IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 January 2025, accepted 15 February 2025, date of publication 24 February 2025, date of current version 4 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3545250

RESEARCH ARTICLE

Enhancing Voice Phishing Detection Using Multilingual Back-Translation and SMOTE: An Empirical Study

MILANDU KEITH MOUSSAVOU BOUSSOUGOU[1], PRINCE HAMANDAWANA[2], AND DONG-JOO PARK[3]

[1]Department of Computer Science and Engineering, Soongsil University, Seoul 06978, Republic of Korea
[2]Department of Software and Computer Engineering, Ajou University, Suwon 16499, Republic of Korea
[3]School of Computer Science and Engineering, Soongsil University, Seoul 06978, Republic of Korea

Corresponding author: Dong-Joo Park (djpark@ssu.ac.kr)

This work was supported by the Ministry of Science and ICT (MSIT), South Korea, through the National Program for Excellence in SW supervised by the Institute of Information and Communications Technology Planning and Evaluation (IITP) under Grant 2024-0-00071.

ABSTRACT With the widespread global trend of voice phishing or vishing attacks, the development of effective detection models using artificial intelligence (AI) has been hindered by the lack of high-quality and large volumes of data. This lack of data reflecting a real vishing scenario often leads to imbalanced datasets and biased detection models. Therefore, we present in this paper a data augmentation (DA) method for expanding the imbalanced Korean call content vishing (KorCCVi) dataset to address the existing data asymmetry problem and enhance the performance of Korean vishing detection. The proposed approach for DA involves using the back-translation (BT) method with three different intermediate languages: English, Chinese, and Japanese. The proposed method offers several advantages over the traditional synthetic minority oversampling technique (SMOTE), which is the main technique used to compare with our multilingual BT approach. Using these two DA techniques, several machine learning (ML) and deep learning (DL) models were trained on the original imbalanced dataset, the dataset balanced with SMOTE and its variants, and the dataset augmented with our method. We analyzed the impact of these DA methods on the performance of the models, demonstrated the benefits of each approach, and suggested the most suitable approach. The performance of the trained models was evaluated using the accuracy, precision, recall, and F1-score metrics. The experimental results demonstrated that the proposed multilingual BT method effectively expands the dataset while preserving its contextual and linguistic characteristics. The average performance of the models revealed that those trained on the augmented dataset outperformed the other models. They achieved F1-scores of 98.91% for the back-translated data, 98.14% for the original data, and 97.23% for SMOTE.

INDEX TERMS Back-translation, data augmentation, machine learning, natural language processing, SMOTE, voice phishing.

I. INTRODUCTION

The continuous proliferation of emerging high-speed Internet-driven technologies has led to a cultural shift in service delivery approaches. With the high penetration rates of smart mobile devices worldwide, traditional on-site service delivery has massively shifted to Internet-driven service delivery platforms, such as Internet banking, Fintech, online shopping, and other related technology-based services. Consequently, this has led to the rise of more sophisticated phishing attacks on users of such platforms, as personal and sensitive data are exchanged over the internet.

Phishing is a cybercrime in which an attacker exploits social engineering and technical strategies to persuade their target to provide personal and sensitive data that can be used to gain unauthorized access [1], [2], [3], [4]. One of the

The associate editor coordinating the review of this manuscript and approving it for publication was Chi-Yuan Chen.

most common types of phishing attacks is voice phishing, or vishing, in which fraudsters try to manipulate the victims over the phone to scam them. The evolution of technologies such as artificial intelligence (AI) and machine learning (ML) has led to more complex and automated vishing techniques such as wardialing [5] and audio-deepfake [6], making it more sophisticated to detect vishing attacks. This raises the need to build efficient, state-of-the-art (SOTA) ML mobile applications that dynamically detect vishing attempts in real-time during the lifespan of a call and alert users of the possibility of a vishing attack.

Because they are based on strict rules, traditional methods such as the call blacklisting approach [7] are not sufficient to find vishing attacks coming from automated ML sources that are more complicated. Some of the most recent solutions, such as HearMeOut [8], embed anti-vishing applications that trace the existence of malicious vishing applications on mobile phones. These anti-vishing apps track changes in call redirection behavior and call modification patterns. An action is taken to block or allow the call to pass the call initiation stage. However, the efficiency of such approaches is limited to detecting vishing using call initiation flow characteristics that would have been modified by the existence of vishing applications on a user's mobile phone. This usually fails when sources of vishing come from sources other than the vishing applications on the user's phone, such as wardialing techniques.

To improve existing vishing protection tools, several studies have suggested using ML and DL methods that can learn the details of the complex patterns of vishing sources on the fly and detect attacks in real time [9], [10], [11], [12], [13]. Nevertheless, these prior works suffer from the inherent challenges of the nature of existing vishing workloads: (i) the lack of a large available public dataset owing to the private nature of phone conversations. (ii) Owing to data-sharing laws, available datasets are specific to a small geographic area (country or region). (iii) The language used in region-specific datasets limits the type of ML techniques used for feature extraction when building ML vishing detection models. Consequently, because of these limitations, prior research adopted shallow ML models for vishing detection—the asymmetry between the training datasets and real-world inference results in poor vishing detection performance in real-time. Moreover, owing to the limited resources available on mobile devices, there is an urgent need for DL techniques that can enhance the efficiency of deep neural network (DNN) models. These enhancements aim to decrease the amount of memory required, lower energy usage, and simplify operations while ensuring the accuracy of voice phishing detection.

To address these challenges, there is a need for a solution that provides a comparatively larger dataset with respect to prior approaches, develops a more efficient DL model that extracts the features of the created dataset, and classifies the conversations in the dataset as either vishing or benign conversations. Data augmentation (DA) is the solution we investigated in this research, as it has proven to be efficient in handling class-imbalanced dataset issues, especially in domains such as image processing and signal analysis, by increasing the dataset's size and improving the performance of the models. However, DA techniques from these domains cannot be directly applied to textual data because of their potential impact on the syntax, grammar, and meaning of original data [14].

As voice phishing datasets are region-based and, therefore, language-specific, we selected South Korea as a case study because of the prevalence of this phenomenon, as summarized in Table 1. This table shows the continued threat of voice phishing and the need for more advanced detection methods. We therefore carefully constructed our proposed vishing dataset from the publicly available records of phone conversations from the Financial Supervisory Service of Korea (FSS) website [15] and the National Institute of the Korean Language (NIKL) website [16].

This research investigates the impact of two techniques, namely the back-translation (BT) DA method and SMOTE balancing method, in detecting Korean voice phishing using the Korean Call Content Vishing (KorCCVi) v2 dataset.

We summarize our contributions as follows:

1) We propose the use of a multilingual BT approach as a DA technique to expand the KorCCVi dataset. English, Japanese, and Chinese are used as intermediate languages to generate three augmented training sets, namely BT-Eng ($Kr \rightleftharpoons En$), BT-Chi ($Kr \rightleftharpoons Ch$), BT-Jap ($Kr \rightleftharpoons Ja$).
2) We constructed a larger dataset by combining the original training set and the three augmented training sets from the BT process (Original + BT-Eng + BT-Chi + BT-Jap). This expanded dataset was referred to as BT-All.
3) We used different augmented datasets from the BT step to train and compare several state-of-the-art ML and DL models. These include the Decision Tree (DT), LightGBM (LGBM), Random Forest (RF), Extreme Gradient Boosting (XGBM), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and Convolutional Neural Networks (CNN). We also used a balanced dataset from SMOTE methods and its variants to train these models.
4) We conducted an empirical analysis to assess the impact of BT data augmentation and SMOTE balancing methods on the model performance in detecting Korean voice phishing. We used evaluation metrics such as the F1-score, precision, recall, and accuracy.

The rest of the paper is organized as follows: Section II introduces the background and motivations of this research, presenting different DA techniques in natural language processing (NLP) to deal with class imbalance. Section III presents the proposed multilingual BT approach. Section IV explains the comprehensive experiments conducted, including the description of the dataset and evaluation of the models' performances on the BT and SMOTE datasets. The

TABLE 1. Grouped summary of voice phishing statistics in Korea (2016-2023)[1]. Aggregation of the total number of cases, damages, arrests, and the number of individuals arrested for both types of fraud (institution impersonation and loan fraud) over the given years.

Year	Total Cases	Total Damages (Billion KRW)	Total Arrests	Total Arrested Individuals
2016	17,040	1,468	11,386	15,566
2017	24,259	2,470	19,618	25,473
2018	34,132	4,040	29,952	37,624
2019	37,667	6,398	39,278	48,713
2020	31,681	6,000	34,051	39,324
2021	30,982	7,744	27,647	26,397
2022	21,832	5,438	24,522	25,030
2023	18,902	4,472	20,991	22,386

[1]Data provided by the Korean National Police Agency (KNPA) [17].

experimental results obtained using the BT and SMOTE methods are examined in Section V, and their values and limits are discussed in Section VI. Finally, we conclude this study in Section VII, where the main conclusion and some perspective work are provided.

II. BACKGROUND AND MOTIVATIONS

This section presents the most recent solutions for detecting voice phishing. The available options consist of conventional and AI-based approaches.

A. EXISTING VISHING DETECTION TECHNIQUES

Extensive prior research has attempted to mitigate vishing, which has proposed different detection and prevention solutions. One conventional and ineffective approach is to filter callers using a simple black-and-white list of phone numbers [7]. This straightforward blacklisting technique relies on analyzing call data graphs constructed using a Markov clustering algorithm to identify phone numbers associated with voice fraud in a cellular network.

The blacklisting method has been improved in several ways [18], [19], [20]. These improvements include ML-based frameworks that combine the traditional black-and-white list and heuristics with up-to-date ML methods. However, because vishing attacks are typically carried out using internet calling technologies such as voice over internet protocol (VoIP), attackers can use dynamically generated phone numbers, rendering the blacklisting strategy ineffective. Furthermore, attackers can easily conceal their true identities and induce them to answer calls by using the caller ID spoofing technique.

To address the challenges of the blacklisting approach, recent ML-based vishing detection approaches have adopted NLP techniques that involve speech recognition using speech-to-text APIs [9], [10], [11]. These techniques perform feature engineering by creating word or sentence embeddings that are fed into detection ML models for vishing classification. These ML-based vishing detection and prevention methods achieved good classification accuracy for vishing datasets. Nevertheless, despite the impressive results achieved by these studies, they suffer from various limitations, including the use of imbalanced datasets and insufficient attention given to the low-resource devices used by end users. Therefore, a clear pattern exists between all current and previous vishing detection solutions: (i) There is a lack of a vast publicly available dataset for training vishing detection models. (ii) Consequently, shallow ML models are used, which present an asymmetry problem between the training data and real-world inference data. This issue ultimately results in a low vishing detection accuracy in real-world applications. (iii) Their ideas cannot be implemented on less powerful mobile devices because no power optimization techniques have been incorporated into their approaches.

B. HANDLING IMBALANCED DATASETS IN NATURAL LANGUAGE PROCESSING

In NLP, the problem of class-imbalanced datasets is a common issue that has led several researchers to find ways to address it when performing tasks such as neural machine translation (NMT) and classification. According to recent survey papers, numerous approaches have been proposed to address this problem. There are data-level and algorithmic-level approaches. Data-level approaches employ diverse data resampling (data balancing) and DA methods to mitigate the extent of the imbalance. In contrast, in algorithmic-level approaches, ML or DL classifiers focus on modifying the learning process to handle the class imbalance by biasing the model's prediction toward the minority class.

Data resampling is an ensemble of methods used to modify a dataset to reduce differences between class sizes. Resampling techniques can be categorized as undersampling, oversampling, and hybrid techniques.

Undersampling methods consist of downsizing the majority class by deleting some instances from the dataset and making the dataset more balanced. The oversampling methods instead upsize the minority class of the dataset by adding new instances. Popular oversampling methods are the synthetic minority over-sampling technique (SMOTE) [21] and adaptive synthetic sampling (ADASYN) [22]. These methods have proven efficient in several studies, such as in toxic comment classification [23] and other classification tasks [24], [25]. Finally, hybrid methods combine undersampling and oversampling methods. They help achieve a balance between removing majority-class instances and creating minority-class instances.

DA techniques are widely recognized methods employed in the field of computer vision [26] to produce more training

data, thus addressing the issue of imbalanced and limited training data. Consequently, the performance of the trained models was enhanced. As recently discussed in survey papers [27], [28], [29], [30] and more specifically for the Korean language [31], different DA methods can be used for NLP tasks. One frequently used method is Easy Data Augmentation (EDA) [32], introduced by Wei and Zou. EDA has a variety of operations to select from and apply to sentences. This includes Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), and Random Deletion (RD).

In addition to EDA, Sennrich et al. introduced the back-translation technique, which is rapidly becoming popular in NLP, to tackle the issue of imbalanced class distribution in datasets. BT is the process of translating a text from the original language (source language) to an intermediate language (target language) and then translating it back to the original language using NMT [33], [34]. Ma and Li [35] used BT augmentation to expand the three Chinese datasets for text classification. The performance evaluation of their approach showed that BT effectively mitigated the uneven distribution of data and improved the classification performance. Vaishali and Ratnavel also created MTDOT [36], a multilingual translation-based discourse analysis method that can find offensive content in Tamil textual data. The MTDOT strategy outperformed the widely used SMOTE class balancing method, improving the F1-score by 65%. Similarly, the successful use of the BT technique combined with paraphrasing by Djamila and Md Saroar in detecting hate speech demonstrates the strengths of this method [37].

The previously mentioned solutions for handling imbalanced datasets are popular and have proven efficient in specific scenarios. After looking at these results, this paper's main contribution is to compare and contrast how well models trained using the BT and SMOTE data augmentation methods can improve Korean vishing detection.

C. CHALLENGES IN VISHING DETECTION

1) DATA ASYMMETRY PROBLEM

Two types of data asymmetry problems are associated with training vishing detection models: the train-inference data asymmetry and an imbalanced workload scenario.

- **Train-inference data asymmetry:** As previously mentioned, existing vishing detection solutions suffer from limited available data for model training. Regarding the variability of the nature of vishing attacks, the training data versus the data used at inference time present a huge disproportion. This usually results in high vishing classification accuracy for both the training and test sets but poor inference classification results. For example, the dataset used in [9] consists of only 140 non-scam and 75 scam conversations. A comparatively larger dataset is required to construct a robust and highly effective vishing detection and prevention model. Rather than shallow ML models, DL can also be developed to produce better vishing classification accuracy during training and inference times.
- **Imbalanced workload scenario:** When one of the classes in a classification problem has an extremely low number of samples, it causes the problem of imbalanced classes for model training. This is a typical case of vishing detection models in which the number of vishing conversation samples is much lower than that of non-vishing conversation samples, thus creating an imbalance between the labeled dataset of the two classes involved in the vishing detection models (non-vishing conversation and vishing conversation). There are positive and negative aspects to this type of data asymmetry. The positive aspect is that the number of non-vishing conversations is proportionally higher in real-world scenarios than the number of vishing conversations during calls. Training an imbalanced labeled dataset can sometimes provide a perfect scenario for real-world inference; however, this is incredibly challenging. The negatives of training a model with imbalanced labeled data are that the model will drastically overfit the much larger non-vishing samples to the extent that it will not be able to learn the real-world scenarios of the vishing data. This causes the model to perform poorly for real-time vishing detection. Traditional techniques, such as undersampling, oversampling, and SMOTE [21] or DA techniques, can be adopted to address imbalanced class problems.

2) LOW-POWER MODEL IMPLEMENTATION FOR BETTER INFERENCE

Training a robust and efficient vishing detection and prevention model requires large amounts of data and DNNs. As much as we want to train the large vishing workloads with DNNs, we should consider the intended less-powerful hardware that the model will be deployed on (primarily mobile phones). DNNs are computationally intensive and energy-hungry, as reflected in the example of VGG-16 [38], which requires 15 billion operations to classify a single image [39]. These computations require significant computational resources and lead to high energy costs. In this regard, a low-power computing vishing model that adopts techniques that allow DNNs to reduce memory requirements, energy consumption, and operations numbers must be constructed without significantly decreasing vishing classification performance.

Many prior works [40], [41], [42] have implemented ideas of low-power DNN models that can be adopted for vishing detection models with a few modifications. There are four categories of techniques presented in these prior studies: (i) Parameter Quantization and Pruning, (ii) Compressed Convolution Filters and Matrix Factorization, (iii) Network Architecture Search, and (iv) Knowledge Distillation. Overall, no single technique is superior to the others. Instead, they are complementary to building an efficient, low-power DNN.

The solutions mentioned above are practical approaches that can be investigated to develop an efficient DL model that matches mobile devices' resource constraints. However, they are only mentioned here as suggestions and are outside the scope of this study.

3) NO SINGLE-UNIFIED VISHING DETECTION MODELS

Thousands of languages are spoken across various geographical locations worldwide. Different languages have substantially diversified vocabulary and different forms of phrasing, inflections, and cultural norms. Consequently, the language modeling tools used are country- or geographically specific to define the problem or task adequately. Therefore, the techniques needed to build vishing detection models should be aligned with standard tools that best fit the geographic/country-specific language. As mentioned earlier in the introduction, South Korea has one of the highest prevalences of vishing crimes. This study uses it as a case study to construct a Korean-language-based vishing detection and prevention model. Hence, in this work, we provide insight into the knowledge of existing Korean tokenization APIs, such as those found in the Korean natural language processing Python package, KoNLPy [43], and associated models that best suit our task. We adopted strategies to build an efficient and fast Korean voice phishing detection DL model.

D. MOTIVATIONS

The motivation to address existing challenges arises from the increasing sophistication and frequency of voice phishing crimes in Korea. Looking at the existing work, traditional detection and prevention methods have proven inadequate for efficiently combating this issue. One of the reasons for this is the limited amount of genuine vishing data available for building robust and efficient AI-based detection systems. Therefore, there is an urgent need for innovative techniques that can effectively augment existing datasets while maintaining the integrity of underlying linguistic and contextual characteristics. This will allow us to train efficient models for detecting and preventing vishing crimes in Korea.

Our research was motivated by the desire to address some of the abovementioned challenges by using back-translation for data augmentation and adequate DL algorithms. This approach solves the dataset asymmetry problem and enhances the detection capabilities of our model for mobile phones. Through this research, we aim to develop an efficient and cost-effective model tailored explicitly for Korean voice phishing detection and, therefore, contribute significantly to the ongoing efforts to combat this type of crime in South Korea.

III. METHODOLOGY

Looking at the KorCCVi v2 dataset, training a model using the current state of the dataset can lead to a biased model with poor performance in real-world scenarios owing to its imbalanced nature. This study suggests using BT to expand the dataset, ensuring that the generated samples retain their meaning, to build a robust model capable of efficiently performing the downstream classification task of Korean voice phishing calls.

The augmented data produced by the BT method were used to train ML and DL models, which were further compared with ML and DL models trained using the SMOTE method to balance the dataset. For this purpose, in the early stages of our study, we split the KorCCVi dataset into training, validation, and test sets. Then, using only the training set, we generated synthetic data employing the BT method with English, Japanese, and Chinese as the intermediate languages. This process yields three different sets of augmented data. Following various data combinations, we then fed these to our ML and DL algorithms to train multiple models.

Similarly, we trained different ML and DL models using the SMOTE method. All the models trained were tested to see how well BT does at adding semantically preserved samples to the dataset and looking into how to improve the detection models. In the last stage of this work, after an empirical analysis of the BT and SMOTE methods, we provide insights into the most suitable method and the conditions under which they are more efficient in vishing detection.

In summary, our study aimed to train different ML and DL models using different combinations of data (augmented and original) and the SMOTE method to assess the effectiveness of the BT method in solving the data asymmetry problem for a better detection model. Fig. 1 illustrates the different steps of the methodology used in this study.

A. DATASET DESCRIPTION

The KorCCVi v2 dataset, which is an enhanced version of the KorCCVi dataset [12], was used in this study. It adopts a data-centric AI approach by integrating high-quality and pertinent raw data to enhance the model's performance. This strategy prioritizes meticulous data gathering, cleaning, and organizing to guarantee its accuracy and relevance to the issue domain.

The KorCCVi dataset is the result of our previous work, and it is affected by the second type of data asymmetry problem, which is called the imbalanced workload scenario, as described in Subsection II-C1 of this paper. Readers can consult our previous research for additional information to thoroughly understand the process of creating the KorCCVi dataset.

The dataset consisted of two distinct classes: the vishing class (represented as '1') and the non-vishing class (represented as '0'). The former consists of authentic phone call transcripts from Korean vishing scams, whereas the latter consists of everyday Korean talks. The KorCCVi v2 dataset comprises 2927 samples, with 695 samples belonging to the vishing class and 2232 samples from the non-vishing class.

Table 2 presents a comprehensive summary of the dataset used in our research and the data sources. We consider it suitable to utilize the KorCCVi v2 dataset for this investigation because it was derived from publicly available datasets

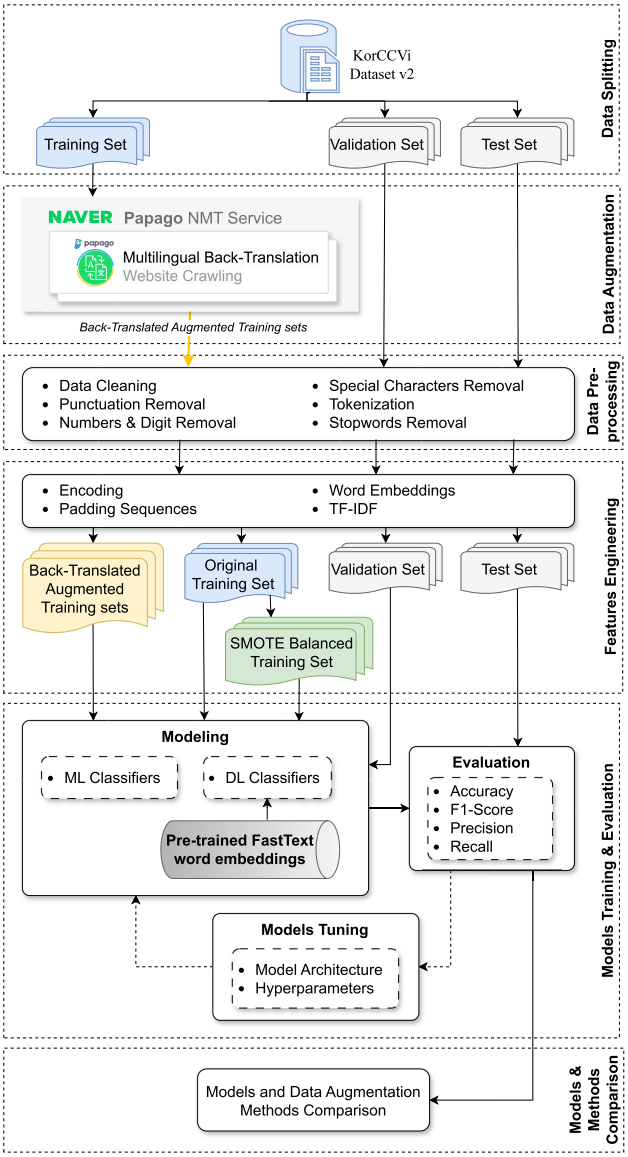


FIGURE 1. The flowchart of the proposed methodology.

that accurately depict a real-life context for detecting vishing attempts. Moreover, owing to its unbalanced character, it is a suitable candidate for evaluating the efficacy of the proposed DA strategy in enhancing the model's performance.

TABLE 2. Description of the KorCCVi v2 dataset.

Source	Class (label)	Samples	Distribution (%)
FSS[1]	Vishing (1)	695	23.75
NIKL[2]	Non-vishing (0)	2232	76.25
Total		**2927**	**100**

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

B. DATASET PREPROCESSING AND PREPARATION

To address the class imbalance of the KorCCVi dataset, we must first preprocess the dataset following common preprocessing steps when dealing with text classification tasks. These meticulous steps included data cleaning and tokenization.

Data cleaning involves eliminating extraneous or duplicate information from the original dataset. The process involves removing numbers, special characters, irrelevant symbols, punctuation marks, and personal information such as phone numbers. It also eliminates superfluous or duplicate information that does not contribute to the comprehension of vishing features.

To tokenize the dataset, we used the morphological analyzer MeCab-ko [44], which is known for its efficient morphological analysis in Korean. To use this method, the cleaned text was broken down into separate tokens or words, which were then used as the basic input units for our model. Furthermore, we removed the stop-words from the resulting data. The process involved the removal of Korean stop-words with minimal semantic significance in the context of vishing.

After preprocessing, we used the Term Frequency-Inverse Document Frequency (TF-IDF) technique to choose and extract the features from the dataset. The TF-IDF method assigns weights to each word based on its frequency within a specific document and its scarcity across all documents in the dataset. This enables us to accurately assess the significance of each word in differentiating vishing occurrences from genuine occurrences. In addition, we partitioned the dataset into training, validation, and test sets, which is a crucial procedure for assessing the model's performance and avoiding overfitting.

After completing the entire dataset preprocessing and preparation process, we applied DA techniques exclusively to the training set. This generates additional samples for the minority class and generates a balanced dataset. The preprocessing phase is essential for optimizing the dataset to enhance the effectiveness of the training and evaluation of our proposed DA strategy.

C. DATASET AUGMENTATION WITH MULTILINGUAL BACK-TRANSLATION

1) INTRODUCTION TO BACK-TRANSLATION

BT uses NMT to translate a text from its original (source) language to a target language and then back to its original language. This process generates synthetic parallel textual data while preserving the original context and meaning, thereby effectively increasing the size of the training dataset. Fig. 2 illustrates the steps involved in our proposed multilingual BT process, highlighting how it can be applied to augment textual data.

In the context of our research on detecting Korean vishing, where authentic vishing data are rare, and the asymmetric nature of the KorCCVi dataset, BT presents an appealing solution to augment our dataset. The detection of vishing strongly depends on understanding the intricacies of deceitful or suspicious conversations, and the limited availability of labeled data in this regard presents a notable obstacle.

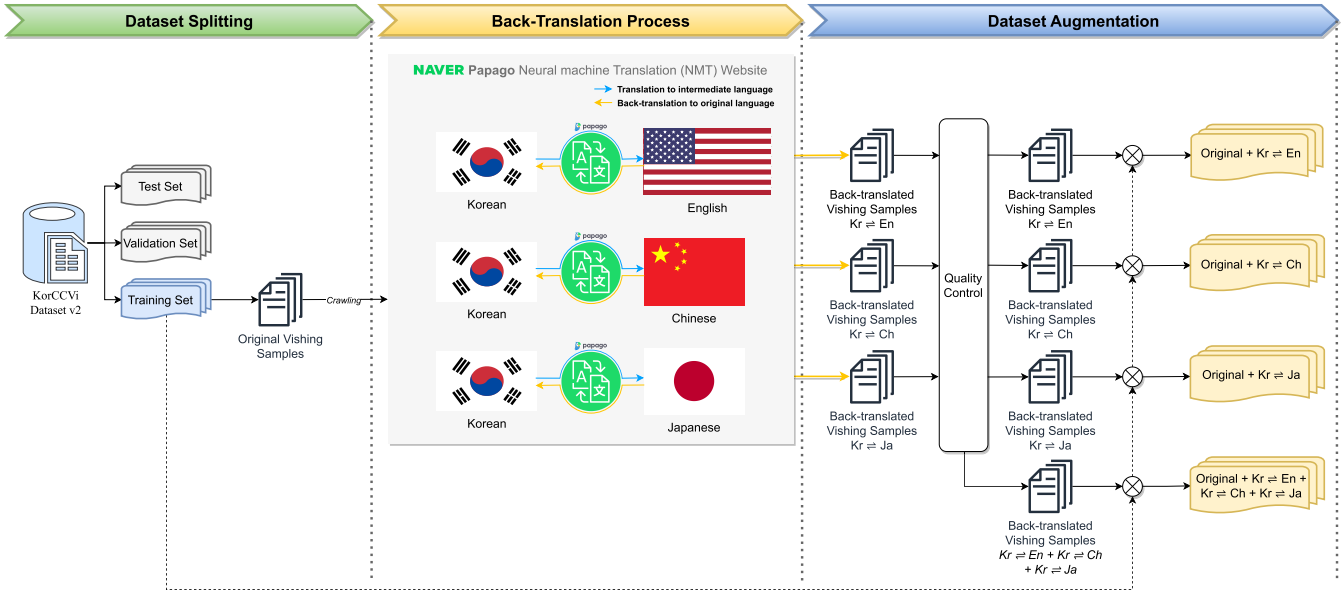


FIGURE 2. Data augmentation flowchart with multilingual back-translation.

By augmenting our dataset using BT, we can generate additional synthetic vishing samples that closely mimic real-world conversations, thereby providing our models with a wider range of unique instances to learn from. Therefore, we can introduce more diversity and variability into the data, thus balancing the distribution of classes and improving the generalization ability of the models. We justify our choice because BT has been proven effective in improving the performance of models in various NLP tasks [14], [45].

2) MULTILINGUAL BACK-TRANSLATION APPLIED TO KORCCVI V2 DATASET

To effectively augment the KorCCVi dataset and address the data asymmetry problem and imbalanced workload scenario, we performed the BT technique by leveraging the translation service Naver Papago [46].

Naver Papago, shortened to Papago, is a multilingual machine translation cloud service provided by Naver Corporation that uses NMT to translate text and speech. Unlike many other translators, Papago uses an NMT to learn from its mistakes and provides natural, context-sensitive translations. The multilingual BT augmented process of the training set is described in Fig. 2 and includes the following steps:

- **Step 1: Selection of Language Pairs:** We carefully selected suitable language pairs for the BT process to ensure the effectiveness of the augmentation process. We chose English as an intermediate language because of its status as a widely spoken language and the availability of reliable machine translation services. In addition to English, we included Chinese and Japanese for multiple reasons. Owing to their linguistic proximity and shared semantic and syntactic characteristics with Korean, these languages are ideal for preserving the meaning and context of sentences during translation. This decision aimed to generate diverse synthetic data and enhance the preservation of the context, nuances, and subtleties in Korean vishing conversations.
- **Step 2: Translation from Korean to Intermediate Languages Using Papago:** In the initial phase of BT, we used the Papago machine translation service through a website scraping process to translate the original Korean vishing training set from Korean to English, Chinese, and Japanese. This crucial step resulted in parallel datasets, in which each sentence in Korean had corresponding translations in these intermediate languages. Papago employs NMT technology, specifically a sequence-to-sequence (seq2seq) architecture [47] with an encoder-decoder and attention mechanism framework. First, the encoder component of the model encodes the input Korean sentences into a fixed-length vector representation, known as encoding. This vector contains the semantic meaning of the sentence. Then, the decoder component performs this encoding and generates a translation in the target language, such as English, Chinese, or Japanese. Papago incorporates an attention mechanism to capture complex linguistic dependencies and is trained on vast multilingual datasets, ensuring versatility. Because of this, Papago's NMT features helped us accurately translate the Korean vishing training set into intermediate languages, giving us high-quality synthetic data to add to the training set while retaining the meaning and context of the original Korean.
- **Step 3: Translation from Intermediate Languages to Korean (Back-Translation):** In the BT step, we employed the Papago service again to translate sentences from intermediate languages (English, Chinese, and Japanese) back into Korean. This step followed the

same technical approach described in the previous step. This is crucial as it generates synthetic Korean sentences that closely resemble the original data in terms of context and meaning.

- **Step 4: Quality Control:** During the quality control phase, we meticulously examined and assessed the synthetic Korean sentences produced in the preceding stages to guarantee their precision and dependability. Native Korean speakers manually reviewed and validated translated sentences for grammar, syntax, and coherence errors. In addition, we compared the synthetic sentences and the original data to verify that they preserved a high level of fidelity in terms of semantic and contextual accuracy. By implementing this rigorous quality control process, we successfully generated synthetic data of exceptional quality, which significantly improved the performance of our detection models.
- **Step 5: Augmentation of the Dataset:** Using the Papago service, back-translated Korean sentences from English, Chinese, and Japanese were added without any problems to the original training set. This enabled us to create four different combinations of datasets. For the first data combination, BT-Eng, the original training set was mixed with English back-translated samples ($Original + Kr \rightleftharpoons En$). The second combination, BT-Chi, was created by combining the original training set and Chinese samples that had been translated backward ($Original + Kr \rightleftharpoons Ch$). Combining the original training set with the back-translated sample from Japanese ($Original + Kr \rightleftharpoons Ja$) made the third combination. Finally, BT-All, the fourth combination, combines the original training set with back-translated samples from English, Chinese, and Japanese ($Original + Kr \rightleftharpoons En + Ch + Ja$). This augmentation process significantly increased the size and diversity of our training set, as it now contains both authentic vishing conversations and synthetic data.

This methodology is justified by its ability to diversify the training set while maintaining the underlying characteristics and context of the original data. By incorporating variations in language expressions, phrasing, and sentence structure, the BT method effectively addresses class imbalance and enhances model generalization.

In the following sections, we examine the benefits of using BT instead of traditional oversampling methods such as SMOTE. We also investigate the possible limitations and factors to consider when using BT for vishing detection data.

3) COMPARATIVE ANALYSIS: BACK-TRANSLATION VS SMOTE

In addition to its efficacy in handling class imbalance and enhancing model generalization, BT provides various benefits compared with conventional oversampling methods such as SMOTE:

- **Preservation of Contextual and Linguistic Structure:** The Papago service effectively maintains the context and semantic coherence of the original data when implementing BT. SMOTE relies on interpolation between existing samples, whereas BT introduces novel instances that are independent of the original data and preserves linguistic nuances and phrasing unique to vishing conversations. Although the SMOTE method can be effective for certain types of data, it may not work well for text data because of the intricacies of natural language, as the interpolation of text can lead to less meaningful and coherent samples [21].
- **Language Variation and Realism:** BT uses a larger parallel corpus than SMOTE, making high-quality synthetic samples possible. We added linguistic diversity by translating the data into English, Chinese, and Japanese and then back into Korean. This diversification enhances the training set by making it more representative of real-world situations and capturing the diversity and complexity of the minority class, leading to the development of more resilient models. Furthermore, the synthetic data produced by BT closely resemble genuine vishing conversations, enhancing the authenticity of the dataset.
- **Scalability:** BT enables effortless dataset expansion by generating numerous synthetic samples without replicating the existing data, unlike SMOTE. This scalability is particularly advantageous when acquiring real-world data is difficult or time-consuming. Furthermore, using multiple intermediate languages increases scalability by accommodating different linguistic contexts and allows us to train models on larger datasets, resulting in improved performance and generalization.

Nevertheless, it is essential to consider the possible disadvantages of employing BT on this vishing detection dataset:

- **Data Dependence:** The effectiveness of BT depends on the availability and accuracy of parallel data for translation, which may not always be easily accessible or reliable. If appropriate translations are not available, the effectiveness of this technique can be limited owing to its reliance on data. Moreover, BT can introduce noise, biases, or errors in the produced samples, which can harm the overall quality and dependability of the training data. This could affect the performance of vishing detection models.
- **Computational Resources:** Generating and training large amounts of translated data can be computationally and time-intensive. This could be a limiting factor for researchers with constrained resources, thus hindering the scalability and practicality of this approach.
- **Limited by Translation Service Scope:** The quality of back-translated data depends heavily on the accuracy of the translation service. BT may inaccurately translate domain-specific or highly specialized terminology present in vishing conversations because of translation

service limitations to specific languages or domains. Domain-specific adaptations are required to address this limitation.

In summary, BT, as implemented through the Papago machine translation service, offers distinct advantages over SMOTE methods for Korean vishing detection. It prioritizes contextual preservation, linguistic diversity, and the generation of realistic synthetic data. However, rigorous attention to the quality and diversity of the initial dataset is necessary for harnessing its full potential. Choosing the most accurate machine translation service is crucial for this approach. In subsequent sections, we evaluate the performance of our vishing detection models trained on the augmented dataset and provide insights into their efficacy.

D. BASELINE CLASSIFICATION ALGORITHMS

In the experimental part of our study, we used diverse ML and DL algorithms to train the baseline models using the original and augmented training sets to establish a performance benchmark for vishing detection. The algorithms we selected were Random Forest (RF), Decision Tree (DT), LightGBM (LGBM), eXtreme Gradient Boosting (XGB), long short-term memory (LSTM), Convolutional Neural Network (CNN), and Bidirectional Long Short-Term Memory (BiLSTM). Our choice of these algorithms is based on their proven effectiveness in various NLP tasks, specifically for text classification.

In addition to the baseline classification models, we investigated the effectiveness of different variants of the SMOTE method in addressing the class imbalance issue in our dataset. The variants we explored included Adaptive Synthetic (ADASYN) [22] and borderline-SMOTE (B-SMOTE) [48]. These specific variants of SMOTE were chosen because they focused on dataset samples that are more challenging to classify and generate potentially more informative samples. Moreover, they have been widely used to address classification with class-imbalance problems. This was done to evaluate their impact on balancing the dataset and their effectiveness in improving the detection of Korean vishing.

We used these baseline models and SMOTE variants as standards to evaluate the effectiveness of our vishing detection methodology. We compared the models' performance on back-translated augmented data with those trained using SMOTE augmentation methods. We can determine which methodology produces better outcomes in detecting Korean vishing by comparing the two augmentation strategies. Ultimately, this will contribute to the development of more robust and accurate detection models.

IV. EXPERIMENTS

In the empirical approach of this work, we trained a total of 128 models: 16 models with the original imbalanced training set, 48 models with the training set balanced via the SMOTE method and its variants, and 64 models with the training set augmented using multilingual BT methods. These models include using FastText word embedding to initialize the embedding layer in the DL architecture and improve the performance of the models. In the Tables, FastText word embedding models are distinguished by the suffix 'FT' in their names.

TABLE 3. Distribution of samples for the training, validation, and test sets.

Sets	Vishing (1)	Non-vishing (0)	Total
Training	485	1562	**2047**
Validation	105	335	**440**
Test	105	335	**440**
Total	**695**	**2232**	**2927**

TABLE 4. Description of the dataset after the augmentation and balancing steps. Imbalance Ratio (IR), Total number of samples, and distribution percentage per class with label 1 for vishing and 0 for non-vishing.

Training set	IR	Samples	Distribution (1, 0) (%)
Original set	3.22	2047	23.7, 76.3
SMOTE-Balanced Sets			
SMOTE	1	3124	50, 50
Borderline-SMOTE	3.22	2047	23.7, 76.3
ADASYN	1	3125	50, 50
BT-Augmented Sets			
BT-Eng	1.61	2532	38.31, 61.69
BT-Chi	1.61	2532	38.31, 61.69
BT-Jap	1.61	2532	38.31, 61.69
BT-All	0.8	3502	55.4, 44.6

A. EXPERIMENTAL SETTINGS AND DETAIL OF THE DATASETS

The KorCCVi dataset was split into training, validation, and test sets at a ratio of 70:15:15. The distribution of the samples per set after splitting the dataset is presented in Table 3.

The details of the training sets are listed in Table 4. This table calculates the training set's imbalance ratio (IR) using the formula outlined in Equation (1). This indicates the number of non-vishing instances outnumber the vishing instances. An IR value closer to 1 indicates a more balanced dataset, whereas a greater ratio indicates a higher imbalance.

$$IR = \frac{|N|}{|V|} \tag{1}$$

where $|N|$ represents the number of non-vishing instances in N, is the set of all non-vishing instances, and $|V|$ represents the number of vishing instances in V, and is the set of all vishing instances.

The ML and DL classifiers were trained on all previously listed training datasets using the hyperparameters presented in Tables 5 and 6. For each ML and DL classifier, the hyperparameters and classifier architectures were optimized through a comprehensive empirical process on the training datasets.

To ensure the training of robust models capable of effectively classifying our data, we employed various configurations to optimize the training process and the performance of the models. We used FastTest word embeddings to capture semantic connections between words and improve the model performance in the embedding layer of the DL classifier architecture. With a learning rate initialized

TABLE 5. The hyperparameters settings for the ML classifiers.

Classifiers	Hyperparameters	Values
DT	random_state	42
RF	random_state	42
	n_estimators	100
XGBoost	random_state	42
	max_depth	3
	learning_rate	0.1
	n_estimators	100
	early_stopping_rounds	10
LightGBM	random_state	42
	max_depth	-1
	learning_rate	0.1
	n_estimators	100
	early_stopping_rounds	10
	num_leaves	31
ADASYN, SMOTE	k_neighbors	5
	sampling_strategy	auto
	random_state	42
B-SMOTE	k_neighbors	5
	sampling_strategy	auto
	random_state	42
	m_neighbors	10

TABLE 6. The hyperparameters settings for the DL classifiers.

Hyperparameters	Values
Word embedding vector dimension	300
Learning rate	$[1 \times 10^{-2}, 1 \times 10^{-3}]$
Optimizer	Adam with Exponential Decay
Decay rate	0.9
SpacialDropout1D rate	[0.2, 0.3, 0.5]
Dropout rate	[0.2, 0.3, 0.5]
Number of epochs	10
Batch size	[16, 32]
Activation function type	ReLu, Sigmoid
Number of convolution filters	50
Convolutional kernel size	[3, 4, 5]
Number of LSTM's hidden units	[64, 32]

at 5×10^{-2}, we used a learning rate schedule applying exponential decay to decrease the learning rate during the training progressively. This helps stabilize the convergence of the model. We selected the Adam optimizer in our setting because of its efficiency and adaptive learning rate capabilities, which are suitable for our textual datasets.

Early stopping was employed to stop the model training if there was no improvement in the validation loss value, thus preventing overfitting and ensuring optimal accuracy. Because the ML classifiers LGBM and XGBoost also include this method, early stopping was used while training the classifiers. In addition to early stopping, we also incorporated *SpacialDropout*1*D* and standard dropout layers in the DL classifier architecture to further regularize the models by randomly dropping subsets of features and neurons, respectively. The training process of the DL models was conducted for 10 epochs, and various training batch sizes were investigated to identify the optimal parameters that balanced the training speed and memory usage. Finally, because our case study is a binary classification task, the *binary_crossentropy* loss function was paired with the sigmoid activation function in the output layer to distinguish between vishing and non-vishing text.

We applied SMOTE oversampling methods using the same default oversampling parameters for the ML and DL classifiers. We used the same *random_state* value for all the classifiers and SMOTE methods.

B. OVERVIEW OF THE BACK-TRANSLATED DATA

Following the process of back-translating the training set of the dataset using the three intermediate languages selected, English, Chinese, and Japanese, as explained in subsection III-C, it is important to overview the resulting back-translated data. Table 7 presents representative samples, including the original Korean transcripts, their back-translated versions across the three languages, and their corresponding English translations. The English translations are included only to help readers understand the meaning of the Korean transcripts and are not part of the dataset. This clear side-by-side comparison demonstrates how BT affects the textual structure depending on the intermediate language. For additional samples, refer to Table 14 in the appendix.

In Fig. 3 and Fig. 4, we highlighted the differences between the original and back-translated Korean transcripts using distinct colors. The red indicates removed words or phrases, the green indicates added content and the unhighlighted text shows unchanged segments. These visualizations enable a detailed analysis of how intermediate languages impact the text structure.

Furthermore, to evaluate the magnitude of the differences between the original and the back-translated Korean transcripts, we conducted a Semantic Textual Similarity (STS) [49] analysis. In the context of voice phishing, the augmented data must preserve the context and semantic integrity of the original text. STS evaluates how similar two texts are in terms of meaning. STS involves generating embeddings for all relevant texts and calculating their similarities. A higher similarity score indicates greater semantically similarity between the compared text. We used the Python module Sentence Transformers (a.k.a. SBERT) [50] , specially with *paraphrase* − *multilingual* − *MiniLM* − *L*12 − *v*2 [51] model. This model was selected due to its superior accuracy in non-English languages, optimal capacity for capturing cross-lingual semantic fidelity, and effective performance when processes such as BT or paraphrase are involved.

We computed the cosine similarity score of the transcripts using their vector representations to measure the semantic similarity between the original and back-translated transcripts. Cosine similarity is frequently used in such scenarios. It measures the cosine of the angle between two vectors in n-dimensional space, providing a metric for the degree of similarity between the texts. The cosine similarity scores of some of the data samples presented in Table 7 are depicted in Fig. 5. The formula is given in Equation (2).

$$CosineSimilarity = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \tag{2}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the vector representations of two text samples, $\mathbf{A} \cdot \mathbf{B}$ is their dot product, and $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ represent the magnitudes (or Euclidean norms).

TABLE 7. Comparison of Original and Back-Translated Korean Transcript Across Languages.

id	Original Korean Transcript	BT-Eng (Kr ⇌ *En*)	BT-Chi (Kr ⇌ *Ch*)	BT-Jap (Kr ⇌ *Jp*)
2400	네. 안녕하세요. 여기 서울중앙지검 김정현 수사관입니다.네. 안녕하세요. 네네. 네. 안녕하세요. 다름이 아니라 본인 개인 정보 유출 사건 관련해서 몇 가지 확인차 전화드렸고요. 네네 혹시 전라도 거주 42세 남성 이성민이라는 사람 알고 계시나요? 아니요. 모릅니다. 저희가 이번에 이성민을 중심으로 한 금융범죄 사기단 검거 당시 현장에서 우리은행, 하나은행 통장 각 하나씩 발견됐는데 통장 관련해서 아시는 거 없으세요? 네. 없습니다. 네. 저희가 확인했을 때는 16년 9월 경기도 광명에서 발급이 됐던데 본인이 발급받으신 거 아니에요? Yes. Hello, this is Investigator Kim Jung-hyun at the Seoul Central District Prosecutors' Office. Yes, hello. Yes, hello. I'm calling to confirm a few things about your personal information leakage. Do you know Lee Sung-min, a 42-year-old man living in Jeolla-do? No, I don't. We found each bank account of Woori Bank and Hana Bank at the time of the arrest of the financial crime fraud group led by Lee Sung-min. Do you know anything about the bank account? Yes, I don't. When we checked, it was issued in Gwangmyeong, Gyeonggi-do in September 2016, but didn't you get it yourself?	안녕하세요. 서울중앙지검 김정현 수사관입니다.네, 안녕하세요. 네, 네, 네. 안녕하세요. 개인정보 유출 관련해서 몇 가지 확인차 전화드렸습니다.전라도에 사는 42세 남성 이성민 씨를 아시나요?아뇨, 모르겠어요.이성민 씨를 중심으로 한 금융범죄 사기단을 검거했을 때 현장에서 우리은행 통장 1개와 하나은행 통장 1개가 발견됐습니다. 당신은 그 은행 계좌에 대해 아는 것이 있습니까?아니요, 없어요. 네. 저희가 확인해보니 2016년 9월에 경기도 광명에서 발행이 되었는데 직접 수령하지 않으셨나요? Hello, this is Investigator Kim Jung-hyun from the Seoul Central District Prosecutors' Office. Yes, hello. Yes, yes, hello. I'm calling to confirm some things about the leakage of personal information. Do you know Lee Sung-min, a 42-year-old man from Jeolla Province? No, I don't. When we arrested a financial crime scam group centered on Lee Sung-min, we found one Woori Bank account and one Hana Bank account at the scene. Do you know anything about the bank account? No, I don't. We checked and it was issued in Gwangmyeong, Gyeonggi-do in September 2016, didn't you receive it yourself?	네. 안녕하세요. 서울중앙지검 수사관 김정현입니다.네. 안녕하세요. 네네네.네. 안녕하세요, 다름이 아니라 본인 개인정보 유출 건에 대해 몇가지 확인차 연락드렸습니다.네, 전라도에 사는 42세 남성 이성민 씨를 아시나요?아니요, 몰랐습니다. 저희가 이번에 이성민씨를 중심으로 한 금융사기단을 검거하면서 현장에서 우리은행, 하나은행 통장을 발견했는데 통장에 대해 알고 계신게 있으신가요?네, 없습니다, 네.저희가 확인했을 때는 16년 9월에 경기도 광명에서 발급했는데 본인이 발급한 거 아닌가요? Yes. Hello, I'm Kim Jung-hyun, an investigator at the Seoul Central District Prosecutors' Office. Hello, yes, yes. Hello, I'm contacting you to check a few things about your personal information leakage. Yes, do you know Lee Sung-min, a 42-year-old man living in Jeolla-do? No, I didn't. We found Woori Bank and Hana Bank bank accounts at the scene while arresting a financial fraud group led by Lee Sung-min. Do you know anything about the bank accounts? Yes, there isn't. When we checked, it was issued by Gwangmyeong, Gyeonggi-do in September 2016, but didn't you issue it yourself?	네. 안녕하세요. 여기 서울중앙지검 김종현 수사관입니다.네. 안녕하세요. 네네. 안녕하세요 다름이 아니라 본인의 개인정보 유출 사건과 관련하여 몇 가지 확인차 전화드렸습니다.네네 전라도 거주 42세 남성 이성민이라는 사람 아세요?아니요, 몰라요.저희가 이번에 이성민을 중심으로 한 금융범죄사기단 검거 당시 현장에서 우리은행, 하나은행 통장을 각각 하나씩 발견하셨는데 통장 관련해서 아시는 것은 없으신가요?네, 없습니다. 네. 저희가 확인해보니 2016년 9월에 경기 광명에서 발행됐는데 본인이 발급된 거 아닌가요? Yes, hello. This is Kim Jong-hyun, an investigator at the Seoul Central District Prosecutors' Office. Yes, hello. Hello, I'm calling to confirm a few things about your personal information leakage case. Yes, do you know Lee Sung-min, a 42-year-old man living in Jeolla-do? No, I don't. At the time of the arrest of the financial crime fraud group led by Lee Sung-min, we found bank accounts of Woori Bank and Hana Bank at the scene. Do you know anything about bank accounts? Yes, there isn't. We checked and it was issued in Gwangmyeong, Gyeonggi-do in September 2016, and didn't you get it issued yourself?
2547	본인이 참고물을 보실 때 평소에 돈을 왜 찾느냐든지 계좌 이체 보낸다든지 물품을 구매했을 때 은행 직원이 이 돈을 왜 찾냐 왜 보내냐 왜 사냐고 질문한 적 있습니까? 개인 자산인데 그런 식으로 할 이유가 없다는 겁니다. 그 질문을 하시는 상황있으면 그런 거 아니다 선 조치를 하신 후에 은행 업무를 다 보시고 나오셔서 본 검사에게 그 분의 성함이라든지 직급을 저에게 말씀해 주시면 됩니다 When you look at the reference materials, have you ever asked why a bank employee is looking for money, why are you sending it, or why are you sending it to a bank when you purchase an item? It's a personal asset, and there's no reason to do it like that. If you're asking that question, it's not like that. After taking precautions, you can finish your banking and come out and tell the prosecutor about the person's name or rank	참고자료를 보면 보통 왜 돈을 찾는지, 계좌이체를 보내는지, 물건을 살 때는 왜 이 돈을 찾는지, 왜 보내는지, 왜 사는지 물어본 적이 있습니까? 그것은 개인적인 자산이고, 그런 식으로 할 이유가 없습니다. 그런 질문을 하는 상황이라면 그렇지 않습니다. 당신이 은행 업무를 마친 후에, 당신은 검사에게 그의 이름이나 직위를 말할 수 있습니다 When you look at the references, have you usually asked why you're looking for money, why you're sending an account transfer, why you're looking for this money when you're buying something? It's a personal asset, and there's no reason to do it that way. Not if you're in a situation where you're asking those questions. After you finish banking, you can tell the prosecutor his name or position.	본인이 참고물을 볼 때 평소에 돈을 찾거나 계좌이체를 하거나 물건을 구매할 때 은행 직원이 왜 이 돈을 찾는지, 왜 부쳤는지, 왜 샀는지 물어 본 적이 있나요? 개인 자산인데 이런 식으로 할 이유가 없다는 겁니다. 그 질문을 하시면 그게 아니라 일단 조치를 취하고 은행 업무를 마치고 나오셔서 본 검사님께 성함이나 직급을 말씀해 주시면 됩니다 When you look at the reference materials, have you ever asked why a bank employee is looking for money, why did you send it, and why did you buy it? It's a personal asset, but there's no reason to do this. If you ask that question, you can take action first and come out of the bank and tell the prosecutor your name or rank.	본인이 참고물을 볼 때 평소에 돈을 왜 인출하는지, 계좌이체를 보내거나 물품을 구입했을 때 은행 직원이 이 돈을 왜 찾는지, 왜 보내는지 질문한 적이 있나요? 개인 자산인데 그렇게 할 이유가 없다는 겨죠. 그 질문을 하시는 상황이 있다면 그렇지 않습니다. 먼저 조치를 취한 후 은행 업무를 다 보고 나오시고 본 검사님께 그분 성함이나 직책을 저에게 말씀해 주시면 됩니다 When you look at the references, have you ever asked why you usually withdraw money, why are bank employees looking for it when you send an account transfer or purchase an item, and why are you sending it? It's a personal asset, and there's no reason to do so. If you are in a situation where you are asking that, then no. You can take the first step, finish your banking and tell me the name or position of the prosecutor.
2707	네이버에 중고나라 사이트 아시죠? 대구지역 있으세요 아시죠? 사이트에서 김재원 일단은 그리 국가의 물건을 싸게 판다고 때물 올려놓고 제일은행 통장으로 입금을 동안 뒤에 작업 많이 추라고 소주 잠적한 상태고요. 물건을 보내지 않았기 때문에 피해자들이 발생을 했고 종필하자 7분으로 하면서 본인 앞으로 사귀자고 수고가 검수가 될 상황입니다. 지금 현재 전국적으로 멜로디 식구들만 180만 원이상 여유가 되어 있고요. 이번 사건의 1층 현미가 있으신 분들은 저희가 구속영장 발부에서 지금 구속 중에 있습니다. 자 여기서 1차 뭐냐면 통신기록 1년 6개월 처리를 했을 때 김병근 통화 내역 이라든지 문자내역이 발견되지 않았기 때문에 일체형인 없다고 생각한다. 전반적인 상황을 녹취록 진행 도와 드릴 거예요. 무슨 말인지 아시죠? 9일에 동의하십니까? Do you know the secondhand country site on Naver? Do you know if there is a Daegu area? Kim Jae-won posted a sale on the site to sell the country's goods cheaply and went into hiding to work a lot later while making a deposit to the bank account of Cheil Bank. The victims occurred because they didn't send the goods, and when they finished it, they decided to go out with them for 7 minutes, and the trouble is being received. Currently, Melody family members alone have more than 1.8 million won nationwide. Those who have brown rice on the first floor of this case are under arrest from the issuance of an arrest warrant. Now, the first thing here is that when I processed the communication record for a year and a half, I don't think there is any in-one because Kim Byung-geun's phone call details or text messages were found. I'm going to help you with the overall situation. You know what I mean, right? Do you agree on the 9th?	네이버 중고나라 사이트 아시죠? 대구에 한 곳이 있습니다. 알고 있죠? 김재원 씨는 현장에서 헐값에 나라 물건을 팔기 위해 매물을 내놓았고, 제일은행 통장에 입금하는 과정에서 피해자들이 잠적해 일을 많이 했습니다. 현재 180만 원이 넘는 돈은 멜로디 가족뿐입니다. 이 사건 1층에 현미가 있다면 현재 구속영장 발부로 구속 상태입니다. 여기서 첫 번째는 1년 반 동안 통신기록이 처리됐을 때 김병근 씨의 통화내역이나 문자메시지 내역이 발견되지 않았기 때문에 올인원은 아니라고 봅니다. 제가 전체 상황을 기록하는 것을 도와드리겠습니다. 제 말 무슨 뜻인지 아시겠어요? 9일에 동의하십니까? Do you know Naver's secondhand country website? There is one in Daegu. You know it, right? Kim Jae-won put up a sale to sell his country's products at a bargain price, and many victims disappeared and worked hard while depositing them into the bank account. Currently, the only money that exceeds 1.8 million won is the Melody family. If there is brown rice on the first floor of this case, he is currently under arrest for issuing an arrest warrant. The first one here is not an all-in-one because no phone calls or text messages were found when the communication records were processed for a year and a half. Let me help you record the entire situation. Do you understand what I mean? Do you agree with the 9th?	네이버 중고나라 사이트 아시죠? 대구 지역이 있나요? 홈페이지에서 김재원이 먼저 나라 물건을 싸게 판다는 이유로 제일은행 통장에 입금하는 동안 뒷일을 더 뛰기 위해 자취를 감췄고, 물건을 보내지 않아 피해자가 생겼고, 종료 7분이 지난 뒤 본인과 사귀느라 고생하는 상황도 있었습니다. 지금 전국적으로 멜로디 가족들만 180만 원 이상의 여유가 있습니다. 이번 사건의 1층 현미에 대해서는 저희가 구속영장 발부 당시 구속 중에 있었습니다. 자, 여기서 1년 6개월의 통신기록을 처리하면서 김병근의 통화내역이나 문자내역이 발견되지 않아 일체형이 없다고 판단했습니다. 전반적인 상황 녹음 도와드리겠습니다. 무슨 말인지 알지? 9일에 동의하십니까? Do you know Naver's secondhand country website? There is one in Daegu. You know it, right? Kim Jae-won put up a sale to sell his country's products at a bargain price, and many victims disappeared and worked hard while depositing them into the bank account. Currently, the only money that exceeds 1.8 million won is the Melody family. If there is brown rice on the first floor of this case, he is currently under arrest for issuing an arrest warrant. The first one here is not an all-in-one because no phone calls or text messages were found when the communication records were processed for a year and a half. Let me help you record the entire situation. Do you understand what I mean? Do you agree with the 9th?	네이버 중고나라 사이트 아시죠? 대구 지역 있어요.아시죠? 사이트에서 김재원, 일단은 그렇게 나랏물건 싸게 팔겠다고 매물로 내놓고 제일은행 통장에 입금하는 동안 나중에 작업을 많이 하라고 소주를 숨겨둔 상태고 물건을 보내지 않아 피해자들이 발생해서 종필하면 7분 만에 본인 앞으로 사귀려고 손이 많이 가는 상황입니다. 지금 현재 전국적으로 멜로디 가족만 180만원 이상 여유가 있습니다. 이번 사건의 1층 현미가 있는 분들은 저희가 구속영장 발부로 지금 구속 중입니다. 여기서 1차는 뭐냐 하면 통신기록 1년 6개월 처리를 했을 때 김병근 통화내역이나 문자내역이 발견되지 않았기 때문에 일체형은 없는 것 같습니다. 전반적인 상황 녹취록 진행 도와드리겠습니다. 무슨 뜻인지 아시죠? 9일에 동의합니까? You know Naver's secondhand country site, right? There's a Daegu area. You know, right? Kim Jae-won, while you put up a sale to sell your country's items cheaply and deposit them in your bank account, you keep soju hidden to work a lot later, and the victims don't send you the items, so you have to go out with them in 7 minutes. Right now, Melody's family alone has more than 1.8 million won nationwide. Those who have brown rice on the first floor of this case are under arrest for issuing an arrest warrant. The first one here is that there seems to be no such thing as Kim Byung-geun's phone call or text message details were found when the communication records were processed for a year and a half. I'll help you with the overall situation transcript. You know what I mean, right? Do you agree on the 9th?
2722	에 그 명의도용사건 때문에 몇 가지 사진 확인 차 연락 드릴 거거든요 혹시 알고 계십니까 남편 분이거든요 모르신다고요? 왜냐하면 얼마 전에 저희 검찰에서 주범으로 이날 금융범죄에 있는데 지금 그 당시 현장에서도 대량의 신용카드와 대포통장 드리벌 경제사업소 하였거든요. 에 그 중에서 지금 명의로 돼 농협이랑 신한은행 통장이 세상에서 통장으로 사용 중인 걸 제가 압수수색했습니다. 에 뭐라고요 통장이 없어요 통장이나 카드를 없다고요 우리 지금 지금 근데 저희가 불법 현장에서 지금 신한은행 통장이 지금 현장에서. 자나요? 아닌가요 아 그래요 그럼 일단 제가 다시 한번 확인하고 전화 드릴게요 서울중앙지검 이요 에 예 아이 준호 수사관입니다 에 예 알겠습니다 다시 연락 주세요 에 일단 한번 확인해 보고 내일 전화 드릴게요. Yes, I'm going to contact you to check some pictures because of the theft of that name. Do you know that? It's your husband. You don't know? Because not long ago, the prosecution was the main culprit in the financial crime on that day, and at that time, there was also a large amount of credit cards and fake bank accounts at the scene. Yes, I raided and searched that Nonghyup and Shinhan Bank's bank accounts are being used as bank accounts in the world. Yes, what? I don't have a bank account. I don't have a bank account. We don't have a bank account or a card right now. Are we sleeping at the illegal site? No, okay. Then I'll check again and call you. This is the Seoul Central District Prosecutors' Office. Yes, yes, I'm Jun-ho Investigator. Yes, yes, okay. Please contact me again after checking. Yes, I'll call you tomorrow.	네, 도난 사건 사진 몇 장 확인하려고 연락드렸습니다. 혹시 아시나요? 그녀는 당신의 남편입니다. 모르신다구요? 왜냐하면 얼마 전까지만 해도 이날 금융범죄의 주범이 우리 검찰이었고 그 당시 현장에는 신용카드와 대포통장이 대량으로 있었습니다. 네, 그중 농협과 신한은행 계좌가 전 세계에서 은행 계좌로 사용되고 있다고 압수수색을 해봤습니다. 뭐라고요? 통장도 없고 카드도 없고 지금 불법 사이트에 신한은행 통장이 있습니다. 너 자요? 아니, 알았어요. 다시 한 번 확인해보고 전화드리겠습니다. 서울중앙지검입니다. 저는 준호 형사입니다. 네, 네. 알겠습니다. 다시 연락주세요. 네, 먼저 확인하고 내일 연락드리겠습니다. Yes, I'm contacting you to check some pictures of the theft. Do you happen to know? She is your husband. You don't know? Because not long ago, the main culprit of financial crimes on this day was our prosecution, and there were a lot of credit cards and fake bank accounts at the scene. Yes, among them, Nonghyup and Shinhan Bank accounts were used as bank accounts all over the world. What? I don't have a bank account, I don't have a card, I have a Shinhan Bank account on an illegal site right now. Are you sleeping? No, okay. I'll check again and call you. This is Seoul Central District Prosecutors' Office. I'm Detective Junho. Yes, yes. Okay. Please contact me again. Yes, I'll check first and contact you tomorrow.	네, 그 명의도용 사건 때문에 사진 몇 장 확인하려고 연락드렸던 거 알고 계셨어요. 남편분이신데 모르셨어요? 얼마 전까지만 해도 우리 검찰의 주범이 그날의 금융범죄였고 지금도 현장에서는 엄청난 양의 신용카드와 대포통장 사업자들이기 때문입니다. 네, 그중에서도 지금 명의의 농협과 신한은행 통장이 세계적으로 통장으로 사용되고 있다는 사실을 가압류했습니다. 네, 무슨 말씀이세요, 통장이 없고 통장이나 카드도대체. 저희는 지금 불법 현장에 있고 신한은행 통장은 현장에 있습니다. 잠들었어요?아니요..그럼 제가 확인해보고 전화드릴게요.저는 서울중앙지검 예예예준호 수사관입니다.네네 알겠습니다.다시 연락주세요.네,확인하고 내일 전화드릴게요. Yes, did you know that I contacted you to check some pictures because of the theft of that name, your husband, didn't you know? Not long ago, the main culprit of our prosecution was the financial crime of the day, and even now, they are a huge amount of credit card and fake bank accounts. Yes, among them, we seized the fact that Nonghyup and Shinhan Bank's bank accounts are being used as bank accounts worldwide. Yes, what do you mean, I don't have a bank account and my bank account or card is on the spot. We are at the illegal site and Shinhan Bank's bank account is on the spot. Did you fall asleep? No. Then I'll check and call you. I'm Ye-jun-ho, an investigator at the Seoul Central District Prosecutors' Office. Yes, okay. Please contact me again. Yes, I'll check and call you tomorrow.	네, 그 명의 도용 사건 때문에 몇 가지 사진 확인차 연락드립니다 혹시 아세요, 남편이거든요.모르신다고요? 왜냐하면 얼마 전 저희 검찰에서 주범으로 이날 금융범죄를 당했는데 지금 그 당시 현장에서도 대량의 신용카드와 대포통장 드리벌경제사업장을 했거든요. 네, 그 중에 지금 명의로 농협과 신한은행 통장이 세상에서 통장으로 쓰이고 있는 걸 제가 압수수색을 했습니다. 네, 뭐예요?통장이 없어요 통장이나 카드가 없는데 저희는 지금 불법 현장에서 신한은행 통장이 현장에서. 자고 있어요?아니에요? 아 그렇군요..그럼 제가 다시 확인하고 전화드릴게요 서울중앙지검입니다.네네 준호 수사관입니다 네 알겠습니다 또 연락주세요 네 일단 한번 확인해보고 내일 전화드릴게요. Yes, I'm contacting you to check some pictures because of the theft of that name. Do you know, it's your husband? You don't know? Because not long ago, the prosecution was the main culprit in financial crimes on that day, and at that time, they also ran a large number of credit cards and fake bank accounts at the site. Yes, I raided Nonghyup and Shinhan Bank's bank accounts under their names. Yes, what is it? I don't have a bank account. I don't have a bank account. I don't have a bank account or a card, but we have a Shinhan Bank account at the illegal site. Is it sleeping? No? Oh, I see. Then I'll check again and call you. This is the Seoul Central District Prosecutors' Office. Yes, this is Junho Lake. Yes, okay. Please contact me again. Yes, please check first and call me tomorrow.

FIGURE 3. Comparison of original and back-translated transcripts across languages (Sample id 2400). Highlighted differences: red for removed words/phrases, green for added content, and no highlighting for unchanged segments.

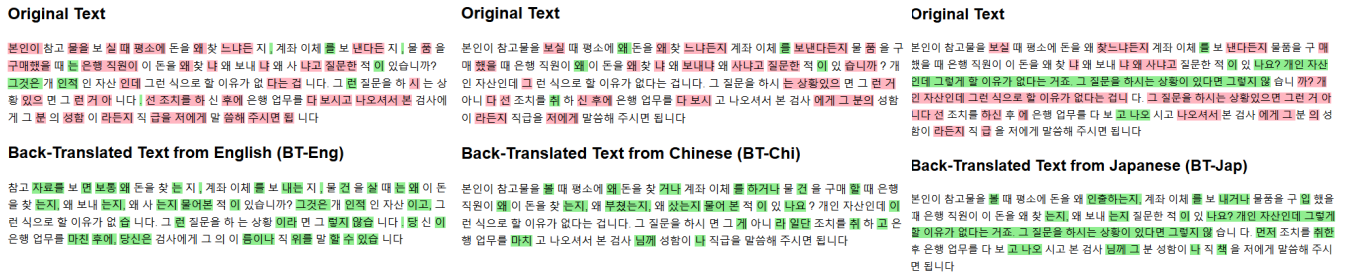


FIGURE 4. Comparison of original and back-translated transcripts across languages (Sample id 2547). Highlighted differences: red for removed words/phrases, green for added content, and no highlighting for unchanged segments.

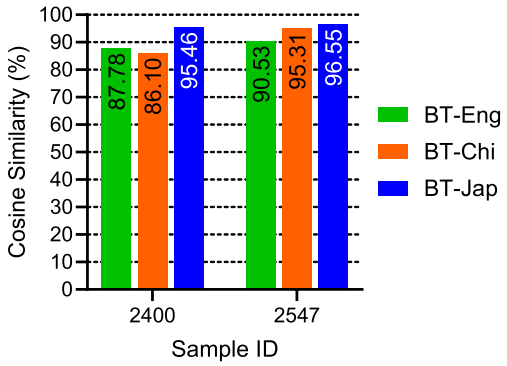


FIGURE 5. Cosine similarity scores between the original and back-translated transcripts for samples 2400 and 2547, using English (BT-Eng), Chinese (BT-Chi), and Japanese (BT-Jap) as intermediate languages.

C. PERFORMANCE EVALUATION

To evaluate the performance of the trained ML and DL models, we employed various evaluation metrics commonly used in the literature, such as accuracy (Acc), precision (Pre), recall (Rec), and F1-score (F1), which are detailed in previous studies [52].

The accuracy evaluates the model's overall accuracy across all classifications. This is mathematically defined in Equation (3).

$$Accuracy = \frac{TP + TN}{TP + TP + FP + FN} \tag{3}$$

Precision measures the accuracy of positive predictions, as given in Equation (4). This shows the model's accuracy for predicting positive instances among those predicted as positive.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall, also known as the model sensitivity, measures the model's strength to predict positive instances. Its formula is given in Equation (5).

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

The F1-score was calculated using a weighted harmonic mean between the precision and recall. This is calculated using Equation (6).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

In Equations (3)–(5), True Positives (TP) is the number of positive samples correctly classified as positive, True Negatives (TN) is the number of negative samples correctly classified as negative, False Negatives (FN) is the number of positive samples incorrectly classified as negative, and False Positives (FP) is the number of negative samples incorrectly classified as positive.

Additionally, we observed the training time of the models to track the total time taken for each model to train the dataset. This significantly affects the practicality of retraining and updating the models in real-world situations.

Because we deal with class-imbalance classification, the accuracy metric can be misleading. Therefore, the F1-score is the most suitable metric. This metric is crucial for vishing detection in Korea, where errors can have significant consequences.

V. RESULTS ANALYSIS

To assess the efficacy of the multilingual BT data augmentation technique for classifying Korean vishing on the imbalanced KorCCVi dataset, we examined how the performance of the models was affected when trained using the original imbalanced training set, the training set balanced with SMOTE methods, and training set augmented with BT. The performance comparison of the models was divided into three categories: the performance of models on the original imbalanced training set, the performance of models on the balanced training set using SMOTE-based methods, and the performance of models on the augmented training sets using the BT method.

A. PERFORMANCE OF MODELS ON THE ORIGINAL IMBALANCED TRAINING SET

Table 8 presents the evaluation metric values for the imbalanced training set. The overall observation shows the high performance of most ML and DL learning models across all metrics, with near-perfect scores in the F1-score metric. These results suggest that the models effectively classified the vishing data, even with the existing class imbalance.

Further analysis of the models' performances showed that all the CNN, RF models, and BiLSTM_FT models achieved a score of 100% across all metrics. The nature of the classifiers used for this classification task can explain these perfect scores, which may indicate overfitting. Tree-based classifiers such as RF and XGB are highly suitable for effectively handling non-linear datasets containing noise. By capturing intricate correlations and patterns within the data, performance is enhanced. However, the XGB model performed the worst across all metrics, with an F1-score of 87.96% and an accuracy of 94.77%.

The use of FastText word embedding during model training resulted in a relatively superior model compared to the model trained without it, as shown in Table 8. However, the trade-off when using FastText word embedding is the increase in the model training time, as testified by StackedBiLSTM_FT, which reached the highest training time of 491.5131536 seconds ($\approx$ 8 minutes).

B. PERFORMANCE OF MODELS ON THE BALANCED TRAINING SET USING SMOTE-BASED METHODS

As shown in Table 9, utilizing of SMOTE techniques to address class imbalance resulted in a slight decrease in performance across all metrics compared to the original imbalanced training set. This result suggests that although SMOTE methods can assist in tackling class imbalance, they may not always improve the model performance for this classification task.

The synthetic minority oversampling process might introduce some noise or generalize the vishing class too much, thus slightly impacting the model's precision and recall. This indicates a trade-off between attaining a balanced dataset and maintaining a high level of model performance.

From the results presented in Table 9, we can observe that the performance of the models varies significantly based on the SMOTE variants and the ML and DL classifiers used. Models such as B-SMOTE_BiLSTM achieved 0% F1, precision, and recall scores, indicating that certain SMOTE variants may not be appropriate for all types of models or require meticulous parameter adjustment to prevent a decline in performance.

When comparing SMOTE variants directly, detailed analysis and insights into the performance of models using these specific variants revealed their impacts on the models. Through this thorough analysis, we can understand their effectiveness and potential limitations in addressing class imbalance for vishing detection.

Table 10 presents the average (mean) performance metrics for each variant after categorizing the models by variant. The standard SMOTE method shows a notable improvement in the models' performance across all metrics out of the three variations, with an F1-score of 97.23% and a slightly higher recall score of 98.85% compared to ADASYN. This demonstrates its remarkable capability to identify the minority class (vishing cases) while upholding high overall accuracy. This can be explained by its balanced capacity to generate synthetic samples that fall well within the feature space of the minority class.

On the other hand, ADASYN demonstrates a higher F1-score of 95.29% than B-SMOTE while achieving similar scores on the different metrics. In contrast, the B-SMOTE variant had the lowest F1 and recall scores, of 91.41% and 89.82%, respectively. This indicates that B-SMOTE may not help detect vishing in specific contexts. Although both ADASYN and B-SMOTE improved model performance to various extents through their distinct impacts on model learning, their effectiveness was highly influenced by how they focused on different aspects of the minority class distribution during synthetic data generation.

Based on the results in Tables 9 and 10, it can be observed that among the three SMOTE variants evaluated, the standard SMOTE technique appears to be the most efficient in maximizing recall without significantly compromising accuracy or precision. Furthermore, it successfully mitigated the challenge of class imbalance with a low risk of overfitting. Consequently, standard SMOTE stands out as the preferred method in scenarios where failing to detect positive cases (false negatives) results in significant costs. This benefit arises from the fact that a higher recall score is correlated with a reduced number of undetected vishing attempts.

C. PERFORMANCE OF MODELS ON THE AUGMENTED TRAINING SETS USING THE BACK-TRANSLATION METHOD

Table 11 presents the performance of the models when trained on different training sets augmented using the BT method. This table shows that using the BT method to expand the dataset demonstrated the greatest variation in performance, with the four models achieving 0% scores across the F1-score, precision, and recall metrics.

TABLE 8. Performance results of models from the original training set.

Models	F1 (%)	Pre (%)	Rec (%)	Acc (%)	Train. Time (Sec)
BiLSTM	99.04	100	98.1	99.55	274.8567834
BiLSTM_FT	100	100	100	100	223.5027118
StackedBiLSTM	99.05	99.05	99.05	99.55	371.9534447
StackedBiLSTM_FT	99.53	99.06	100	99.77	491.5131536
LSTM	99.52	100	99.05	99.77	97.13456535
LSTM_FT	98.59	97.22	100	99.32	101.8722425
StackedLSTM	99.52	100	99.05	99.77	306.5574243
StackedLSTM_FT	99.52	100	99.05	99.77	218.8158045
CNN	100	100	100	100	22.63832974
CNN_FT	100	100	100	100	16.56168246
CNN_multiple	100	100	100	100	23.70107746
CNN_multiple_FT	100	100	100	100	10.99250627
DT	95.73	95.28	96.19	97.95	0.929406
LGBM	91.84	98.9	85.71	96.36	0.717216
RF	100	100	100	100	0.797958
XGB	87.96	97.67	80	94.77	0.751058

TABLE 9. Performance results of models from balanced training sets using SMOTE methods.

Models	F1 (%)	Pre (%)	Rec (%)	Acc (%)	Train. Time (Sec)
SMOTE_BiLSTM	95.41	92.04	99.05	97.73	410.6074381
SMOTE_BiLSTM_FT	98.56	99.04	98.1	99.32	472.3166828
SMOTE_StackedBiLSTM	96.23	95.33	97.14	98.18	906.4738114
SMOTE_StackedBiLSTM_FT	95.02	90.52	100	97.5	881.6497316
SMOTE_LSTM	94.44	91.89	97.14	97.27	111.2039208
SMOTE_LSTM_FT	97.67	95.45	100	98.86	231.3143926
SMOTE_StackedLSTM	98.56	99.04	98.1	99.32	238.5504985
SMOTE_StackedLSTM_FT	96.71	95.37	98.1	98.41	465.3939018
SMOTE_CNN	99.53	99.06	100	99.77	36.1450603
SMOTE_CNN_FT	99.05	99.05	99.05	99.55	17.10092044
SMOTE_CNN_multiple	98.59	97.22	100	99.32	23.86045241
SMOTE_CNN_multiple_FT	99.06	98.13	100	99.55	27.12808537
SMOTE_DT	98.1	98.1	98.1	99.09	1.000865
SMOTE_LGBM	94.06	90.35	98.1	97.05	0.485247
SMOTE_RF	100	100	100	100	1.073024
SMOTE_XGB	94.63	97	92.38	97.5	0.708748
ADASYN_BiLSTM	95.15	97.03	93.33	97.73	400.2655375
ADASYN_BiLSTM_FT	95.02	90.52	100	97.5	497.1454952
ADASYN_StackedBiLSTM	94.98	91.23	99.05	97.5	891.8344269
ADASYN_StackedBiLSTM_FT	91.23	84.55	99.05	95.45	779.4458518
ADASYN_LSTM	95.24	95.24	95.24	97.73	151.2420125
ADASYN_LSTM_FT	94.55	90.43	99.05	97.27	183.3880255
ADASYN_StackedLSTM	90.35	83.74	98.1	95	349.4510062
ADASYN_StackedLSTM_FT	90.04	82.54	99.05	94.77	507.1576664
ADASYN_CNN	99.52	100	99.05	99.77	40.59035587
ADASYN_CNN_FT	100	100	100	100	21.89558148
ADASYN_CNN_multiple	97.63	97.17	98.1	98.86	38.08505416
ADASYN_CNN_multiple_FT	98.56	99.04	98.1	99.32	33.50612307
ADASYN_DT	94.12	96.97	91.43	97.27	1.015568
ADASYN_LGBM	95.65	97.06	94.29	97.95	0.466931
ADASYN_RF	100	100	100	100	1.048213
ADASYN_XGB	92.61	95.92	89.52	96.59	0.730055
B-SMOTE_BiLSTM	0	0	0	76.14	163.7667863
B-SMOTE_BiLSTM_FT	88.89	100	80	95.23	135.2125974
B-SMOTE_StackedBiLSTM	99.52	100	99.05	99.77	638.4121702
B-SMOTE_StackedBiLSTM_FT	100	100	100	100	444.5153925
B-SMOTE_LSTM	98.55	100	97.14	99.32	160.9183226
B-SMOTE_LSTM_FT	100	100	100	100	179.0679028
B-SMOTE_StackedLSTM	99.04	100	98.1	99.55	259.4103954
B-SMOTE_StackedLSTM_FT	99.52	100	99.05	99.77	295.6391838
B-SMOTE_CNN	100	100	100	100	25.61211133
B-SMOTE_CNN_FT	100	100	100	100	20.19499874
B-SMOTE_CNN_multiple	100	100	100	100	27.71963596
B-SMOTE_CNN_multiple_FT	100	100	100	100	21.95742083
B-SMOTE_DT	97.17	96.26	98.1	98.64	1.260827
B-SMOTE_LGBM	91.84	98.9	85.71	96.36	0.439928
B-SMOTE_RF	100	100	100	100	1.130728
B-SMOTE_XGB	87.96	97.67	80	94.77	0.576587

These poor results can be interpreted as a failure or anomaly in training these specific classifiers with the augmented training set. This suggests that the augmentation procedure might have both positive and negative impacts on the models' performances, revealing the challenges of using BT for DA in vishing scenarios. Nevertheless, we can

TABLE 10. Average of performance metrics for models using the SMOTE variants.

Variant	F1 (%)	Pre (%)	Rec (%)	Acc (%)	Train. Time (Sec)
ADASYN	95.29	93.84	97.09	97.67	243.58
B-SMOTE	91.41	93.30	89.82	97.47	148.49
SMOTE	97.23	96.10	98.45	98.65	239.06

also observe the same results as the imbalanced training set for all CNN classifiers, which achieved F1-scores of 100%, suggesting overfitting.

To further evaluate the impact of each augmented training set, we categorized the models' performance based on the language used to augment the training set and then calculated and compared the average performance metrics for each category. Table 12 lists the average performance metrics for the ML and DL models trained using each BT language and their combinations.

When looking at the performance of individual languages, BT-Eng achieved the highest precision and F1-scores among them, with 93.09% and 91.54%, respectively. It can be deduced that employing English as an intermediate language in BT to augment our KorCCVi imbalanced dataset achieves a satisfactory trade-off between precision and recall. Like BT-Eng, BT-Chi showed a slightly lower precision score of 92.19%

As shown in Table 12, the combination of all three intermediate languages remarkably produced the highest scores across all metrics, with an F1-score of 97.06%, outperforming the individual language variants by a significant margin. These results demonstrate a synergistic effect, where using several languages introduces diversity in the training set, improving the generalization of the models and accurately detecting vishing samples.

The potential of multilingual BT to significantly improve the performance of models in detecting vishing was demonstrated by the BT-All training set. Combining English, Chinese, and Japanese seems to produce a more diverse and rich training set, increasing the robustness of the models owing to the introduced linguistic features. Furthermore, while individual languages demonstrate varying degrees of efficacy, the combined approach surpasses them. This highlights the benefits of utilizing diverse linguistic viewpoints in DA strategies with BT.

D. COMPARATIVE ANALYSIS OF MODEL PERFORMANCE ON DIFFERENT TRAINING SETS

To compare the models' overall effectiveness and performance impact using different DA methods, including the original training set, SMOTE variants for balancing, and BT for augmentation, we compared the best average performance metrics across all the methods. In addition to the calculated average of the performance metrics for the original training set, Table 13 displays the best methods selected from the SMOTE variants and BT methods previously discussed in Table 10 and Table 12, respectively. The best performance for each metric is highlighted in bold font. Fig. 6 provides a clear visualization of the average performance across different training sets from the different DA methods used.

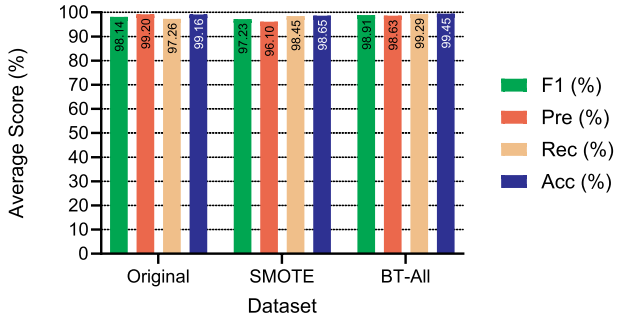


FIGURE 6. Average performance across the different training sets from the different DA methods used.

Despite the class imbalance of the original training set, it produced models with robust performance, providing a solid baseline. The models performed admirably well at detecting vishing. One potential explanation for this is that these models can recognize meaningful patterns. Alternatively, it could be that the nature of the data is such that even the minority vishing class provides sufficient information for effective learning. Furthermore, the models trained on the original training set exhibited the highest precision score of 99.20%. This means that the models made more accurate predictions of voice phishing when using the original training set. When compared, the precision score for the standard SMOTE decreased slightly to 96.10%, resulting in a slight reduction in the F1-score and accuracy. However, it improves the recall, highlighting its effectiveness in increasing the sensitivity of the models to the vishing class. Finally, combining all intermediate languages in BT-All outperformed the original training set across all metrics, achieving a 98.91% score for the F1-score metric. The BT augmentation approach significantly enhances the model performance using a mixture of all languages. However, its average training time indicates that it requires a higher computational cost than the original training set.

Table 13 also shows that there is a significant variation in the training time among the different DA techniques. The models trained using SMOTE or BT training data took longer to train, reflecting the increased complexity or size of the training set after balancing and augmentation, respectively. This training time parameter is important when selecting an appropriate DA technique.

VI. DISCUSSION

Experiments were conducted to assess the impact of the proposed multilingual BT augmentation method and SMOTE oversampling approaches in addressing the class imbalance issue while detecting Korean voice phishing. The comparative analysis of the experimental results highlights the potential of using multilingual BT as a robust strategy to improve the ML and DL models in vishing detection. This suggests that BT can be advantageous when dealing with data asymmetry problems, especially in scenarios with

TABLE 11. Performance results of models from augmented training sets using back-translation method.

Models	F1 (%)	Pre (%)	Rec (%)	Acc (%)	Train. Time (Sec)
BT-Eng_BiLSTM	0	0	0	76.14	190.2709141
BT-Eng_BiLSTM_FT	99.53	99.06	100	99.77	307.0828648
BT-Eng_StackedBiLSTM	96.59	99	94.29	98.41	806.3310618
BT-Eng_StackedBiLSTM_FT	100	100	100	100	751.7603724
BT-Eng_LSTM	98.58	98.11	99.05	99.32	209.8615992
BT-Eng_LSTM_FT	99.04	100	98.1	99.55	202.1593161
BT-Eng_StackedLSTM	98.11	97.2	99.05	99.09	232.7306464
BT-Eng_StackedLSTM_FT	90.05	100	81.9	95.68	231.1589048
BT-Eng_CNN	100	100	100	100	26.30193043
BT-Eng_CNN_FT	100	100	100	100	22.84735632
BT-Eng_CNN_multiple	100	100	100	100	25.98731446
BT-Eng_CNN_multiple_FT	100	100	100	100	24.83230925
BT-Eng_DT	97.12	98.06	96.19	98.64	0.290666
BT-Eng_LGBM	90.05	100	81.9	95.68	0.563866
BT-Eng_RF	100	100	100	100	0.91374
BT-Eng_XGB	95.61	98	93.33	97.95	0.695538
BT-Chi_BiLSTM	90.36	96.74	84.76	95.68	541.4114716
BT-Chi_BiLSTM_FT	100	100	100	100	899.9027667
BT-Chi_StackedBiLSTM	81.25	89.66	74.29	91.82	684.9375842
BT-Chi_StackedBiLSTM_FT	100	100	100	100	1714.177927
BT-Chi_LSTM	99.52	100	99.05	99.77	325.8149447
BT-Chi_LSTM_FT	100	100	100	100	411.1921742
BT-Chi_StackedLSTM	99.52	100	99.05	99.77	434.7488525
BT-Chi_StackedLSTM_FT	100	100	100	100	775.6249232
BT-Chi_CNN	100	100	100	100	68.95088792
BT-Chi_CNN_FT	100	100	100	100	59.86479592
BT-Chi_CNN_multiple	100	100	100	100	87.39903927
BT-Chi_CNN_multiple_FT	100	100	100	100	74.14881468
BT-Chi_DT	95.24	95.24	95.24	97.73	0.292645
BT-Chi_LGBM	0	0	0	76.14	0.399925
BT-Chi_RF	100	100	100	100	0.865057
BT-Chi_XGB	94.34	93.46	95.24	97.27	0.627985
BT-Jap_BiLSTM	99.52	100	99.05	99.77	153.1282606
BT-Jap_BiLSTM_FT	100	100	100	100	301.4052434
BT-Jap_StackedBiLSTM	99.52	100	99.05	99.77	744.415369
BT-Jap_StackedBiLSTM_FT	100	100	100	100	776.2948854
BT-Jap_LSTM	100	100	100	100	202.0255437
BT-Jap_LSTM_FT	100	100	100	100	123.844197
BT-Jap_StackedLSTM	0	0	0	76.14	223.824043
BT-Jap_StackedLSTM_FT	96.33	92.92	100	98.18	237.41927
BT-Jap_CNN	100	100	100	100	24.1204648
BT-Jap_CNN_FT	100	100	100	100	20.69722128
BT-Jap_CNN_multiple	100	100	100	100	26.61035633
BT-Jap_CNN_multiple_FT	100	100	100	100	23.15041876
BT-Jap_DT	97.14	97.14	97.14	98.64	0.33637
BT-Jap_LGBM	0	0	0	76.14	0.411856
BT-Jap_RF	100	100	100	100	0.848492
BT-Jap_XGB	96.19	96.19	96.19	98.18	0.586397
All_BiLSTM	98.08	99.03	97.14	99.09	367.6199691
All_BiLSTM_FT	100	100	100	100	510.3603396
All_StackedBiLSTM	99.52	100	99.05	99.77	1026.850421
All_StackedBiLSTM_FT	100	100	100	100	519.8209453
All_LSTM	99.04	100	98.1	99.55	171.9817603
All_LSTM_FT	99.05	99.05	99.05	99.55	143.4220402
All_StackedLSTM	98.56	99.04	98.1	99.32	263.0430174
All_StackedLSTM_FT	100	100	100	100	511.9564698
All_CNN	100	100	100	100	30.84882116
All_CNN_FT	100	100	100	100	27.49281001
All_CNN_multiple	100	100	100	100	34.8369658
All_CNN_multiple_FT	100	100	100	100	31.4400785
All_DT	99.52	100	99.05	99.77	0.673019
All_LGBM	89.66	81.89	99.05	94.55	0.468471
All_RF	100	100	100	100	1.32094
All_XGB	99.05	99.05	99.05	99.55	0.762876

imbalanced workloads. However, using back-translated data to train the models can lead to overfitting. Therefore, carefully selecting the intermediate languages for DA through BT is crucial.

TABLE 12. Average performance metrics for models trained using training sets from the multilingual back-translated method.

BT Method	F1 (%)	Pre (%)	Rec (%)	Acc (%)	Train. Time (Sec)
BT-Eng	91.54	93.09	90.24	97.51	189.61
BT-Chi	91.26	92.19	90.48	97.39	380.02
BT-Jap	86.79	86.64	86.96	96.68	178.69
BT-All	98.91	98.63	99.29	99.45	227.68

TABLE 13. Comparative analysis of models' performance across the data preparation techniques.

Training Set	F1 (%)	Pre (%)	Rec (%)	Acc (%)	Train. Time (Sec)
Original	98.14	**99.20**	97.26	99.16	**135.21**
SMOTE	97.23	96.10	98.45	98.65	239.06
BT-All	**98.91**	98.63	**99.29**	**99.45**	227.68

While the original training set forms a strong foundation for model training, its recall score must be improved. SMOTE and BT effectively addressed this limitation and demonstrated a superior recall score, showcasing their significant advantages in handling imbalanced datasets.

Nonetheless, while these results are promising, there are certain limitations to the proposed method that need to be addressed for future improvements. These limitations are discussed in the subsections below.

A. VARIABILITY ACROSS LANGUAGES FOR BACK-TRANSLATION

Our experiments revealed significant model performance variations depending on the BT process's intermediate languages. This research demonstrates that different languages introduce varying levels of semantic fidelity and augmentation quality during BT.

Table 11 highlights the differences in F1-scores, precision, and recall for models trained using BT with English, Chinese, and Japanese as intermediate languages. Despite English belonging to the Indo-European language family, with minimal linguistic similarity to Korean in terms of grammar or vocabulary, the results show that BT with English as the intermediate language consistently produced the most reliable improvements across all metrics. It achieved the highest F1-scores and precision values, which can be attributed to the robustness of the translation tools we used.

Conversely, Japanese BT introduced greater variability in model performance, which is unexpected given Japanese and Korean syntactic and morphological similarities. This variability might result from potential limitations in the translation service used, namely Papago. On the other hand, while sharing some vocabulary with Korean, Chinese BT demonstrated intermediate model performance. It balanced semantic fidelity and diversity while maintaining competitive model performance.

This variability highlights the importance of strategic language and translation service selection in the BT process. The choice of intermediate languages and translation tools can significantly impact the quality and effectiveness of the augmented dataset, thereby influencing model performance. To optimize the augmentation process, future work could explore automated methods for selecting intermediate languages based on their impact on model performance and investigate other translation services. Based on insights from Table 11, it is evident that a thoughtful choice of intermediate languages can significantly influence the robustness and effectiveness of the proposed method.

B. LIMITATIONS OF THE PROPOSED METHOD

Despite its promising results, the proposed method has some limitations that must be acknowledged.

1) IMPACT OF DATA VOLUME ON MODEL PERFORMANCE

One major limitation is the scalability of the method when applied to significantly larger datasets. While the multilingual BT approach effectively increases the size of the training data and mitigates class imbalance, it requires greater computational resources during the data augmentation and results in prolonged model training times.

This may provide difficulties in resource-limited settings, such as real-time implementation on mobile devices. Additionally, as the data volume increases, the likelihood of overfitting intensifies due to the inclusion of redundant or noisy back-translated data. Therefore, it is important to maintain the diversity and quality of the augmented data to ensure model performance at scale.

2) DEPENDENCY ON TRANSLATION SERVICES

The BT approach heavily relies on the accuracy and reliability of external translation services. Although high-quality systems such as Papago were utilized, these services lack optimization for domain-specific vocabulary, potentially leading to inaccuracies in the augmented data, especially for technical jargon associated with voice phishing.

Furthermore, our analysis revealed that the effectiveness of BT varies significantly across different language pairs. For instance, translations through languages sharing similar linguistic features with Korean showed lower model performance. However, we obtained higher model performance when using languages with substantially different grammatical structures or writing systems. This variability suggests that the selection of intermediate languages should be carefully considered based on both linguistic proximity to the source language and the specific requirements of the voice phishing detection task.

3) GENERALIZABILITY ACROSS LANGUAGES

This study primarily focuses on Korean voice phishing data, limiting the findings' applicability to other languages or multilingual contexts. The effectiveness of the proposed method in languages with limited translation tools has yet to be investigated, necessitating further research into cross-lingual applications.

4) REAL-WORLD APPLICABILITY

The experiments were conducted on a clean and structured dataset. However, real-world data frequently contain noise, incomplete transcripts, and mixed dialogues, which were not

TABLE 14. Comparison of original and back-translated Korean transcript across languages (Additional samples).

id	Original Korean Transcript	BT-Eng (Kr ⇌ *En*)	BT-Chi (Kr ⇌ *Ch*)	BT-Jap (Kr ⇌ *Jp*)
2694	잠원역에서 불법 인터넷 도박 사이트 권유했던 만 35세 남정인 김응술 한 사람 하시는 겁니까? 주무셔 왜냐면요? 제가 지난 달에 서울 마포 지역에서 재벌을 받고요 만 35세 남정림 김응수 일당을 건강했는데 현장에서 대량의 대포통장 보안카드 전화 드렸습니다. 메종엠오 니 명의로 된 신한은행 거 같고요. 농협 은행 계좌 포함이 되어서 사실관계확인서 제가 지금 연락드립니다. Are you talking about a 35-year-old male Jeong-in and Kim Yong-sul who recommended illegal Internet gambling sites at Jamwon Station? Are you sleeping because? Last month, I received a chaebol in Mapo, Seoul, and I called a large number of borrowed-name bank security cards at the site. I think it's Shinhan Bank under the name of Maison M.O. I'm contacting you now because the Nonghyup Bank account is included.	잠원역에서 불법 인터넷 도박 사이트를 추천한 35세 남성 김응술 씨를 하고 계십니까? 자는 이유는? 지난달 서울 마포에서 재벌을 받았는데 35살 남정림, 김응수 씨 일행과 함께 건강했는데요, 보안카드를 대거 들고 전화드렸습니다. 메종엠오 산하 신한은행인 것 같습니다. 농협은행 계좌가 포함된 사실확인서 관련하여 연락드립니다. Are you Kim Yong-sul, a 35-year-old man who recommended illegal Internet gambling sites at Jamwon Station? Why are you sleeping? I received a chaebol in Mapo, Seoul last month and I'm calling with a large number of security cards. I think it's Shinhan Bank under Maison M.O. I'm contacting you about the fact-finding certificate that includes Nonghyup Bank's account.	잠원역에서 불법 인터넷 도박사이트를 권유하는 만 35세 남자 김응술씨 한 명 말씀하시는 건가요? 주무세요 왜죠? 제가 지난달 서울 마포 지역에서 재벌을 받았는데 35살 남정림 김응수 씨 일당이 건강해서 현장에서 대포통장 보안카드 전화를 많이 했습니다. 메종올 명의의 신한은행인 것 같습니다. 농협은행 계좌가 포함되어 있어서 지금 사실관계 확인서 확인차 연락드렸습니다. Are you talking about Kim Yong-sul, a 35-year-old man who recommends illegal Internet gambling sites at Jamwon Station? Are you sleeping? Why? I received a chaebol in Mapo, Seoul last month, and the 35-year-old Nam Jung-rim and Kim Yong-soo's group were healthy, so I made a lot of calls with fake bank accounts on the spot. I think it's Shinhan Bank under Maison Ohm's name. I'm contacting you to check the facts now because the Nonghyup Bank account is included.	잠원역에서 불법 인터넷 도박 사이트를 추천한 만 35살 남정인 김영술 혼자입니까? 왜냐하면 자요? 제가 지난달 서울 마포 지역에서 재벌을 받아 만 35세 남정림, 김영수 일당을 잘 지내고 있었는데 현장에서 대량의 대포통장 보안카드를 전화했습니다. 메종엠오니 명의의 신한은행이라고 생각합니다. 농협은행 계좌가 포함되어 있어서 사실관계 확인서를 지금 연락드리겠습니다. Are you alone, Kim Young-sul, a 35-year-old male who recommended illegal Internet gambling sites at Jamwon Station? Because are you sleeping? I received a chaebol in Mapo, Seoul last month and was doing well with a 35-year-old male Jeong-rim and a Young-soo Kim, and I called a large number of fake bank accounts security cards at the site. I think it's Shinhan Bank under Maison Mioni's name. Nonghyup Bank's account is included, so I'll contact you now with the confirmation of the facts.
2754	개인 정보가 너무 니혼으로 괜찮은지 그 건에 대해서도 상세하게 조사를 해 드릴 겁니다. 일단은 선생님께서 정해서 저녁 4시 되면 무한 피해자가 증명하는게 가장 시급한 괜찮은 거 같고요. 선생님 계좌 계좌 김영철 진짜 그런 긍정적인 거래는 없었는지 있는 작은 일을 통해서 계산동 부분에 대해서 확인해 드릴 거고요. 여보세요 선생님 앞으로 세탁기 불법 자금 계획을 해 드릴 겁니다. 사장님께서 선생님이 전혀 모르는 다른 계좌 개설이 기울거나 그리고 대출 신청 선생님 앞으로 더 뒤에서 없어져 드릴 거고요. We'll also investigate in detail whether your personal information is too good with Nikon. First of all, I think it's most urgent for the unlimited victim to prove it at 4 p.m. Teacher's account Kim Young-cheol, we'll check the checkout area through small things to see if there was such a positive transaction. Hello, teacher, we'll plan illegal money for the washing machine in the future. The boss will either open another account that you don't know at all, or the loan application teacher will disappear from behind.	개인정보가 너무 니혼인지에 대해서도 자세히 조사하겠습니다. 일단 무한 피해자들이 오후 4시에 결정해서 입증하는 게 가장 시급하다고 생각합니다. 고객님, 김영철 고객님의 계좌에서 계산동 부분에 대해 이런 긍정적인 거래가 있었는지 작은 문제를 통해 확인해 보겠습니다. 안녕하세요. 나는 세탁기의 불법 자금 계획을 세울 것입니다. 주인은 당신이 전혀 모르는 다른 계좌를 개설하려는 경향이 있을 것이고, 대출을 신청하는 선생님은 뒤에서 사라질 것입니다. I will also investigate in detail whether the personal information is too Nikon. First of all, I think it is most urgent for unlimited victims to decide and prove it at 4 p.m. Dear Mr. Kim, I will check through a small matter if there was such a positive transaction for the checkout part in your account. Hello, I will make an illegal fund plan for the washing machine. The owner will tend to open another account that you have no idea about, and the teacher who applies for the loan will disappear behind the scenes.	니혼의 개인정보가 너무 좋은지 자세히 알아보겠습니다. 일단 선생님은 저녁 4시가 되면 무한 피해자 증명이 가장 절실하다고 결정했습니다. 선생님 계좌 김영철은 정말 이런 적극적인 거래가 없었나요, 작은 일로 결제동 부분을 확인해 보겠습니다. 여보세요, 선생님, 나중에 세탁기 불법 자금 계획을 알려드리겠습니다. 사장님은 전혀 모르시는 다른 계좌 개설 방식 때문에 기울어지거나 대출 신청 선생님이 나중에 뒤로 사라지실 거예요. Let me take a closer look at whether Nikon's personal information is too good. First of all, you decided that by 4 p.m., unlimited victim proof is the most urgent thing. Did Kim Young-chul really not have this active transaction, I will check the payment process for a small matter. Hello, sir, I will let you know the illegal money plan for the washing machine later. The boss will either lean because of the other way to open the account he does no know about, or the loan application teacher will disappear later.	개인정보가 니혼으로 좋은 것인지, 그 건에 대해서도 자세히 조사하겠습니다. 우선은 선생님이 결정해서 저녁 4시가 되면 무한 피해자가 증명하는 것이 가장 시급한 괜찮다고 생각합니다. 선생님 계좌의 김영철, 정말 그런 긍정적인 거래는 없었는지 작은 것을 통해 계산동 부분에 대해 확인하겠습니다. 여보세요, 선생님, 지금부터 세탁기 불법 자금 계획을 하겠습니다. 사장님이 선생님이 전혀 모르는 다른 계좌 개설이 기울어지거나 그리고 대출 신청 선생님은 앞으로 더 뒤에서 없어져 드릴게요. I will also look into the matter in detail to see if the personal information is good in Nikon. First of all, I think the most urgent thing is to decide by the teacher and prove it by the infinite victim at 4 p.m. Kim Young-chul in your account, I will check the part of the checkout to see if there really was such a positive deal. Hello, sir, I will start the illegal money plan for the washing machine from now on. If you open another account that you don't know at all is tilted, or the loan application teacher will disappear from behind.
2906	예상 하시는데요 예 저희 쪽으로 입출금이 가능한 계좌임대 주시면 제가 한 달에 250만 생각 좀 있으세요? 불법이야 불법은 뭐든지 다 불법 여사님이 하셔 담배 피워도 물건이구요. 하지만 두 번째로 집에 집에 하잖아요. 받는 거 통장도 안 받고 입출금이 가능한 제가 계좌 안 받고 있는데 하더라고 연락 주면 돼요 예. You're expecting it. Yes, if you could rent an account that allows us to deposit and withdraw money, do you have any thoughts on 2.5 million a month? It's illegal. Everything illegal is done by an illegal woman. Even if you smoke, it's a thing. But it's my second time at home. I don't have a bank account and I can deposit and withdraw money. I don't have an account, so you can contact me.	당신은 그것을 기대하고 있습니다. 네, 입출금이 가능한 계좌를 대여해주시면 한 달에 250만 건은 어떠세요? 불법입니다. 불법적인 모든 것은 불법적인 여성에 의해 행해집니다. 담배를 피운다 해도 그것은 중요한 것입니다. 하지만 저는 집에 두 번째입니다. 저는 은행 계좌가 없고 입출금이 가능합니다. 나는 계정이 없으니 당신은 나에게 연락할 수 있습니다. You are looking forward to it. Yes, how about 2.5 million cases a month if you lend me an account where I can deposit and withdraw money? It's illegal. Everything that is illegal is done by an illegal woman. Even if you smoke, it's important. But I'm second at home. I don't have a bank account and I can deposit and withdraw money. I don't have an account, so you can contact me.	예상하셨군요, 네, 입출금 가능한 계좌로 대여해주시면 한달에 250만원 생각하시나요? 불법이고 뭐고 다 불법인 여자가 하는 거고 담배도 물건이에요. 그런데 두 번째는 집에서 하잖아요. 받을 때 통장도 안 받고 입출금이 되는 저는 계좌번호 안 받으니까 연락 주시면 됩니다. You expected it. Yes, if you lend it to an account where you can deposit and withdraw money, do you think of 2.5 million won a month? It's illegal or anything, it's done by an illegal woman, and cigarettes are also an item. But the second one is done at home. I don't even receive a bank account when I receive it, and I don't receive an account number, so you can contact me.	예상되는데 네, 여기 입출금 가능한 계좌를 임대해주시면 저는 한 달에 250만 생각이 있습니까? 불법이야, 불법은 뭐든지 불법 부인이 하고 담배를 피워도 물건이에요. 그런데 두 번째로 집에서 하잖아요. 통장도 안받고 입출금이 가능한 제가 계좌를 못받았는데 해도 연락주세요 네. I'm expecting it. Yes, if you can rent this account, do I have thoughts of 2.5 million a month? It's illegal, anything illegal is illegal even if your wife does it and you smoke. But it's the second time you do it at home. Please contact me even if I didn't receive my account, which can be deposited and withdrawn without receiving my bank account. Yes.
2919	내가 지금부터 민감한 이야기할 거니까 사람 없는 곳으로 와요. 전화받아요.네 네 경고하는데 경찰에 신고하거나 주위 사람에게 알리면 애 배떼지 칼로 쑤셔버립니다. 내가 어디 있는지 xx 그걸 왜 나한테 물어보냐. 니 애 어딨는지도 모르냐.아줌마? 아줌마 내가 원하는 건 단순히 돈이야 돈. 난 돈이 필요하다.나한테 돈 얼마나 해줄 수 있냐요. 최대한? 아이가 어딨냐고? 우리 집 밑 이에요.아니 어딘지 왜 자꾸 위치 묻냐고 아 돈 준비하라니까. I'm going to talk about sensitive things from now on, so come to a place where there are no people. Answer the phone. Yes, yes. I warn you, but if you call the police or notify people around you, you'll stab them with a knife. Where am I xx why are you asking me that. I don't even know where your kid is.Ma'am? Ma'am, what I want is simply money. I need money.How much money can you give me. As much as possible? Where's the kid? It's under my house.Why do you keep asking where I am? I told you to prepare money.	지금부터 민감한 이야기를 할 테니까 사람이 없는 곳으로 오세요. 전화를 받아요. 네네. 경고합니다만, 경찰에 신고하거나 주변 사람들에게 알리면 칼로 찌르실 겁니다. 내가 어디 xx인지 왜 나한테 물어봐요. 당신 아이가 어딨는지도 몰라요.부인? 부인, 제가 원하는 것은 그저 돈입니다. 돈이 필요해요.당신은 나에게 얼마나 많은 돈을 줄 수 있습니까. 가능한 한 많이? 애는 어디 있어요? 우리 집 밑에 있어요.왜 자꾸 내가 어디냐고 물어봐요? 나는 당신에게 돈을 준비하라고 말했습니다. We will talk about sensitive things from now on, so come where there are no people. Answer the phone. Yes. I warn you, but if you call the police or notify others, they will stab you. Why do you ask me where I am xx? I don't even know where your child is. Ma'am? Ma'am, all I want is money. I need money. How much money can you give me? As much as possible? Where is the child? He is under my house. Why do you keep asking me where I am? I told you to prepare the money.	지금부터 민감한 이야기를 할테니 아무도 없는 곳으로 오세요. 전화 받으세요. 네, 경고합니다. 경찰에 신고하거나 주변에 알리면 아랫배 칼로 찌를 것입니다. 내가 어디에 있는지 xx 그걸 왜 나한테 물어봐요. 당신의 아이가 어디에 있는지도 모른다다.아줌마? 아줌마 내가 원하는 건 돈뿐이에요. 돈이 필요해요. 얼마 줄 수 있어요. 최대한? 아이는 어디 있어요? 우리 집 아래에요. 아니 왜 자꾸 위치를 물어봐요. 돈 준비하라고 했잖아요. We're going to talk about sensitive issues from now on, so come where nobody is. Pick up the phone. Yes, I warn you. Call the police or let others know, and they'll stab you in the bottom of your stomach. Why are you asking me where I am? I don't even know where your child is. I don't know where he is. Grandma, all I want is money. I need money. How much can I give you? As much as possible? Where is the child? Downstairs. Why do you keep asking me about the location? I told you to prepare the money.	내가 지금부터 민감한 이야기를 할 테니까 사람 없는 곳으로 오세요. 전화를 받습니다. 네, 경고하는데 경찰에 신고하거나 주변 사람들에게 알리면 아기 떼의 칼로 웅크리고 맙니다. 제가 어디에 있는지 xx 그걸 왜 저한테 물어보시는 거죠? 네 아이가 어디 있는지도 몰라요?아줌마? 아줌마, 내가 원하는 건 단순히 돈이에요. 나는 돈이 필요해요.저에게 얼마나 돈을 지불해 주실 수 있나요?최대한? 아이가 어디 있냐고요? 저희 집 아래입니다.아니, 어딘지 왜 자꾸 위치를 물어봐요.아, 돈 준비하래요. I'm going to talk about sensitive issues from now on, so please come where there are no people. I answer the phone. Yes, I call the police or notify people around me when I warn them, and they will crouch down with a knife. Why are you asking me where I am? You don't even know where your child is? Grandma? All I want is money. I need money. How much can you pay me? As much as possible? Where is my child? He's under my house. Why do you keep asking me where I am? Oh, get the money ready.

fully investigated in this study. This gap may constrain the model's robustness when deployed in real-world situations.

Addressing these limitations will enhance the proposed approach's scalability, applicability, and practicality.

VII. CONCLUSION AND FUTURE WORK

This study explored the impact of data augmentation methods on different ML and DL models for detecting Korean voice phishing on imbalanced datasets. The proposed augmentation method is a multilingual back-translation method to address the imbalanced workload scenario problem in the KorCCVi dataset. The analysis examined the impact of imbalanced, augmented, and balanced datasets on model performance. The empirical results indicate that ML models, such as DT and XGB, perform poorly on the imbalanced dataset but perform better on the balanced and augmented datasets. Despite the imbalance in the dataset, both the ML and DL models performed comparatively well compared with those on the balanced datasets. Moreover, compared with the overfitting issue observed with the CNN algorithm on the imbalanced and augmented training sets, the traditional SMOTE method resulted in more stable DL models that generalized well on the test set. This observation suggests that balancing the dataset with SMOTE reduces the overfitting probability during the model training. In addition, the overfitting issue highlights the importance of selecting the appropriate intermediate language when using the BT augmentation method. Regardless, the average performance revealed that the models trained using the proposed multilingual BT outperformed those trained on imbalanced and

balanced data. This is because of the preservation of the contextual and linguistic structures of the original dataset and the generation of diverse new samples. Nevertheless, while BT and SMOTE methods can improve model performance in imbalanced workloads, they may have trade-offs in terms of computational resources.

There are several avenues for future research. Additional languages can be explored to fine-tune the proposed multilingual BT process, expand the imbalanced dataset, and evaluate the performance of this method. Another avenue is to explore the impact of other DA methods, such as Generative Adversarial Networks (GAN), to address the imbalance problem of the dataset and analyze its performance in vishing detection. The GAN method can also be combined with the multilingual BT method to improve dataset size.

APPENDIX
DATASET BEFORE AND AFTER THE BACK-TRANSLATION PROCESS

This appendix presents a comparison of the original Korean transcript with its back-translated versions through intermediate languages (English, Chinese, and Japanese). Table 14 highlights how back-translation impacts semantic and structural consistency, demonstrating its role in the proposed data augmentation approach for improving voice phishing detection models.

REFERENCES

[1] APWG. (2024). *Phishing Activity Trends Reports, 1st Quarter 2024*. [Online]. Available: https://apwg.org/trendsreports

[2] Proofpoint. (2021). *2024 State of the Phish*. [Online]. Available: https://www.proofpoint.com/sites/default/files/threat-reports/pfpt-us-tr-state-of-the-phish-2024.pdf

[3] Transnexus.com. (2024). *Telecom Fraud*. [Online]. Available: https://transnexus.com/whitepapers/introduction-to-telecom-fraud

[4] Phishing. (2024). *Phishing What is Phishing?*. [Online]. Available: http://www.phishing.org/what-is-phishing

[5] Digital-Journal. (2019). *Businesses Now Face a 'Vishing' Threats (Includes Interview)*. [Online]. Available: https://www.digitaljournal.com/tech-science/businesses-are-coping-with-phising-but-how-about-vishing-threat/article/564188

[6] Z. Khanjani, G. Watson, and V. P. Janeja, "How deep are the fakes? Focusing on audio deepfake: A survey," 2021, *arXiv:2111.14203*.

[7] N. Jiang, Y. Jin, A. Skudlark, W.-L. Hsu, G. Jacobson, S. Prakasam, and Z.-L. Zhang, "Isolating and analyzing fraud activities in a large cellular network via voice call graph analysis," in *Proc. 10th Int. Conf. Mobile Syst., Appl., Services*, New York, NY, USA, Jun. 2012, pp. 253–266. [Online]. Available: https://dl.acm.org/doi/10.1145/2307636.2307660

[8] J. Kim, J. Kim, S. Wi, Y. Kim, and S. Son, "HearMeOut: Detecting voice phishing activities in Android," in *Proc. 20th Annu. Int. Conf. Mobile Syst., Appl. Services*, New York, NY, USA, Jun. 2022, pp. 422–435. [Online]. Available: https://dl.acm.org/doi/10.1145/3498361.3538939

[9] A. Derakhshan, I. G. Harris, and M. Behzadi, "Detecting telephone-based social engineering attacks using scam signatures," in *Proc. ACM Workshop Secur. Privacy Anal.*, New York, NY, USA, Apr. 2021, pp. 67–73. [Online]. Available: https://dl.acm.org/doi/10.1145/3445970.3451152

[10] J.-W. Kim, G.-W. Hong, and H. Chang, "Voice recognition and document classification-based data analysis for voice phishing detection," *Hum.-Centric Comput. Inf. Sci.*, vol. 11, Mar. 2021, Art. no. 2. [Online]. Available: http://hcisj.com/articles/?HCIS202111002

[11] M. Lee and E. Park, "Real-time Korean voice phishing detection based on machine learning approaches," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 7, pp. 8173–8184, Jul. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s12652-021-03587-x

[12] 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, Jul. 2023. [Online]. Available: https://www.mdpi.com/2227-7390/11/14/3217

[13] S. Yu, Y. Kwon, M. Kim, and K. Lee, "Korean voice phishing detection applying NER with key tags and sentence-level N-Gram," *IEEE Access*, vol. 12, pp. 52951–52962, 2024. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10496052

[14] D. R. Beddiar, M. S. Jahan, and M. Oussalah, "Data expansion using back translation and paraphrasing for hate speech detection," 2021, *arXiv:2106.04681*.

[15] (2024). *Voice Phishing Protector Voice Phishing Experience Center*. [Online]. Available: https://www.fss.or.kr/fss/bbs/B0000203/list.do?menuNo=200686

[16] Nat. Inst. Korean Lang. (2024). *Everyone's Corpus*. [Online]. Available: https://corpus.korean.go.kr/

[17] Nat. Police Agency Voice Phishing Statist. (2024). *Public Data Portal—The Official E-government Website of the Republic of Korea*. [Online]. Available: https://www.data.go.kr/data/15063815/fileData.do

[18] G. Bottazzi, E. Casalicchio, D. Cingolani, F. Marturana, and M. Piu, "MP-shield: A framework for phishing detection in mobile devices," in *Proc. IEEE Int. Conf. Comput. Inf. Technol.; Ubiquitous Comput. Commun.; Dependable, Autonomic Secure Comput.; Pervasive Intell. Comput.*, Oct. 2015, pp. 1977–1983. [Online]. Available: https://ieeexplore.ieee.org/document/7363339

[19] H. Li, X. Xu, C. Liu, T. Ren, K. Wu, X. Cao, W. Zhang, Y. Yu, and D. Song, "A machine learning approach to prevent malicious calls over telephony networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2018, pp. 53–69. [Online]. Available: https://ieeexplore.ieee.org/document/8418596

[20] C. M. R. da Silva, E. L. Feitosa, and V. C. Garcia, "Heuristic-based strategy for phishing prediction: A survey of URL-based approach," *Comput. Secur.*, vol. 88, Jan. 2020, Art. no. 101613. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404819301622

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002. [Online]. Available: https://www.jair.org/index.php/jair/article/view/10302

[22] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell.*, Jun. 2008, pp. 1322–1328. [Online]. Available: https://ieeexplore.ieee.org/document/4633969

[23] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, "Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model," *IEEE Access*, vol. 9, pp. 78621–78634, 2021. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9440474

[24] M. Kim and K.-B. Hwang, "An empirical evaluation of sampling methods for the classification of imbalanced data," *PLoS ONE*, vol. 17, no. 7, Jul. 2022, Art. no. e0271260, doi: 10.1371/journal.pone.0271260.

[25] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Appl. Soft Comput.*, vol. 76, pp. 380–389, Mar. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494618307130

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 60, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., May 2017, pp. 84–90. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[27] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," 2021, *arXiv:2105.03075*.

[28] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: A survey," *AI Open*, vol. 3, pp. 71–90, Mar. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651022000080

[29] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A survey on data augmentation for text classification," *ACM Comput. Surv.*, vol. 55, no. 7, pp. 1–39, Dec. 2022, doi: 10.1145/3544558.

[30] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *J. Big Data*, vol. 8, no. 1, p. 101, Dec. 2021, doi: 10.1186/s40537-021-00492-0.

[31] D. T. Vu, G. Yu, C. Lee, and J. Kim, "Text data augmentation for the Korean language," *Appl. Sci.*, vol. 12, no. 7, p. 3425, Mar. 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/7/3425

[32] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Nov. 2019, pp. 6382–6388. [Online]. Available: https://aclanthology.org/D19-1670/

[33] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, K. Erk and N. A. Smith, Eds., 2016, pp. 86–96. [Online]. Available: https://aclanthology.org/P16-1009/

[34] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Nov. 2018, pp. 489–500. [Online]. Available: https://aclanthology.org/D18-1045/

[35] J. Ma and L. Li, "Data augmentation for Chinese text classification using back-translation," *J. Phys., Conf. Ser.*, vol. 1651, no. 1, Nov. 2020, Art. no. 012039, doi: 10.1088/1742-6596/1651/1/012039.

[36] V. Ganganwar and R. Rajalakshmi, "MTDOT: A multilingual translation-based data augmentation technique for offensive content identification in Tamil text data," *Electronics*, vol. 11, no. 21, p. 3574, Nov. 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/21/3574

[37] D. R. Beddiar, M. S. Jahan, and M. Oussalah, "Data expansion using back translation and paraphrasing for hate speech detection," *Online Social Netw. Media*, vol. 24, Feb. 2021, Art. no. 100153, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468696421000355

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[39] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding," in *Proc. 4th Int. Conf. Learn. Represent.*, San Juan, Puerto Rico, Y. Bengio and Y. LeCun, Eds., 2016, pp. 1–15.

[40] S. Alyamkin et al., "Low-power computer vision: Status, challenges, opportunities," 2019, *arXiv:1904.07714*.

[41] R. Ding, Z. Liu, T.-W. Chin, D. Marculescu, R. D., and Blanton, "FLightNNs: Lightweight quantized deep neural networks for fast and accurate inference," 2019, *arXiv:1904.02835*.

[42] A. Goel, Z. Liu, and R. D. Blanton, "CompactNet: High accuracy deep neural network optimized for on-chip implementation," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 4723–4729. [Online]. Available: https://ieeexplore.ieee.org/document/8622329

[43] KoNLPy. (2024). *Python Package for Korean Natural Language Processing*. [Online]. Available: https://github.com/konlpy/konlpy

[44] T. Kudo. (2006). *MeCab: Yet Another Part-of-Speech and Morphological Analyzer*. [Online]. Available: http://taku910.github.io/mecab/

[45] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 1–12.

[46] Naver Corp. (2024). *Naver Papago: AI Translation Service*. [Online]. Available: https://papago.naver.com/

[47] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014, *arXiv:1409.3215*.

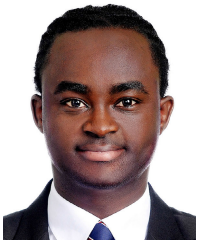
[48] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*, D.-S. Huang, X.-P. Zhang, and G.-B. Huang, Eds., Berlin, Germany: Springer, 2005, pp. 878–887, doi: 10.1007/11538059_91.

[49] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation," in *Proc. 11th Int. Workshop Semantic Eval. (SemEval)*, 2017, pp. 1–14. [Online]. Available: https://aclanthology.org/S17-2001/

[50] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2019, pp. 1–12.

[51] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1–14.

[52] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.



MILANDU KEITH MOUSSAVOU BOUSSOUGOU received the B.S. degree in computer science, specializing in network, database, and web, from the Institut Supérieur de Technologie (IST), Libreville, Gabon, in 2014, and the M.S. degree in computer science and engineering from Soongsil University, Seoul, South Korea, in February 2021, where he is currently pursuing the Ph.D. degree in computer science and engineering.

In 2017, he was granted a scholarship from the Global Korea Scholarship (GKS) program to pursue his studies in South Korea. From 2020 to 2021, he was a Developer with WeCrest Ltd., Seoul, and as a Network Engineer with Huawei Technologies Gabon SARL, from 2016 to 2017, and CFAO Technologies, from 2015 to 2016. His research interests include the application of AI and NLP to build solutions for security and cybersecurity.

Mr. Moussavou Boussougou has received several prestigious awards and honors, including the Merit Scholarship from the Korean Government, the Ph.D. Merit Scholarship from Soongsil University, and the Big Data and Artificial Intelligence Excellence Scholarship from the Daewoong Foundation.

PRINCE HAMANDAWANA received the B.S.C. degree (Hons.) in computer science from the National University of Science and Technology (NUST), Bulawayo, Zimbabwe, in 2010, and the Ph.D. degree in artificial intelligence from Ajou University, Suwon, South Korea, in 2020.

He was a Network Engineer with Econet Wireless, from 2008 to 2011, and Liquid Telecom, from 2011 to 2016. He is currently an Assistant Professor with the Department of Software and Computer Engineering, Ajou University. His research interests include operating systems, file and storage systems, smart storage utilization, and distributed systems for AI and ML.

DONG-JOO PARK received the B.S. and M.S. degrees from the Computer Engineering Department, Seoul National University, in February 1995 and February 1997, respectively, and the Ph.D. degree from the School of Computer Science and Engineering, Seoul National University, in August 2001.

He is currently a Professor with the School of Computer Science and Engineering, Soongsil University, Seoul, South Korea. His research interests include flash memory-based DBMSs, multimedia databases, and database systems.