-fold authentication to counter voice phishing. Their system detects synthetic or deepfake (voice clone) and verifies the sender's identity. Importantly, it does so while ensuring privacy, as voice data preprocessing occurs locally on the user's device, avoiding the need for the data to be stored on detection servers. However, despite its impressive performance on the ASVspoof 2019 dataset, concerns about the model's generalizability, computational demands, and a lack of comparison with other methods limit the study's comprehensiveness. Alongside this work on privacy-centric voice phishing detection, Yoon et al. [36] proposed an innovative federated learning-based approach that ensured user data privacy while improving detection accuracy. While the study showcased promising results with an emphasis on privacy preservation and communication efficiency, it mainly focused on the accuracy of the detection algorithm, leaving room for a comprehensive evaluation using other performance metrics such as recall, precision, and F1-score. Despite these limitations, the authors made a noteworthy contribution to privacy-protecting phishing detection, setting the stage for future research.

The blacklisting and whitelisting-based detection strategy for Korean voice phishing detection was investigated by Tran et al. [5] in their study. The authors proposed an intelligent system, iCaMs, to identify callers and prevent voice phishing damage. The system is based on a global phone book with a black and whitelist of phone numbers and implemented in a client–server architecture, where the client is a mobile application, and the web server applies ML to validate numbers. However, the common drawback of blacklisting or whitelisting-based detection approaches is that they are inefficient when encountering new phone numbers. To cope with this drawback, Jeong and Lim [29] addressed telecommunication finance fraud accidents, also called voice phishing, by designing an intelligence-based detection model. They combined blacklisting-based, scenario-based rule models and CNN as AI algorithms, and the paper suggests a hybrid model that aims to improve the accuracy of detecting abnormal financial transactions.

Although many proposed solutions rely on features such as voice, phone numbers, telecommunications data, and financial transactions, voice phishing detection has also been explored using call content. This method typically treats the problem as a binary text classification task. For instance, Zhao et al. [30] proposed an Android application to detect voice phishing in China by dynamically analyzing and classifying the call content. The novel proposed system extracts the features, leverages NLP techniques, and classifies them using ML algorithms and detection rules. However, despite its outperforming results compared to the blacklisting or whitelisting strategies, the study relied on low real-world fraudulent phone call samples. That same year, Peng and Lin [31] presented similar research using label propagation community detection algorithm (LPA) to analyze fraud phone calls. Likewise, Kim et al. [8] carried out voice recognition on real-world voice phishing data and subsequently compared the accuracy of Korean voice phishing text classification across several different method combinations. However, the limited size of their dataset resulted in a maximum accuracy of only 61% and an F1 score of 74%, which are noticeably low figures.

In our prior studies, we proposed a real-time Korean voice phishing detection approach using NLP, ML, and DL models [7]. The approach aimed to efficiently detect voice

phishing by performing a speech recognition task on incoming calls. We then analyzed the call content using a model trained on the KorCCVi dataset, built with real-world voice phishing data. In a further study, we benchmarked the classification performance of several ML algorithms on the KorCCVi dataset [9]. Concurrently, Lee and Park [6] studied a real-time Korean voice phishing detection system using basic ML models, motivated by rapidly training ML models for real-time Korean voice phishing detection. Kim et al. [32] mainly utilized DL techniques and sentiment analysis to determine whether a call is voice phishing. They presented a model that combines Deep Neural Networks (DNN), CNN, and BiLSTM to classify fraudulent calls. By incorporating sentiment analysis from both text and voice perspectives, the proposed model aims to enhance the reliability of the detection results by considering the emotional state of voice phishing perpetrators.

In India, Kale et al. [33] used Naive Bayes and CNN algorithms in their studies to analyze the conversation transcripts' intent and classify the fraudulent call. Derakhshan et al. [28] presented a novel approach called the Anti-Social Engineering Tool (ASsET) to detect social engineering attacks in telephone scams. By analyzing the semantic content of conversations, the ASsET approach identifies patterns in speech acts and utilizes word embedding techniques to detect scam signatures. The proposed approach demonstrates high accuracy in detecting scam calls. In addition, the availability of a telephone-based scam dataset generated through a human subject study further contributes to the research on social engineering attacks. However, this telephone-based scam conversation dataset may not reflect reality as it does not contain real-world scam conversations.

Moreover, the most recent studies analyzing call content have also leveraged the power of pre-trained language models (LMs). For the first time, Moussavou and Park [34] explored the use of the Korean LM, KoBERT, for detecting Korean voice phishing. The proposed approach demonstrated superior results, which can be attributed to the fine-tuning of the Korean LM using the KorCVVi dataset. Concurrently, Yang et al. [35] trained their models using KoBERT and achieved impressive accuracy. Besides the detection, their solution provides an educational content section to improve the users' ability to differentiate between normal and phishing calls.

### *2.2. Convolutional Neural Networks and Hybrid Approaches*

Despite the excellent performance of basic ML and DL approaches in detecting Korean voice phishing, several studies presented in the Section 2.1 agreed that, although the DL approach has a non-negligible model training time as a drawback, it has more potential to detect voice phishing accurately. Several researchers have already proven the potential of DL algorithms in phone spam detection and text classification tasks. However, using different existing DL algorithms for NLP tasks has revealed limitations, such as dealing with long sequences and a large number of parameters. As a result, using hybrid ANNs or, more recently, foundation models (e.g., Transformers model, GPT model) [37] is an emerging trend to overcome the drawbacks of DL algorithms and improve the performance in NLP tasks such as text.

Among the DL algorithms, CNN-based methods have found immense success in various complex applications. This success is mainly due to their ability to directly extract learned features from raw data, thereby maximizing classification accuracy. This high level of performance has made CNNs an attractive option when constructing hybrid architectures to tackle complicated engineering applications. In this context, 1D CNNs, a variant of conventional 2D CNNs, have swiftly gained researchers' attention since their introduction in 2015 by Kiranyaz et al. [38] for an electrocardiogram (ECG) classification and monitoring system. These compact and adaptive 1D CNNs offer advantages and superiority over their 2D counterparts, especially when processing sequential and 1D signal data.

Furthermore, numerous studies have attested to the benefits of 1D CNNs in a broad range of domains. These include NLP tasks such as sentiment analysis, text classification, machine translation, and speech recognition, as well as in the healthcare and various engineering sectors such as civil, environmental, mechanical, and electrical [39,40]. The success

and advantages of CNNs, particularly 1D CNNs for sequential data and 1D signals, have significantly contributed to their integration into hybrid architectures.

To overcome the limitations of ANNs, such as CNNs and Recurrent Neural Networks (RNNs), researchers have proposed various hybrid models, which have proven efficient for detecting phishing attacks and NLP tasks, especially those applying attention mechanisms. Fang et al. [41] proposed a new phishing email detection model named THEMIS, based on a recurrent convolutional neural networks (RCNN) model with multilevel vectors and attention mechanisms. Although the proposed model achieved an accuracy score of 99.84%, it is mainly designed for phishing email detection and no other phishing types. Huang et al. [42] proposed PhishingNet, a deep learning-based approach for timely detecting phishing Uniform Resource Locators (URLs) to improve phishing URL detection. PhishingNet used CNN and attention-based hierarchical RNN modules, achieving an accuracy score of 97.90%.

Zhou et al. proposed [43] Hybrid Attention Networks (HANs) for Chinese short text classification, applying RNN and CNN to extract semantic features at the word and character level and using attention mechanisms with context for each level. On the contrary, Hao et al. [44] proposed a novel framework called Mutual-Attention Convolutional Neural Networks to avoid the feature information loss observed in Ref. [43]. Furthermore, Deng et al. presented a new model called attention-based BiLSTM fused CNN with a gating mechanism (ABLG-CNN) to address the challenge of Chinese long text classification [45].

In the sentiment analysis task, Jang et al. [46] proposed a novel hybrid model combining Word2vec, CNN, BiLSTM, and attention mechanism, leveraging LSTM and CNN's distinct advantages to classify sentiment on the Internet Movie Database (IMDB) dataset. More recently, to address the problem of classifying long text for sentiment analysis, Kamyab et al. [47] proposed a novel hybrid model called ACR-SA, an attention-based model using two-channel CNN and Bi-RNN. The authors jointly used BiLSTM and Bidirectional Gated Recurrent Unit (BiGRU) to achieve better results in large and small datasets.

This study builds upon the success and effectiveness of CNNs and 1D CNNs as demonstrated in various applications and the trend of incorporating them into hybrid architectures for solving complex tasks. We aim to employ a hybrid ANN architecture combined with the attention mechanism to build an efficient Korean voice phishing detection model.

## 3. Methodology

This study proposes a novel Korean Voice Phishing detection model, which is improved using a hybrid ANN architecture composed of a 1D CNN, BiLSTM, and hierarchical attention mechanism layers. The proposed architecture consists of six blocks: input block, word embedding block, features extraction block, sequence learning block, attention layer block, and classification block, as illustrated in Figure 1.

In the proposed approach, the word embedding technique FastText [48–50] trains the word feature vector representation from the preprocessed input data while the 1D CNN layer extracts local features. Then, the features obtained by the 1D CNN layer are fed to the BiLSTM layer, which extracts the long dependencies features. Following the sequence learning block, the attention layer receives the BiLSTM output and assigns particular attention to each word. This attention mechanism is crucial to perform the final classification, as it weighs each word according to its relevance and contribution to the voice phishing context. The detailed classification process employed by our proposed model is demonstrated in Figure 2. Furthermore, for a more granular understanding of our methodology, each of these blocks is discussed separately and in detail in the following subsections, offering a comprehensive view of the working mechanism of our architecture.

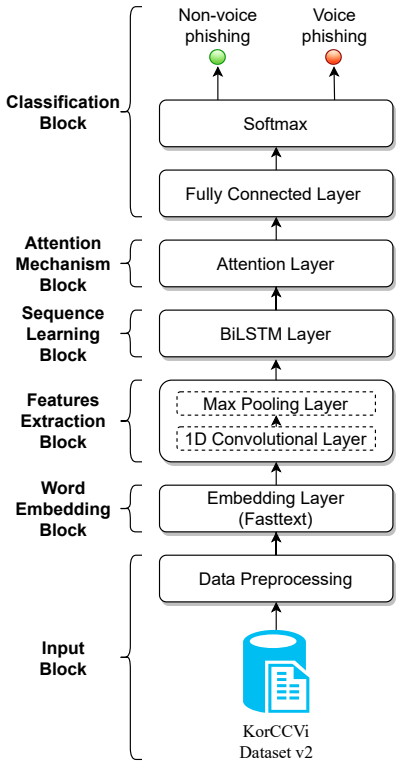


**Figure 1.** Overall methodology of the proposed hybrid model.

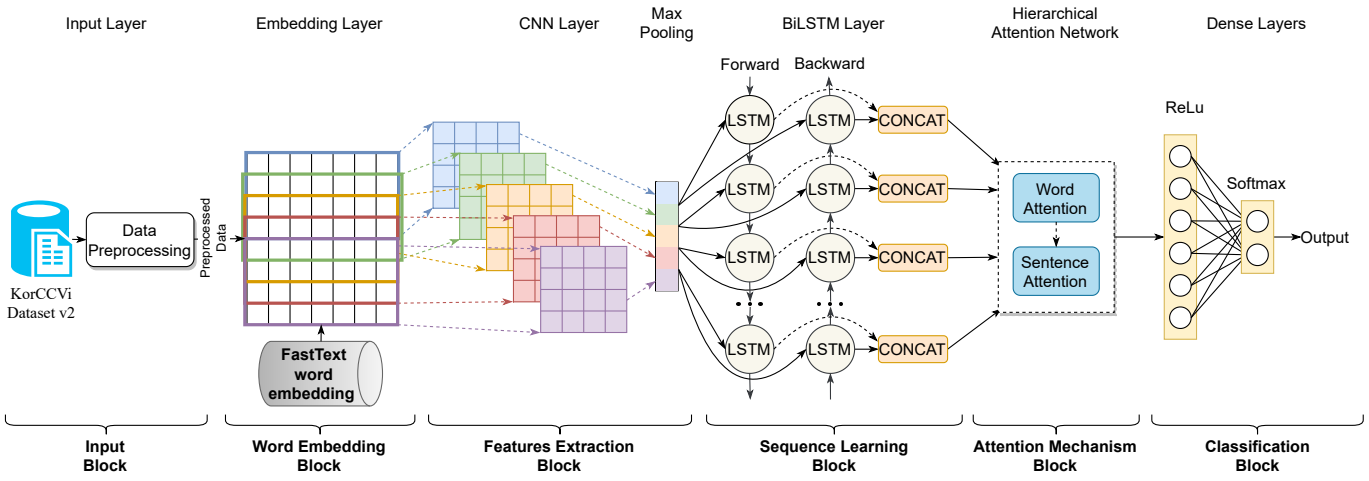


**Figure 2.** Flowchart of the proposed attention-based 1D CNN-BiLSTM architecture.

### *3.1. Input Block*

The dataset used in this study contains raw data, and a preprocessing step must be performed first on these raw data before any further steps. Indeed, to fully unleash the potential of any DL algorithm and build an accurate model, it is crucial to spend the necessary amount of time preparing the dataset through preprocessing, as it significantly impacts the model's performance [51]. This study follows the best practices in text preprocessing recommended by NLP experts [37,52] to reduce the models' training time and address the dimensionality problem of the feature matrix.

The input block layer contains the list of tokens generated from the preprocessing step of our voice phishing dataset. The preprocessing tasks performed on the collected raw data from the KorCCVi dataset for the voice phishing and non-voice phishing classes are detailed as follows:

- Transcription and Audio Processing: We used Google's Cloud Speech-to-Text API [53] to transcribe our voice data into textual format. To generate the most accurate transcriptions possible, we converted the audio channel and format and manually adjusted the audio files for optimal audibility and length. If necessary, we segmented audio files containing multiple voice phishing attacks and manually compared optimized transcript versions to select the most accurate version. All detailed processes can be found in our prior work [54].
- Data Cleaning: This step involved the removal of any irrelevant or redundant information from the raw data. It includes eliminating personal information such as phone numbers, punctuation marks, special characters, and digits. This cleaning task ensures that we remove irrelevant or redundant information that does not benefit in understanding voice phishing characteristics.
- Tokenization: While performing the cleaning tasks, we tokenized the dataset using the morphological analyzer MeCab-ko [55] due to its high-speed morphological analysis. This process breaks down the cleaned text into individual tokens or words, which serve as our model's basic input units. Several tokenization strategies are available for the Korean language, and their impact on a DL-based voice phishing detection model presents a potential research topic of interest to scholars. This comparative study could uncover this application's most effective tokenization method.
- Removal of Stop Words: This step involved the removal of Korean stop words that carry little semantic weight in the context of voice phishing.

After completing these preprocessing steps, all the tokens created from the dataset were encoded and stored in an NPZ file using NumPy. This file was subsequently used to build the word embedding. These meticulous preprocessing steps ensure that our model is fed with high-quality, relevant input, facilitating effective learning and superior performance in voice phishing detection.

### *3.2. Word Embedding Block*

The word embedding block contains word feature vectors representing the inputted list of tokens or words. The embedding efficiently represents text in a multidimensional space, capturing the contextual relationships between words and enabling in-depth feature extraction.

The purpose of the word embedding layer is to map the segmented tokens into word feature vectors using a pre-trained embedding dictionary or matrix. We consider the input sequence of tokens represented as $D = \{t_1, t_2, t_3, \ldots, t_M\}$, where $t_1$ represents a token in $D$, and $M$ the number of tokens in $D$. For each token in $D$, we look up the pre-trained embedding matrix $W$ as a table and dictionary and map each token in the sequence to its corresponding word vector. With $W \in \mathbb{R}^{d \times V}$, where $d$ represents the word embedding dimension and $V$ the number of tokens in the token list (i.e., vocabulary size).

We employed Korean pre-trained word vectors as an embedding matrix provided by FastText to build the embedding word feature vectors. FastText provides a spatial representation of 157 different languages trained on massive datasets. Since it implements the sub-word tokenization method, it improves the quality of representation for rare words and out-of-vocabulary words, a feature often absent in other word embedding methods such as Word2Vec or GloVe [48–50]. It also handles the issue of word disambiguation more effectively by considering the context of words. These attributes of FastText allow us to generate word embeddings that effectively translate our text data into a numerical format that our model can understand and learn from. This method is particularly beneficial for languages with heavy morphemic reliance, such as Korean.

Compared to traditional methods such as one-hot encoding, using word embedding as input for our neural networks has the advantage of mitigating the high-dimensionality problem. These embeddings capture the words' semantic meaning and context, forming the bedrock of our model's understanding of voice phishing. The word embedding block is

pivotal in creating a dense, feature-rich representation of the input data, thus enabling our model to perform effective voice phishing detection.

### *3.3. Features Extraction Block*

In the features extraction block, the embedding word feature vectors are fed to the 1D CNN as input to perform a textual feature extraction with the CNN. The conventional architecture of a CNN comprises the input layer, the convolutional layer, the pooling layer, the fully connected layer, and finally, the output layer. However, as our architecture is hybrid, the CNN's layers, as mentioned earlier, are not aligned back-to-back.

In the convolutional layer of our 1D CNN, the convolution kernel (filter) only slides in vertical order toward the embedding matrix's length to transform the inputted word feature vectors to generate the feature maps. Then, the nonlinear activation function performs another transformation on these feature maps to generate the output of the 1D convolutional layer. Explained in simpler terms, during the 1D convolution operations, scalar multiplications and additions are performed on the output of the embedding layer to extract the contextual features.

The role of the pooling operation is to sample the output of the 1D convolutional layer and reduce the number of needed features and the convolution vector's size to prevent overfitting. This operation reduces the number of parameters of our neural network, thereby decreasing the computational complexity. Different pooling methods exist, such as maximum pooling (or max pooling), global pooling, and average pooling. This study uses the max pooling method subtracting the largest scalar value from the result vector obtained during each convolution operation.

In this study, the 1D convolution operations are performed using only one kernel of size 3, equal to the length of the 1D convolution window. The choice of a 1D CNN layer and a unique kernel size in this work is supported by numerous reasons, as detailed by Kiranyaz et al. [39]. A 1D CNN layer is particularly suited to handling sequence data, such as our text data, as it can efficiently manage inputs of varying lengths and automatically learn spatial hierarchies of features. Given the nature of the data we handle, this ability is crucial in our context. Although using kernels of various sizes may lead to a better model performance by capturing adjacent words, using a single kernel size helps maintain a balance between the model's computational performance and its practical feasibility. Notably, the computational complexity of a 1D CNN is significantly lower than that of a 2D CNN, making it a more efficient choice for small datasets. The reduced complexity and unique kernel size collectively help reduce the training time and the number of parameters to be trained, thereby decreasing the overall model complexity. This aspect is particularly relevant in our case, considering the small dataset we work with. Additionally, compact 1D CNNs are well-suited for real-time and low-cost applications due to their low computational requirements, especially on mobile or hand-held devices with limited computational power. Therefore, using a 2D CNN in our hybrid architecture may not be suitable for real-time voice phishing detection applications targeting mobile phones, which are low-power/low-memory devices [39].

### *3.4. Sequence Learning Block*

The BiLSTM [56] layer consists of two parallel LSTM layers that function in opposite directions and are used in the sequence learning block. The first layer receives as input the feature vector from the Max pooling operation in the 1D CNN layer and propagates it in the forward direction. The second layer propagates backward to capture the long-term dependencies in the left and right contexts. This bidirectional processing allows the model to capture information from both past and future contexts, enhancing its ability to learn complex patterns in the input data.

Mathematically, the forward LSTM layer computes the hidden state ($h_t^f$) at time step $t$ using the following equations:

Input gate:

$$i_t^f = \sigma(W_i^f x_t + U_i^f h_{t-1}^f + b_i^f) \quad (1)$$

Forget gate:

$$f_t^f = \sigma(W_f^f x_t + U_f^f h_{t-1}^f + b_f^f) \quad (2)$$

Output gate:

$$o_t^f = \sigma(W_o^f x_t + U_o^f h_{t-1}^f + b_o^f) \quad (3)$$

Cell state update:

$$c_t^f = f_t^f \odot c_{t-1}^f + i_t^f \odot \tanh(W_c^f x_t + U_c^f h_{t-1}^f + b_c^f) \quad (4)$$

Hidden state update:

$$h_t^f = o_t^f \odot \tanh(c_t^f) \quad (5)$$

Similarly, the backward LSTM layer computes the hidden state ($h_t^b$) at time step $t$ using analogous equations:

Input gate:

$$i_t^b = \sigma(W_i^b x_t + U_i^b h_{t+1}^b + b_i^b) \quad (6)$$

Forget gate:

$$f_t^b = \sigma(W_f^b x_t + U_f^b h_{t+1}^b + b_f^b) \quad (7)$$

Output gate:

$$o_t^b = \sigma(W_o^b x_t + U_o^b h_{t+1}^b + b_o^b) \quad (8)$$

Cell state update:

$$c_t^b = f_t^b \odot c_{t+1}^b + i_t^b \odot \tanh(W_c^b x_t + U_c^b h_{t+1}^b + b_c^b) \quad (9)$$

Hidden state update:

$$h_t^b = o_t^b \odot \tanh(c_t^b) \quad (10)$$

In these equations, $x_t$ represents the input at time step $t$, $\sigma$ is the sigmoid activation function, tanh is the hyperbolic tangent activation function, and $\odot$ denotes element-wise multiplication. $W$ and $U$ denote weight matrices, while $b$ represents bias terms.

Finally, the forward and backward hidden states ($h_t^f$ and $h_t^b$) are concatenated to form the final output at each time step: $h_t = [h_t^f; h_t^b]$. Thus, a BiLSTM neural network offers superior performance for text classification tasks and is more powerful than a unary LSTM layer. Although a BiLSTM neural network can capture a text sequence's preceding and subsequent contextual information, it still struggles to capture all the dependencies of every previous word. In the structure of the proposed hybrid network, the attention mechanism block is there to address this issue.

### *3.5. Attention Mechanism Block*

The attention layer is directly plugged into the output layer of BiLSTM, and the final combined hidden state from the BiLSTM is fed to the attention mechanism. The attention mechanism focuses on each token or word of the input sequence and analyzes each word's semantic correlation. In other words, it finds how each word relates to all the other words in the sequence. Then, the attention mechanism allocates different attention weights (or attention scores) to the semantic coding of the word vector, which will help to select the word vector that highly correlates with the classification's final stage.

There are multiple variants of the attention mechanism that differ in the way the attention scores are calculated. The hybrid DL architecture proposed in this paper uses the Hierarchical Attention Networks (HAN) variant proposed by Yang et al. [57] for document classification. The HAN method considers the hierarchical structure of documents

(document—sentences—words) and consists of two levels of attention: word-level attention and sentence-level attention.

Each word in a sentence is assigned an importance weight at the word-level attention. The word-level attention is computed using a bidirectional GRU (BiGRU) layer to encode the input sentence. In Equation (11), let $h_{it}$ denote the hidden state of the BiGRU at time step $i$ for a given sentence. The word-level context vector (the word's weight), $u_{it}$, is calculated by applying a learned weight matrix $W_w$ and bias term $b_w$ to the hidden state:

$$u_{it} = \tanh(W_w h_{it} + b_w) \tag{11}$$

Next, the importance weight $\alpha_{it}$ for each word is computed using a softmax function, which assesses the word's importance as the degree to which $u_{it}$ is similar to a word-level context vector $u_w$. The total number of time steps in the input sequence is denoted by $T$. Equation (12) denotes $\alpha_{it}$:

$$\alpha_{it} = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)} \tag{12}$$

The sentence vector $s_i$ is then obtained by taking a weighted sum of the hidden states, as denoted in Equation (13):

$$s_i = \sum_t \alpha_{it} h_{it} \tag{13}$$

A similar process is applied to a sequence of sentence vectors at the sentence-level attention. The sentence-level attention is also computed using a BiGRU layer to encode the input document. In Equation (14), let $h_i$ denote the hidden state of the BiGRU at time step $i$ for a given document. The sentence-level context vector, $u_i$, is calculated by applying a learned weight matrix $W_s$ and bias term $b_s$ to the hidden state:

$$u_i = \tanh(W_s h_i + b_s) \tag{14}$$

The importance weight $\alpha_i$ for each sentence is computed using a softmax function denoted in Equation (15):

$$\alpha_i = \frac{\exp(u_i^T u_s)}{\sum_i \exp(u_i^T u_s)} \tag{15}$$

Finally, the document vector $v$ in Equation (16), which summarizes all the information of sentences in a document, is obtained by taking a weighted sum of the sentence-level hidden states:

$$v = \sum_i \alpha_i h_i \tag{16}$$

The HAN effectively captures the hierarchical structure of text data by applying attention mechanisms at both the word and sentence levels, resulting in more accurate and contextually rich representations of the input data. The final output of the attention mechanism block is a document vector of features, which is used to perform the final classification.

### *3.6. Classification Block*

The classification block performs the Korean voice phishing classification task. This block comprises a fully connected layer (or dense layer) with the activation function ReLu, which is connected to the final classification dense output layer with two neurons (i.e., one output neuron for each prediction class, voice phishing, and non-voice phishing). The document vector $v$ from the attention layer is fed as input to the first dense layer to perform a nonlinear transformation. Then, the last classification dense layer uses the categorical cross-entropy loss function and the Softmax activation function to convert the previous vector of numbers into a vector of probabilities and output the distribution

probability for each label for the inputted Korean text. The probability is calculated using the Softmax function as demonstrated in Equation (17):

$$P = softmax(W_c v + b_c) \tag{17}$$

## 4. Experiments and Experimental Results

This section presents the dataset used in our experiment, the environmental setup, the different baselines trained and compared, the evaluation metrics, and the experimental results analysis.

### *4.1. Dataset Details*

The dataset used in the experimental part of this paper is the KorCCVi dataset in its extended version, herein referred to as KorCCVi v2. This dataset aligns with the newly data-centric AI approach, focusing on enhancing model performance by including high-quality and relevant raw data. The data-centric AI approach, introduced by Professor Andrew Ng, emphasizes the value of meticulously prepared high-quality data over the exclusive improvement of code and algorithmic methodologies, typically seen in the model-centric AI approach [18].

The KorCCVi v2 dataset comprises two classes, labeled as '1' for voice phishing and '0' for non-voice phishing. The voice phishing class includes transcripts from real-world Korean voice phishing phone calls, whereas the non-voice phishing class incorporates transcripts of regular day-to-day Korean conversations. To better illustrate the nature of these classes, Table A1 in the Appendix A provides an extract of the KorCCVi v2 Dataset with English Translation as the ground truth. This representation provides insights into the real-world examples for each class and acts as a guide to understanding the classification of our model. However, it is essential to clarify that the English translations are not part of the original dataset. They are presented alongside the Korean version solely to enable readers to understand the content of the transcripts.

The KorCCVi v2 dataset exhibits an imbalance in its data distribution, meaning that the number of samples per class is irregularly distributed. It comprises 695 samples within the voice phishing class and 2232 in the non-voice phishing class, totaling 2927 samples, as presented in Table 2.

**Table 2.** KorCCVi v2 dataset description.

| Source | Class (Label) | Samples | Percentage |
|---|---|---|---|
| FSS [1] | Voice phishing (1) | 695 | 23.7% |
| NIKL [2] | Non-voice phishing (0) | 2232 | 76.3% |
| | Total | 2927 | 100% |

[1] The Financial Supervisory Service of Korea (FSS); [2] The National Institute of the Korean Language (NIKL) website.

In addition, the whole dataset was split during the experiments as 80% for the training set and 10% for the test set. The validation set represents 10% of the training set data. Table 3 presents the dataset used in the experiment.

**Table 3.** Dataset distribution.

| Training Set | Validation Set | Test Set | Total |
|---|---|---|---|
| 2370 | 264 | 293 | 2927 |

### *4.2. Experimental Setup*

This subsection describes the hardware and software environment of the experiments conducted in this study.

In this paper, we conducted all the experiments on a computer running on Ubuntu 20.04.1 LTS, with an environment including CUDA 11.1, TensorFlow 2.8.0, and Python programming language. The proposed hybrid architecture and the selected baselines were processed on an NVIDIA GeForce RTX 3090 GPU.

For the word embedding build with FastText, we set the size of the embedding vector to 300. All the models were trained over 10 epochs using a batch size of 64, a learning rate of $1 \times 10^{-3}$, and a learning decay of $1 \times 10^{-10}$ optimized with the Adam optimizer. For regularization, we applied a spatial dropout set at 0.2 for the word embeddings and a dropout set at 0.1 for the 1D CNN and the attention layers. We use 32 convolutional filters with a unique kernel size of 3, and the output of the convolutional layer fits into Max pooling with a pool size of 2. The number of hidden units in the first LSTM layer is 64, and for the second LSTM layer, it is 32. We used a categorical cross-entropy loss function during the training. Early stopping was utilized to stop the model's training when there was no improvement in the loss value and, therefore, to prevent the model from overfitting the data and to maintain the best accuracy.

In this study, we did not optimize the hyperparameters to their optimal values but used fixed settings, as shown in Table 4. The reasons for this choice are multifold. Firstly, hyperparameter optimization can require substantial computational resources and time. Secondly, we aimed to prevent overfitting to our training data, thus preserving the model's generalization ability and its robustness. Finally, our goal was to highlight the efficacy of our proposed approach rather than the optimization of hyperparameters. This decision, however, does not preclude the possibility of enhanced performance through more rigorous hyperparameter tuning in future studies.

**Table 4.** The hyperparameters settings.

| Hyperparameters | Values |
|---|---|
| Word embedding vector dimension | 300 |
| Number of convolution filters | 32 |
| Convolutional kernel size | 3 |
| Number of Pooling | 1 |
| Pooling size | 2 |
| Number of Dropout | 2 |
| Spatial dropout rate | 0.2 |
| Dropout rate | 0.1 |
| Number of LSTM's hidden units | (64, 32) |
| Number of attention mechanism | 1 |
| Number of dense layers | 2 |
| Activation function type | ReLu, tanh, Softmax |
| Number of epochs | 10 |
| Batch size | 64 |
| Learning rate | $1 \times 10^{-3}$ |
| Learning decay | $1 \times 10^{-10}$ |
| Optimizer | Adam |

### *4.3. Baseline Models*

To rigorously evaluate the performances of our proposed attention-based 1D CNN-BiLSTM model, we juxtaposed it with four classical ANN baseline approaches: 1D CNN, LSTM, BiLSTM, and a hybrid approach, the 1D CNN-BiLSTM model. The main difference between our proposed approach and the 1D CNN-BiLSTM model is the absence of the attention mechanism layer in the architecture of the 1D CNN-BiLSTM baseline model. All the baseline models were trained with the same hyperparameter settings and the same dataset distribution. These baseline models, widely recognized in the field, provided a robust comparison for our proposed model's performance.

### *4.4. Evaluation Metrics*

Four evaluation metrics were selected to evaluate the performance of the Attention-based 1D CNN-BiLSTM model we propose in this paper and the other baseline models trained: accuracy, F1 score, precision, and recall.

The accuracy metric is the proportion of correct predictions over the total number of samples evaluated in the test set. Precision is the percentage of positive predictions, which is usually used along with recall, the ratio of positive instances that are correctly classified. The F1 score is the harmonic mean of precision and recall. These evaluation metrics are defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TP} + \text{FP} + \text{FN}} \tag{18}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{19}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

where in Equations (18)–(20), True Positives (TP) represent the number of positive samples correctly classified as positive, True Negatives (TN) represent the number of negative samples correctly classified as negative, False Negatives (FN) represent the number of positive samples incorrectly classified as negative, False Positives (FP) represent the number of negative samples incorrectly classified as positive.

### *4.5. Experiment Results Analysis*

Four classical baseline models (1D CNN, LSTM, BiLSTM) and a hybrid model (1D CNN-BiLSTM) were trained along with the proposed Attention-based 1D CNN-BiLSTM models for performance comparison purposes on the classification of Korean voice phishing attacks. During the training of all these models, the same KorCCVi v2 dataset and dataset distribution was adopted. Figures 3 and 4 and Table 5 show a comprehensive comparison of all the models' performance using previously defined evaluation metrics.

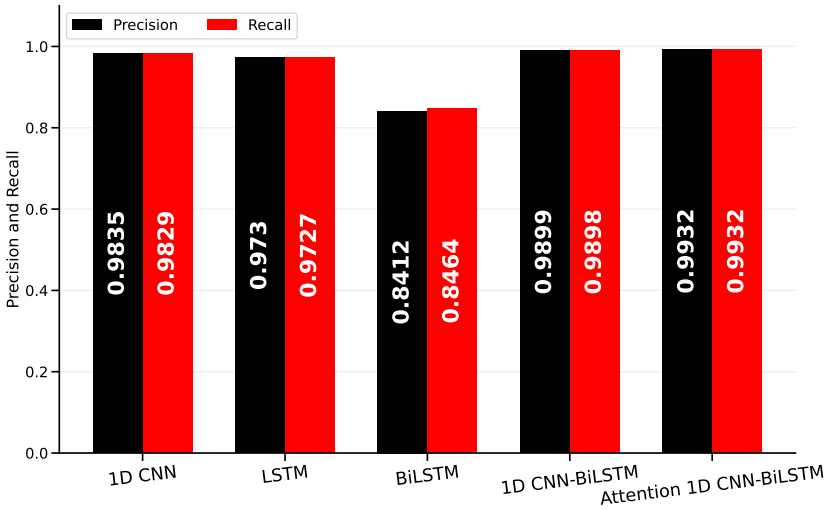


**Figure 3.** The comparison results of the precision and recall scores across baseline models.

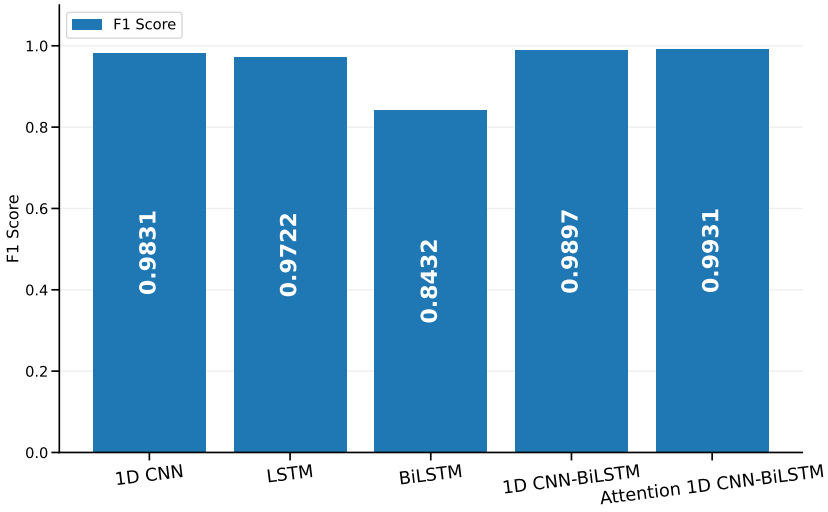


**Figure 4.** The comparison results of the F1 scores of the baseline models.

**Table 5.** Comparative performance metrics of the baseline models.

| Models | Evaluation Metrics | | | | Trainable Parameters | Training Time in Second |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 Score | Accuracy | | |
| 1D CNN | 98.35% | 98.29% | 98.31% | 98.29% | 429,756 | 18.61 |
| LSTM | 97.30% | 97.27% | 97.22% | 97.27% | 108,098 | 63.07 |
| BiLSTM | 84.12% | 84.64% | 84.32% | 84.64% | 232,386 | 143.34 |
| 1D CNN-BiLSTM | 98.99% | 98.98% | 98.97% | 98.98% | 149,436 | 131.69 |
| Attention-based 1D CNN-BiLSTM | 99.32% | 99.32% | 99.31% | 99.32% | 153,660 | 137.83 |

As presented in Figure 3, the 1D CNN and 1D CNN-BiLSTM models outperform others with closely matched precision and recall scores, suggesting their ability to maintain a balanced performance on both measures. Notably, our Attention 1D CNN-BiLSTM exhibits the highest scores on both precision and recall, indicating its excellent capability in correctly classifying voice phishing attempts and limiting the number of false positives.

Contrarily, the BiLSTM model has substantially lower precision and recall scores than other models, pointing to its limitations in effectively identifying voice phishing cases without raising too many false alarms. The LSTM model, although performing better than BiLSTM, still falls behind the 1D CNN-based models, reflecting the enhanced effectiveness of convolutional structures in processing voice phishing data.

The F1 score was also highest for the Attention 1D CNN-BiLSTM model, as presented in Figure 4. This further corroborates its superior performance observed in Figure 3. The 1D CNN-BiLSTM and 1D CNN models also show competitive F1 scores, while on the other end, the BiLSTM model confirms its underperformance with the lowest F1 score.

Table 5 presents that our proposed model leads with the highest score across all evaluation metrics, with an accuracy and F1 score of 99.32% and 99.31%, respectively. This superior performance results from the presence of the HAN in our architecture, providing more context to the model and significantly improves the classification performance. Besides, the 1D CNN-BiLSTM model trained without the attention mechanism showed a remarkable balance between performance and complexity. It achieved the second best performance among all the other baselines, with an accuracy of 98.98% and a F1 score of 98.97%. The hybrid models' overall performances are better compared to the classical models using DL algorithms. Hence, this result corroborates the potential of hybrid approaches when performing NLP tasks such as text classification.

The performance of the 1D CNN model is comparatively similar to the 1D CNN-BiLSTM model, with an accuracy of 98.29% and an F1 score of 98.31%. The model's capacity

to extract relevant features from the word embeddings can explain this good performance. Table 5 shows that the 1D CNN model demonstrates the fastest training time of 18.61 s among all models, despite its relatively larger number of trainable parameters (429,756). When comparing the complexity of the models, it becomes evident that the 1D CNN model has achieved a notable performance, making it an ideal competitor to our proposed model.

We observed that the model's complexity significantly impacts its performance. The BiLSTM model is a perfect example, as it has achieved the worst performance across all evaluation metrics compared to the other models. Moreover, it exhibited the highest training time of 143.34 s, despite having 232,386 trainable parameters, which is not the highest among the models. During the model training, we observed that the BiLSTM and LSTM models had their training stopped by the early stopping function because of poor improvement of the learning model. The progress of the validation accuracy and the validation loss over 10 epochs for all the compared models are presented in Figures 5 and 6, respectively.

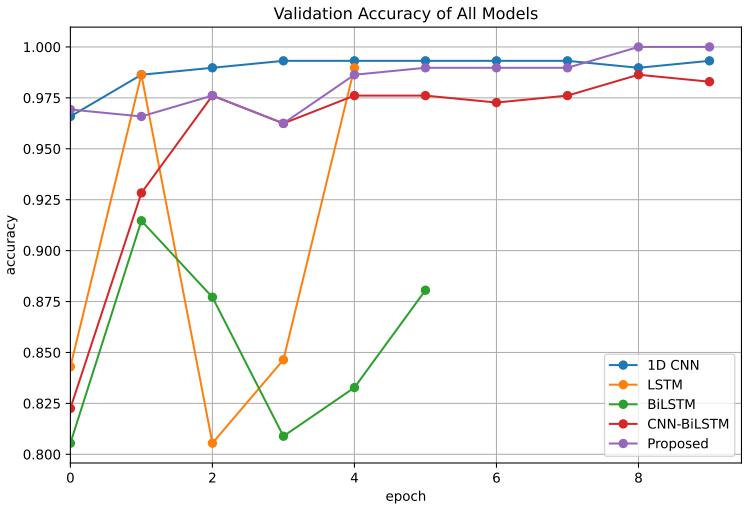


**Figure 5.** Validation accuracy vs, epochs of all models.

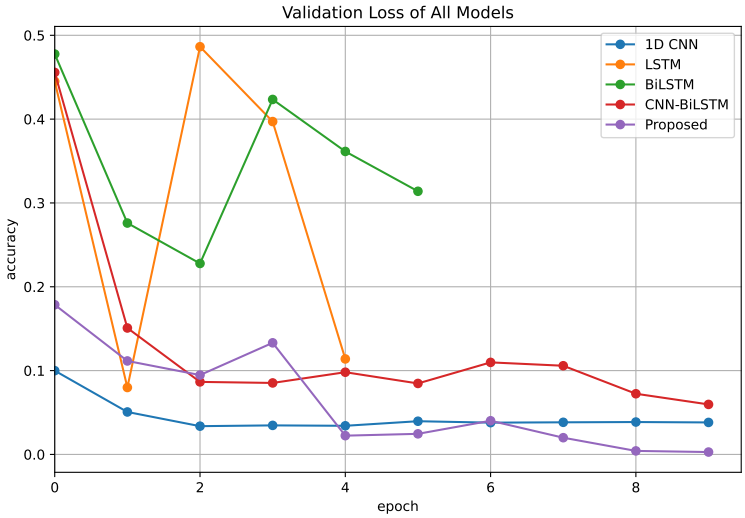


**Figure 6.** Validation loss vs. epochs of all models.

Taken together, our Attention 1D CNN-BiLSTM model demonstrates the most promising performance in voice phishing detection, whereas the BiLSTM model presents the least effectiveness among the assessed models. These results suggest that applying attention mechanisms and convolutional structures on LSTM could effectively enhance the performance.

### *4.6. Comparative Analysis*

The battle against voice phishing has attracted substantial research attention globally, yielding numerous strategies to mitigate this problem, as presented in Section 2. To perform a rigorous and unbiased evaluation of our proposed method and validate its results, we have only compared its performance with some existing methods specifically aimed at detecting voice phishing in Korea.

As presented in Table 6, the methods selected for this comparative analysis meet two crucial criteria: they address the issue of voice phishing in Korea and utilize the same evaluation metrics (i.e., F1 score, accuracy) that we employed to assess our proposed method. However, it is important to note that a direct comparison with some methods becomes challenging due to their reported results' lack of performance metrics (F1 score and accuracy). Nonetheless, the comparison provides valuable insights into the relative effectiveness of different approaches.

**Table 6.** Comparison with the existing approaches on Korean voice phishing detection.

| Ref. | Methods | Embeddings | Datasets (Total Samples) | Evaluation Metrics | |
|---|---|---|---|---|---|
| | | | | F1 Score | Accuracy |
| [5] | ML | None | Global phone book | - * | - * |
| [36] | Federated Learning | None | KorCCVi v2 (2927) | - * | - * |
| [32] | CNN, BiLSTM | None | - * | - * | - * |
| [6] | SVM, Logistic Regression, Decision Tree, Random Forest, XGB | TF-IDF | FSS + NIKL (2847) | 100% | 100% |
| [7] | Random Forest, XGBoost, LGBM, and CatBoost, Linear SVC, RNN, BiLSTM, GRU | TF-IDF, FastText | KorCCVi v1 (1218) | 99.43% | 99.45% |
| [8] | LSA, K-means | Doc2Vec, TF-IDF | FSS | 74% | 61% |
| [34] | KoBERT | KoBERT | KorCCVi v1 (1218) | 99.57% | 99.60% |
| [35] | KoBERT | KoBERT | FSS + AI Hub | - * | 97.86% |
| Ours | Attention-based 1D CNN-BiLSTM | FastText | KorCCVi v2 (2927) | 99.31% | 99.32% |

* Details not provided by the authors.

In addition, the basis of our comparison includes the following key factors. Firstly, the choice of the embedding technique is vital as it greatly affects the feature representation and, thereby, the model's overall performance. Secondly, the variance in the dataset sources employed by the compared methods offers a chance to examine the robustness and adaptability of our model. Specifically, all the studies, except for the research of Tran et al. [5], used the same primary source, the FSS, for their voice phishing samples. However, they adopted different sources for their non-voice phishing samples. This diversity, particularly in the source of non-voice phishing samples, contributes to our understanding of the model's effectiveness in handling diverse data. Lastly, the size of the dataset is considered, as our method employs a slightly larger but high-quality dataset.

## 5. Discussion of the Results

Our proposed model demonstrates a highly competitive performance compared to a range of existing methodologies in the field of voice phishing detection in Korea. Our model achieves an F1 Score of 99.31% and an accuracy of 99.32%. When delving into a more detailed comparison, several noteworthy observations arise.

Our proposed model demonstrates a highly competitive performance compared to a range of existing methodologies in the field of voice phishing detection in Korea. Our model achieves an F1 score of 99.31% and an accuracy of 99.32%. When delving into a more detailed comparison, several noteworthy observations arise. Three significant studies [5,32,36], despite not employing any embedding techniques or reporting specific performance metrics, collectively demonstrated a range of methodologies in voice phishing detection. These spanned from applying ML techniques to innovative privacy preservation strategies and

even traditional DL architectures. Considering these diverse perspectives, our study introduces advanced DL techniques alongside FastText embeddings. We have achieved meaningful performance enhancement by applying this novel combination to a dataset of real-world voice phishing data.

Lee and Park's work [6] used traditional ML models such as SVM and Logistic Regression along with a TF-IDF embedding, reporting a perfect score of 100% in terms of accuracy and F1 score. However, this extraordinary performance could be due to overfitting. Our model, on the other hand, implements an attention mechanism to reduce overfitting, resulting in more generalizable and robust performance. The same embedding technique and FastText embedding on the KorCCVi v1 dataset were employed in our previous work using various ML and DL models [7]. Although this study used a smaller dataset than the current one, it demonstrated the effectiveness of the approach, reporting an F1 score of 99.43% and an accuracy of 99.45%.

In contrast to our proposed method, the method of Kim et al. [8], which utilized LSA and K-means with Doc2Vec and TF-IDF embeddings on the FSS dataset, reported an F1 score of 74% and an accuracy of 61%. Among all the methods we compared, this demonstrates the lowest performance, further highlighting the advantages of our model, which leverages the strengths of DL to extract more complex and abstract features from the data.

The model suggested in our recent work [34] uses the language model KoBERT on the KorCCVi v1 dataset, achieving an F1 score of 99.57% and an accuracy of 99.60%. This study illustrates the strength of using pre-trained language models such as KoBERT. Remarkably, our proposed model, which uses FastText embeddings, shows comparable performance. Furthermore, Yang et al. [35] also used KoBERT on a dataset composed of FSS and AI Hub data. However, despite using the same pre-trained language model, their reported model's accuracy is slightly lower than our recent work [34] and the current study, with an accuracy of 97.86%. This contrast in results highlights the efficacy of our model over theirs, which combines 1D CNN, BiLSTM, and an attention mechanism.

These comparative analyses show the importance of suitable model architecture and feature extraction methods. It is also apparent that the quality and size of the dataset used for training can heavily influence a model's performance.

An essential aspect that validates the superiority of our proposed hybrid ANN architecture is our novel complementary approach combining data-centric and model-centric AI methodologies. This approach addresses the challenge of limited dataset availability by strategically leveraging our model architecture to extract high-quality features from the data. Moreover, our strategy further enhances the model's robustness by effectively using attention mechanisms to focus on key aspects of the input data. This results in a more accurate representation of the features relevant to Korean voice phishing detection. This method ensures that we are not only relying on the quantity of the data but also making the most of the quality of our data.

However, while our hybrid Attention-based 1D CNN-BiLSTM model requires greater computational resources and longer training times than simpler ML models, its comparative performance indicates the value of this trade-off. Our proposed method is a pioneering approach combining DL algorithms to enhance feature learning for voice phishing detection in Korea. Then unique advantage of our proposed method lies in its ability to capture both local and global contextual information from the input data , making it exceptionally effective for voice phishing detection in the Korean language.

## 6. Conclusions and Future Works

Despite their limits, classical approaches using Machine Learning or Deep Learning algorithms have shown acceptable performances in detecting Korean voice phishing attacks. Researchers have widely explored numerous hybrid artificial neural network approaches in NLP tasks to overcome these limits and to improve the model's performance. However, to our knowledge, no hybrid model is proposed to detect Korean voice phishing attacks.

This paper proposes an attention-based 1D CNN-BiLSTM model for detecting Korean voice phishing by classifying phone call transcripts. The hierarchical attention networks, the core component of the proposed architecture, significantly enhance model learning through word and sentence attention. The experiments were conducted on the KorCCVi v2 dataset, and the experimental results demonstrated that the performance of our proposed model is comparable to that of other baseline and hybrid models with 99.32% accuracy and 99.31% F1 score.

However, despite its superior performance, our proposed approach has limitations. The size of our dataset is relatively small, which may lead to a model underperforming in the real world. As our model is yet to be implemented into a mobile application for real-world testing, this can also be considered a limitation of our approach. Moreover, all the hyperparameters used during our experiments were randomly selected and fixed without any fine-tuning process. Using fine-tuned hyperparameters may slightly improve the performance of all the models compared in this paper. In our future work, we plan to increase the size of our dataset through data augmentation or by collecting real-world data through a federated learning approach. Additionally, to optimize the performance of our model, we will further investigate the use of multiple kernel sizes in the convolution stage and employ fine-tuned hyperparameters. Another later stage of our work will be to deploy our model on mobile applications (e.g., Android and iOS).

**Author Contributions:** Conceptualization, M.K.M.B. and D.-J.P.; methodology, M.K.M.B.; software, M.K.M.B.; validation, D.-J.P.; formal analysis, M.K.M.B.; investigation, M.K.M.B.; resources, M.K.M.B.; data curation, M.K.M.B.; writing – original draft, M.K.M.B. and D.-J.P.; writing – review and editing, M.K.M.B. and D.-J.P.; visualization, M.K.M.B.; supervision, D.-J.P.; project administration, D.-J.P.; funding acquisition, D.-J.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW (2018-0-00209) supervised by the IITP (Institute of Information & communications Technology Planning & Evaluation).

**Data Availability Statement:** All relevant data generated and analyzed during the current study have been made publicly available on GitHub at the following URL: https://github.com/selfcontrol7/Korean_Voice_Phishing_Detection (accessed on 10 May 2023). In addition, the codes used for the data preprocessing and model training in this study are publicly available in the directory Attention of the same GitHub repository.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Extract of the KorCCVi v2 Dataset along with the English Translation

**Table A1.** Extract of the KorCCVi v2 Dataset.

| ID | Transcript | Label |
|---|---|---|
| 2403 | 다만 아직까지 피해자라고 증명할 증거가 없으셔서 피해자 입증 조사 도와드리려고 하는 부분이고요. 본인 같은 경우는 1차적인 혐의점이 없으셔서 녹취 조사로 진행이 되실 겁니다. 본인을 대신해 법원에 제출될 서류기 때문에 주위 잡음이나 제3자가 있는 공간에서 녹취조사 하시면 안되시고요. 실례지만 지금 직장이신가요?네 여보세요? 여보세요? 저기 잘 안들리는데요.<br>However, there is no evidence to prove that you are a victim yet, so I am trying to help you investigate the victim. In your case, there is no primary suspicion, so the recording investigation will proceed. Since it is a document to be submitted to the court on your behalf, do not record and investigate in a space where there is noise around you or a third party. Excuse me, are you at work? hello? hello? I can't hear you well there. | 1 |
| 669 | 건물주들이 어 말도 안 되는 가격을 말도 안 되게 가격을 아 보증금을 올리고 있는 그런 얘기가 많이 나오는데요. 그 런 점에서 그게 맞는 합당하다고 생각하시나요? 합당하지 않다고 생각합니다. 건물주들도 세입자들이 있어야만 돈을 벌고 수익을 얻기 때문에 서로 공생관계에 있다고 생각합니다. 앗 너무 지나친 월세 인상은 세입자들에게 큰 부담을 느끼게 됩니다. 더군다나 요즘 같은 코로나 사태 때 소비가 줄어든 가 가게들에게 월세를 그대로 받는다는 것은 큰 재앙과도 같습니다. 그래서 건물주가 내려주는 세 세인 만큼 나라에서 그 세액을 공제해주기도 한 하는 정책을 시행하고 있습니다. [TRUNCATED]<br>There are a lot of stories about landlords raising their deposits at ridiculous prices. Do you think it is reasonable in that respect? I don't think it's worthy. I think that building owners are in a symbiotic relationship with each other because they make money and earn profits only when there are tenants. Oh, the excessive monthly rent increase puts a great burden on tenants. Moreover, it is like a big disaster to receive the monthly rent as it is from the shops that have reduced consumption during the corona crisis these days. Therefore, we are implementing a policy that allows the government to deduct the tax amount as much as it is the tax paid by the building owner. [TRUNCATED] | 0 |
| 2461 | 일단은 교육 김형석 주간문경 되었는데요. 증명서 일당들이 아직 되지 않았어요. 그렇기 때문에 또 다른 대포통장 바닐라 소득이 있어서 저희가 진행 중이고요네 그리고 금융감독원 뒤에서 이제 금일 내로 될 건데요 진행 안 했을 때 보니까 직접 계좌 말고 혹시나 본인께서 모르고 또 발견이 되잖아요. 그러면 금융감독원 보내는 모르는게 사건에 대해서는 불법 계좌로 없고요. 본인이 모르면 불법계좌 건에 대해서는 사건 종결 되기 전까지 진행하도록 할 건데요.<br>First of all, it became a weekly reading for education Kim Hyung-seok. The certificates haven't been done yet. That's why there is another cannon account vanilla income, so we're in the process. And behind the Financial Supervisory Service, it will be done today. Then, there is no illegal account for the case that the Financial Supervisory Service does not know. If you do not know, we will proceed with the illegal account case until the case is closed. | 1 |
| 1805 | 그 김정은이 트럼프한테 이 서로 원하는 내용을 논의해 보자고 했다는데 그 그거에 대해서 어떻게 생각해? 서로 원하는 내용을 논의해 보자는 거는 서로가 뭘 원하는지 알고 어느 정도 알고 있는 거 같고 그래서 뭔가 계속 글로 주고받으면은 해결되는 게 없을 것 같으니깐 서로 만나서 대화를 통해서 이 문제를 해결해 나가자는 것 같다고 생각해. 너는 계속 김정은의 이런 행동은 행동이 비핵화를 시키고 우리나라가 통일될 수 있을 거라고 생각해? 나는 잘 모르겠어. 뭔가 김정은이 이렇게 계속 한다고 해서 어떻게 일이 풀릴지도 잘 모르겠고 일이 더 커질 수도 있다고 생각해. 너는 이제 막 협상이 열리고 이러잖아. [TRUNCATED]<br>He Kim Jong-un asked Trump to discuss what he wanted with each other. He What do you think about that? Let's discuss what we want with each other seems to know what each other wants and know to some extent, so I think it's like we're going to meet each other and solve this problem through conversation because nothing seems to be resolved if we keep exchanging texts. Do you continue to think that Kim Jong-un's actions will lead to denuclearization and Korea to be reunified? I'm not sure. Just because Kim Jong-un continues like this, I don't know how things will work out, and I think things can get bigger. You've just opened negotiations and you're like this. [TRUNCATED] | 0 |

## References


1. Hernandez, J. That Panicky Call from a Relative? It Could Be a Thief Using a Voice Clone, FTC Warns. 2023. Available online: https://www.npr.org/2023/03/22/1165448073/voice-clones-ai-scams-ftc (accessed on 15 March 2023).
2. Stupp, C. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. 2019. Available online: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (accessed on 15 March 2023).
3. Brewster, T. Fraudsters Cloned Company Director's Voice In $35 Million Heist, Police Find. 2021. Available online: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=3c9c45107559 (accessed on 15 March 2023).
4. Jun-bae, S. [Crime Safety] Voice Phishing Status, Types, Trends and Implications for Countermeasures. 2022. Available online: https://kostat.go.kr/board.es?mid=a90104010311&bid=12312&act=view&list_no=422196&tag=&nPage=1&ref_bid= (accessed on 15 March 2023).
5. Tran, M.H.; Hoai, T.H.L.; Choo, H. A Third-Party Intelligent System for Preventing Call Phishing and Message Scams. In Proceedings of the Communications in Computer and Information Science, Online, 5–6 November 2020; Springer: Singapore, 2020; Volume 1306, pp. 486–492. [CrossRef]
6. Lee, M.; Park, E. Real-time Korean voice phishing detection based on machine learning approaches. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *14*, 8173–8184. [CrossRef]
7. Moussavou Boussougou, M.K.; Park, D.J. A Real-time Efficient Detection Technique of Voice Phishing with AI. In *Proceedings of the Korean Institute of Information Scientists and Engineers Korea Computer Congress*; Korean Institute of Information Scientists: Jeju, Republic of Korea, 2021; pp. 768–770. Available online: https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE10583070 (accessed on 20 May 2022).
8. Kim, J.W.; Hong, G.W.; Chang, H. Voice Recognition and Document Classification-Based Data Analysis for Voice Phishing Detection. *Hum. Centric Comput. Inf. Sci.* **2021**, *11*. [CrossRef]
9. Moussavou Boussougou, M.K.; Jin, S.; Chang, D.; Park, D.J. Korean Voice Phishing Text Classification Performance Analysis Using Machine Learning Techniques. In Proceedings of the Korea Information Processing Society Conference, Yeosu, Republic of Korea, 4–6 November 2021; pp. 297–299. [CrossRef]
10. Alpaydin, E. *Introduction to Machine Learning*; MIT Press: Cambridge, MA, USA, 2020.
11. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
12. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. Available online: http://www.deeplearningbook.org (accessed on 10 June 2023).
13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
14. Sarker, I.H. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* **2021**, *2*, 160. [CrossRef] [PubMed]
15. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.S.; Asari, V.K. A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics* **2019**, *8*, 292. [CrossRef]
16. Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M.P.; Shyu, M.L.; Chen, S.C.; Iyengar, S.S. A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
17. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
18. Ng, A. A Chat with Andrew on MLOps: From Model-Centric to Data-Centric AI. 2021. Available online: https://www.youtube.com/live/06-AZXmwHjo (accessed on 25 March 2021).
19. Hamid, O.H. Data-Centric and Model-Centric AI: Twin Drivers of Compact and Robust Industry 4.0 Solutions. *Appl. Sci.* **2023**, *13*, 2753. [CrossRef]
20. Moussavou Boussougou, M.K.; Park, M.G.; Park, D.J. An Attention-Based CNN-BiLSTM Model for Korean Voice Phishing Detection. In *Proceedings of the Korean Institute of Information Scientists and Engineers Korea Computer Congress*; Korean Institute of Information Scientists: Jeju, Republic of Korea, 2022; pp. 1139–1141. Available online: https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11113590 (accessed on 15 March 2023).
21. Basit, A.; Zafar, M.; Liu, X.; Javed, A.R.; Jalil, Z.; Kifayat, K. A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommun. Syst. Model. Anal. Des. Manag.* **2021**, *76*, 139–154. [CrossRef]
22. Tang, L.; Mahmoud, Q.H. A Survey of Machine Learning-Based Solutions for Phishing Website Detection. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 672–694. [CrossRef]
23. Goel, D.; Jain, A.K. Mobile phishing attacks and defence mechanisms: State of art and open research challenges. *Comput. Secur.* **2018**, *73*, 519–544. [CrossRef]
24. Aleroud, A.; Zhou, L. Phishing environments, techniques, and countermeasures: A survey. *Comput. Secur.* **2017**, *68*, 160–196. [CrossRef]
25. Das, A.; Baki, S.; El Aassal, A.; Verma, R.; Dunbar, A. SoK: A Comprehensive Reexamination of Phishing Research From the Security Perspective. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 671–708. [CrossRef]
26. Song, J.; Kim, H.; Gkelias, A. iVisher: Real-Time Detection of Caller ID Spoofing. *ETRI J.* **2014**, *36*, 865–875. [CrossRef]
27. Kang, Y.; Kim, W.; Lim, S.; Kim, H.; Seo, H. DeepDetection: Privacy-Enhanced Deep Voice Detection and User Authentication for Preventing Voice Phishing. *Appl. Sci.* **2022**, *12*, 11109. [CrossRef]

28. Derakhshan, A.; Harris, I.G.; Behzadi, M. Detecting Telephone-Based Social Engineering Attacks Using Scam Signatures. In Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics (IWSPA '21), New York, NY, USA, 26–28 April 2021; pp. 67–73. [CrossRef]
29. Jeong, E.S.; Lim, J.I. Study on Intelligence (AI) Detection Model about Telecommunication Finance Fraud Accident. *J. Korea Inst. Inf. Secur. Cryptol.* **2019**, *29*, 149–164. [CrossRef]
30. Zhao, Q.; Chen, K.; Li, T.; Yang, Y.; Wang, X. Detecting telecommunication fraud by understanding the contents of a call. *Cybersecurity* **2018**, *1*, 8. [CrossRef]
31. Peng, L.; Lin, R. Fraud Phone Calls Analysis Based on Label Propagation Community Detection Algorithm. In Proceedings of the 2018 IEEE World Congress on Services (SERVICES), San Francisco, CA, USA, 2–7 July 2018; pp. 23–24. [CrossRef]
32. Kim, W.W.; Kang, Y.J.; Kim, H.J.; Yang, Y.J.; Oh, Y.J.; Lee, M.W.; Lim, S.J.; Seo, H.J. Determination of voice phishing based on deep learning and sentiment analysis. In Proceedings of the Korea Information Processing Society Conference, Online, 14–15 May 2021; pp. 811–814. [CrossRef]
33. Kale, N.; Kochrekar, S.; Mote, R.; Dholay, S. Classification of Fraud Calls by Intent Analysis of Call Transcripts. In Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 6–8 July 2021; pp. 1–6. [CrossRef]
34. Moussavou Boussougou, M.K.; Park, D.J. Exploiting Korean Language Model to Improve Korean Voice Phishing Detection. *KIPS Trans. Softw. Data Eng.* **2022**, *11*, 437–446. [CrossRef]
35. Yang, J.; Lee, C.; Kim, S.B. Development and Utilization of Voice Phishing Prevention Service through KoBERT-based Voice Call Analysis. *KIISE Trans. Comput. Pract.* **2023**, *29*, 205–213. [CrossRef]
36. Yoon, J.Y.; Choi, B.J. Privacy-Friendly Phishing Attack Detection Using Personalized Federated Learning. In Proceedings of the Intelligent Human Computer Interaction, Copenhagen, Denmark, 23–28 July 2023; Zaynidinov, H., Singh, M., Tiwary, U.S., Singh, D., Eds.; Springer: Cham, Switzerland, 2023; pp. 460–465. [CrossRef]
37. Rothman, D. *Transformers for Natural Language Processing*, 2nd ed.; Packt: Birmingham, UK, 2022; p. 564.
38. Kiranyaz, S.; Ince, T.; Gabbouj, M. Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 664–675. [CrossRef]
39. Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D.J. 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **2021**, *151*, 107398. [CrossRef]
40. Junior, R.F.R.; dos Santos Areias, I.A.; Campos, M.M.; Teixeira, C.E.; da Silva, L.E.B.; Gomes, G.F. Fault detection and diagnosis in electric motors using 1d convolutional neural networks with multi-channel vibration signals. *Measurement* **2022**, *190*, 110759. [CrossRef]
41. Fang, Y.; Zhang, C.; Huang, C.; Liu, L.; Yang, Y. Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism. *IEEE Access* **2019**, *7*, 56329–56340. [CrossRef]
42. Huang, Y.; Yang, Q.; Qin, J.; Wen, W. Phishing URL Detection via CNN and Attention-Based Hierarchical RNN. In Proceedings of the 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), Rotorua, New Zealand, 5–8 August 2019; pp. 112–119. [CrossRef]
43. Zhou, Y.; Xu, J.; Cao, J.; Xu, B.; Li, C.; Xu, B. Hybrid Attention Networks for Chinese Short Text Classification. *Comput. Sist.* **2018**, *21*, 759–769. [CrossRef]
44. Hao, M.; Xu, B.; Liang, J.Y.; Zhang, B.W.; Yin, X.C. Chinese Short Text Classification with Mutual-Attention Convolutional Neural Networks. *ACM Trans. Asian -Low-Resour. Lang. Inf. Process.* **2020**, *19*, 1–13. [CrossRef]
45. Deng, J.; Cheng, L.; Wang, Z. Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification. *Comput. Speech Lang.* **2021**, *68*, 101182. [CrossRef]
46. Jang, B.; Kim, M.; Harerimana, G.; Kang, S.u.; Kim, J.W. Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism. *Appl. Sci.* **2020**, *10*, 5841. [CrossRef]
47. Kamyab, M.; Liu, G.; Rasool, A.; Adjeisah, M. ACR-SA: Attention-based deep model through two-channel CNN and Bi-RNN for sentiment analysis. *PeerJ Comput. Sci.* **2022**, *8*, e877. [CrossRef]
48. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
49. Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; Mikolov, T. Learning Word Vectors for 157 Languages. *arXiv* **2018**, arXiv:1802.06893.
50. Mikolov, T.; Grave, E.; Bojanowski, P.; Puhrsch, C.; Joulin, A. Advances in Pre-Training Distributed Word Representations. *arXiv* **2017**, arXiv:1712.09405.
51. Uysal, A.K.; Gunal, S. The impact of preprocessing on text classification. *Inf. Process. Manag.* **2014**, *50*, 104–112. [CrossRef]
52. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2019**, *21*, 1–67.
53. Google. Speech-to-Text: Automatic Speech Recognition. Available online: https://cloud.google.com/speech-to-text (accessed on 23 May 2022).
54. Moussavou Boussougou, M.K. An Artificial Intelligent Approach to Detect Voice Phishing Crime by Analyzing the Call Content: A Case Study on Voice Phishing Crime in South Korea. Master's Thesis, Soongsil University, Seoul, Republic of Korea, 2021. Available online: https://www.riss.kr/link?id=T15765846 (accessed on 23 May 2022).

55. Kudo, T. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. 2005. Available online: http://taku910.github.io/mecab/ (accessed on 23 May 2022).
56. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef]
57. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical Attention Networks for Document Classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489. [CrossRef]