Hybrid CNN-GRU Framework with Integrated Pre-trained Language Transformer for SMS Phishing Detection

Rubaiath E Ulfath
Department of Computer Science and Engineering,
Chittagong University of Engineering & Technology
Chittagong-4349, Bangladesh

Hamed Alqahtani
College of Computer Science
King Khalid University
Abha-62529, Saudi Arabia

Mohammad Hammoudeh
Department of Computing & Math, Manchester
Metropolitan University,
Manchester, UK

Iqbal H. Sarker
Department of Computer Science and Engineering,
Chittagong University of Engineering & Technology
Chittagong-4349, Bangladesh

ABSTRACT

Smartphones are prone to SMS phishing due to the rapid growth in the availability of smart mobile technologies driven by Internet connections. Also, detecting *phishing SMS* is a challenging task due to the unstructured nature of SMS text data with non-linear complex correlations. In this concern, considering the recent advancements in the domain of cybersecurity, we have proposed a *hybrid deep learning framework* that extracts robust features from SMS texts followed by an automatic detection of Phishing SMS. Due to combining the potential capability of individual models into one hybrid framework, it has outperformed various other individual machine learning and deep learning models. The proposed Phishing Detection framework is an effective hybrid combination of pretrained transformer model, MPNet (Masked and Permuted Language Modeling), with supervised ConvNets (CNN) and Bi-directional Gated Recurrent Units (GRU). It is intended to successfully detect unstructured short phishing text messages that contain complex patterns.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Information extraction**; **Natural language processing**; **Artificial intelligence**.

KEYWORDS

Smishing, Deep learning, NLP, AI, Cybersecurity

ACM Reference Format:
Rubaiath E Ulfath, Hamed Alqahtani, Mohammad Hammoudeh, and Iqbal H. Sarker. 2021. Hybrid CNN-GRU Framework with Integrated Pre-trained Language Transformer for SMS Phishing Detection. In *The 5th International Conference on Future Networks & Distributed Systems (ICFNDS 2021), December 15–16, 2021, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3508072.3508109

ICFNDS 2021, December 15–16, 2021, Dubai, United Arab Emirates

ACM ISBN 978-1-4503-8734-7/21/12…$15.00
https://doi.org/10.1145/3508072.3508109

1 INTRODUCTION

Smartphones have become one of the most important aspects of our daily lives. Almost everyone utilizes smartphone-like gadgets that enable internet connectivity, from the corporate world to the home. This provides phishers with a new opportunity to begin phishing via SMS. Fake text messages with links that resemble authentic and legitimate but are malicious in origin are delivered by SMS with the intent of stealing personal information, committing fraud, and spreading malicious smartphone viruses. Attackers now take different strategies to deceive smartphone users and trap them to steal their personal information by harming their privacy. Fake delivery notification, tax scams, fake two-factor authentication message [13] has become a very common way choice for smishing attackers.

SMS phishing, often known as smishing, is a type of phishing scam that targets smartphone users. Smishing attacks have become widespread in recent years [22]. Smartphones have been accessible to individuals of all ages and socioeconomic groups as a result of the rapid progress of smart and advanced technology in the previous decade [5].At the very beginning of the pandemic situation all over the world caused by COVID-19, reports showed that 44 percent of US citizens reported an increase in scam texting and calling within the first two weeks of the countrywide quarantine period [19]. As shown in a 2018 survey by the cloud-infrastructure company Wandera, 17% of its corporate users were exposed to phishing URLs on their smartphones. In contrast, just 15% of consumers got a scam email, and 16% received phishing URLs via social networking applications [4]. According to Proofpoint, a software security company, SMS-based scams have increased by 328% since the middle of 2020. It's because ordinary smartphone users with little awareness of the internet and phishing schemes are easily duped by these SMS-based phishing attacks. And attackers are coming up with new ways to fool users and lure them into a smishing attack to steal sensitive personal information or infect their mobile devices with malware. In 2020, the Bank of Ireland was obliged to pay €800,000 to over 300 bank clients as a result of a single smishing scheme [8]. SMS has recently become the most widely used data service on the globe. SMS is vital for corporate communications because the world sent 8.3 trillion [3] SMS messages in 2017, and 690 billion SMS messages are delivered weekly. According to Tatango's report [1], SMS Spam affects 68 percent of mobile phone users.

 doi:10.20944/preprints202202.0024.v1

Artificial intelligence researchers have been researching the possibility of machine learning and deep learning-based methods combined with natural language processing to resist smishing attacks [16]. Many state-of-the-art studies [23] [30] have been undertaken with notable results that have aided in the identification of phishing using AI technologies. However, individual deep learning or machine learning algorithms many times proved inefficient for handling larger text corpus with greater variability in features whereas hybrid models that overcome the limitations of individual models tend to perform better. Deep learning models [24] usually contain a great number of parameters which are needed to be learned during the training process. In this process to avoid an overfitting scenario the models are needed to be trained on significantly larger datasets. Building large-scale labeled datasets, on the other hand, is difficult for most NLP activities due to the high costs of annotation, especially for syntax and semantically related tasks. Pre-trained models trained in large corpora can solve this issue by unsupervised training and feature extraction. Transformer-based pre-trained models have shown great potential for unsupervised feature extraction in the domain of Natural Language Processing in recent years.

SMS texts data are difficult to process and unstructured in this manner. It's tough to tell the difference between phishing and legitimate SMS because of the non-linearity involved in the processing and analyzing SMS text data. Computationally, extracting significant features from text data is similarly time-consuming. As a result, proper detection of phishing SMS is a challenging problem in the AI-Driven Security [25] area. Therefore, the research question addressed in this work is: "RQ: How can we handle non-linearity, variability in-text features and minimize computing complexity while discovering robust discriminating characteristics to combat ever-increasing phishing attacks?"

In this context, to answer this research question, we have developed an integrated deep hybrid framework and proposed a compact architecture that can take raw SMS text data and by making use of certain data preprocessing steps, unsupervised feature extraction with Transformer based pretrained model along with the supervised integration of Lightweight ConvNets (CNN) and Bidirectional Gated Recurrent Units (GRU) leading to phishing SMS detection with fully connected blocks. The key contributions of this study are, as follows:

- Integration of transformer based pretrained feature extraction method by incorporating MPNet which takes advantage of both masked language modeling and permuted language modeling to generate robust embedding of SMS texts to complement smishing detection.
- Proposed a hybrid deep learning framework combining MPNet, ConvNets(CNN), Bidirectional Gated Recurrent Unit (GRU) for the detection short phishing texts.
- Performance analysis of proposed hybrid framework's effectiveness against individual deep models and popular machine learning models by experimenting on benchmark datasets.

The remainder of this paper is organized as follows. Section 2 offers details about the related works. In section 3, the proposed methodology has been described. In section 4, we present a comprehensive performance evaluation of our proposed hybrid model against individual established deep learning and machine learning approaches. Finally, section 5 presents the main conclusions and outlines for future works.

2 LITERATURE REVIEW

In recent, traditional machine learning-based methods utilizing handcrafted text features have been incorporated by Artificial Intelligence researchers for detecting Phishing SMS adopting the domain of AI-Driven Cybersecurity [25]. The study of [27] investigated feature selections strategies based on statistical significance using multiple correlation techniques and feature selection improved the performance of both tree-based and linear classifiers. By incorporating the Naive Bayes probabilistic model, Mishra et al. [18] constructed an efficient model that decreases false positives in assessing SMS contents and URL characteristics. Boukari et al. [7] explored the possibility of a machine learning-based detection system for smishing attacks that sends out an early warning to consumers. It can also be used to carry out phishing and vishing scams. Utilizing machine learning methods, multiple researchers have suggested novel and state-of-the-art smishing classifiers [11, 28]. Sahar et al. [6] have developed a system that includes numerous machine learning (ML) based classifiers that are built to utilize three classification methods – Naive Bayes (NB), Support Vector Machine (SVM), and Naive Bayes Multinomial (NBM) – as well as five preprocessing and feature extraction methods.

Traditional feature extraction methods and traditional machine learning algorithms, on the other hand, are not fully capable of capturing distinguishing features from more complex and unstructured SMS Texts, corroborating the detection of Phishing SMS. Artificial Intelligence researchers recognize the importance of incorporating deep learning methods to best support SMS Phishing detection in this context. Followed by this, several Artificial Intelligence researchers have studied the potential of deep learning algorithms such as CNN, LSTM, and Transformer based models for AI-Driven Cybersecurity. In the study of Jain et al. [14] the Long Short Term Memory (LSTM), a variation of the Recursive Neural Network (RNN), is used for spam classification. Gomaa [12] have compared the outcomes of seven different deep neural network architectures and six traditional machine learning classifiers.

Because of their intricate and state-of-the-art dense architecture, deep learning-based models such as CNN (Convolutional Neural Network) can perform better in feature extraction for text analysis and text classification-based issues. Because of its efficient encoding and sequence learning, LSTM (Long Short Term Memory) based architectures can solve text classification challenges. In this context, For smishing detection in Arabic and English texts, Ghourabi et al. [10] suggested a hybrid CNN-LSTM architecture. The intricate structure of deep learning-based models makes interpreting the underlying differentiating features that improve classification performance difficult at times. Content-based SMS classification for the Turkish Language was performed utilizing machine learning and deep learning methods to filter out undesirable texts, in the study of Karasoy et al. [15].In [32], the authors proposed a discrete

hidden Markov model for detecting SMS spam, which is the first study to use word order information to detect spam SMS.

Language Transformer models have been playing a central role in text processing and text analysis in the realm of Natural Language Processing in recent times due to their massive potential for robust text embedding. An optimized Transformer based model for detecting SMS spam messages has been proposed by Xiaoxu et al. [17] and evaluated the proposed model on benchmarking datasets. Sergio et al. [21] look into whether language models that are sensitive to the semantics and context of words, such as Google's BERT, can be used to resist this adversarial attack. A lightweight deep learning model is a blessing for the SMS Phishing detection domain, as Smart Phone like devices has lower computational resources due to their size and usability. Wei et al. [31] have proposed a novel lightweight deep neural model for SMS spam detection which is Lightweight Gated Recurrent Unit (LGRU). Furthermore, the authors have used enhanced semantics retrieved from external knowledge (WordNet) to aid in the understanding of SMS text inputs for better classification.

However, a single model may not be able to capture all the dependencies and complex correlations of unstructured SMS Phishing texts. Hybrid Deep Learning models that overcome the limitations of individual models may help in this concern. The major goal of this research is to present a lightweight, fast, and efficient integrated hybrid framework that takes advantage of the vast potential of both the transformer pretrained model and the state-of-the-art deep learning models.

3 METHODOLOGY

In this study, we have proposed a hybrid deep learning model for solving the detection of SMS phishing text with high efficiency. A graphical representation of the hybrid deep learning model architecture is illustrated in figure 1.

For the training stage, we have taken Raw SMS text data as input followed by certain text preprocessing steps including stop words removal and lemmatization. Articles and pronouns are typically categorized as stop words. The technique of gathering together the inflected forms of a word so that they may be studied as a single item, designated by the word's lemma, or dictionary form, is known as lemmatization in linguistics. Both stop words removal and lemmatization are two very important preprocessing steps for natural language processing (NLP) applications. Furthermore, textual data embedding is an important step, which is a learned text representation in which words with similar meanings are represented similarly. The algorithm for the training process of our model is given in Algorithm 1.

The complete layer by layer architecture of proposed framework which is implemented utilizing Tensorflow library is presented as a summary including the defined hyper parameters and input-output shapes of each layer is presented in Table 1.

Pretrained language model [20] based text embedding has the capability of enhancing the performance of deep learning modules used for text classification problems. Considering the great necessity of a strong representation of text data into numerical one, we have incorporated pretrained language modeling for extracting robust text features using MPNet [26]. Pre-training on a large text corpus can aid with downstream tasks by learning universal language representations. It also improves model initialization, which in turn improves generalization performance and speeds convergence on the target task. Furthermore, it can be thought of as a form of regularization that prevents over-fitting on small datasets.

Algorithm 1 Algorithm Design of Proposed Hybrid Deep Learning Framework

Input: $|\mathbb{N}|$ represents number of training SMS samples, with input in $\{n_1, n_2, \ldots, n_{|\mathbb{N}|}\}$, where n_i represents individual SMS text with a label associated with it, indicating whether the SMS is phishing or legitimate.

```
for Each n_i in ℕ do
    Remove stopwords from corpus.
    Apply WordNet lemmatizer to lemmatize.
end for
for Each n_i in ℕ do
    Extract feature using Pretrained MPNet.
end for
for Each Epoch do
    Extract feature from tokenized ℕ using CNN.
    Extract feature from tokenized ℕ using Bi Directional GRU.
    Combine MPNet, CNN and GRU based features.
    Feed features to fully connected layers for SMS Text classification.
end for
```

Output: Detection of Phishing SMS

In natural language processing research, pretrained language models [20] has been a hot topic. These models, such as BERT [9], are fine-tuned on downstream tasks to enhance accuracy after being trained on large-scale language corpora with well-planned pretraining objectives. Masked language modeling (MLM), which is used in BERT [9], and permuted language modeling (PLM), which is used in XLNet[34], are two examples of pretraining targets . BERT proposes masked language modeling (MLM) [9], which masks some tokens with a masked symbol [M] at random and predicts the masked tokens given the remaining tokens. If we mask tokens w2 and w4 from a sequence w=(w_1, w_2, w_3, w_4, w_5), the masked sequence becomes (w_1, [M], w_3, [M], w_5). To forecast w_2 and w_4, MLM promotes the model to extract better representations [29]. In XLNet, permuted language modeling (PLM) [34] is presented, in which a sequence is randomly permuted and the tokens in the right part (forecasted part) are predicted in an auto-regressive manner. For example, given a sequence w=(w_1, w_2, w_3, w_4, w_5), PLM predicts w_2 and w_4 auto-regressively conditioned on (w_1, w_3, w_5, w_2, w_4) (w_1, w_3, w_5) [29]. The researchers of MPNet [26] have combined the benefits of MLM and PLM while avoiding their drawbacks to create MPNet [26], a more powerful pretrained model. In consideration of the robust embedding capacity, fast processing, and lightweight nature of the MPNet model, we elected this model as a good-fit embedding strategy for the SMS phishing domain.

Moreover, CNN based models can be incorporated into Natural Language Processing for extracting non-linear and robust contiguous text features [10]. Textual data, on the other hand, is always given in sequences, and the order in which it is presented is critical. Because the meaning of a sentence varies when the order of

 doi:10.20944/preprints202202.0024.v1

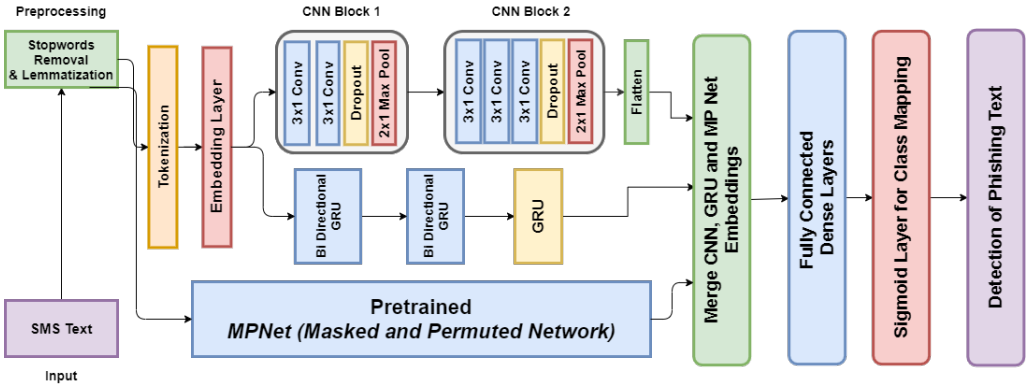

Figure 1: Diagram of Proposed Hybrid Deep Learning Framework

the words in that sentence changes, we therefore remark that the sentence information is stored in both the words and the order of the words in that sentence. Recurrent Neural Networks (RNN) are extremely popular for capturing these type of sequential features of text data to support text classification problems. LSTM (Long Short Term Memory) and GRU (Gated Recurrent Unit) are the two most popular and effective RNN architectures [24] in the text classification domain. However, GRU units are comparatively simple, lightweight and computationally more efficient and more faster than LSTM units [33]. Empirically, GRU has evinced better performance than LSTM for short text and smaller datasets. Yet, individual models can fail to capture sufficient number of eloquent complex text features from multidimensional aspects. To address this problem, the proposed hybrid deep learning framework includes a CNN and a GRU module with a pretrained language transformer in an integrated manner for optimizing the feature extraction approach for SMS Phishing Texts in motivation of individual models complementing one another when combined.

In the beginning, the raw SMS text data is incorporated into MPNet.Then pretrained MPNet model generates 768-dimensional vectors for each SMS text sample in the training dataset. After that, the words are tokenized and incorporated into CNN and bidirectional GRU units. The CNN unit of our hybrid framework consists of two CNN blocks. The two blocks of CNN is designed in a lightweight manner with a filter size of 3x1 for each Conv layer and the number of filters is 32 and 64 respectively for each Conv layer of block 1 and 2. In each of the Conv Blocks, a dropout layer of 0.5 thresholds is incorporated, following that the resultant feature vector obtained from CNN blocks is flattened for integration with fully connected layers. In the GRU units, two bi-directional GRU blocks with 256 nodes followed by one uni-directional GRU block containing 64 nodes are designed and integrated into our proposed hybrid framework. Then, the three types of features extracted from MPNet, GRU, and CNN-based models have been concatenated into a single feature vector for feeding it as an input for fully connected layers to classify phishing and legitimate SMS. In our proposed hybrid deep learning framework 3 fully connected layers have been incorporated with the number of nodes accordingly 1024, 1024, 512 with a learning rate of 0.001 (Adam Optimizer), and the loss function is binary cross-entropy loss.

The feature vectors are merged and given to the fully connected block 1. In the fully connected block 1 the feature vectors are given to the dense layer with 1024 nodes, followed by a batch normalization layer and the Leaky ReLU activation function is used for solving nonlinear problems with alpha value of 0.01. After that 50% dropout is applied. Then the output is given to the next fully connected block 2. It also follows the same steps. Then the output of the previous block is given to the next final fully connected block's dense layer along with 512 nodes followed by batch normalization. After that Leaky ReLU activation function is used similarly to the previous blocks. Then again 50% dropout has been applied. After completing all the steps of fully connected layers the output is given to the sigmoid layer for doing the final classification. After that, it will determine whether the SMS is phishing or legitimate.

4 EXPERIMENTAL RESULTS

In this section, we have evaluated the efficiency of our proposed hybrid framework against individual deep models and popular machine learning models by experimenting with several benchmark SMS Phishing Datasets.

4.1 Dataset Description

We have collected our dataset [2] from the UCI machine learning repository, which is a well-known source for datasets used in machine learning research. It comprises 5572 data instances, of which 4825 are "legitimate" (legitimate SMS) and 747 are "phishing" (Fake or phishing SMS). We have used stratified sampling to divide our dataset into training and testing sets for machine learning classifiers, with an 80:20 hold-out validation ratio. This dataset has been used for both training and testing our hybrid deep learning framework. To validate the efficiency of our proposed hybrid deep learning framework, we have experimented with another popular dataset in the domain of SMS Phishing named British English SMS Corpora Dataset which contains 875 SMS samples labeled Phishing SMS and Legitimate SMS. A Wordcloud analysis illustrating 100 top most frequent words of Phishing SMS from UCI SMS Dataset [2] is depicted in Figure 2.

Figure 2 illustrates that Phishing SMS usually comprises numbers, as well as alluring terms such as "free", "winner," "reward,"

 doi:10.20944/preprints202202.0024.v1

Table 1: The Complete Layer by layer summary of proposed Hybrid Architecture

Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	[(None, 300)]	0	
embedding (Embedding)	(None, 300, 300)	30000000	input_2[0][0]
conv1d (Conv1D)	(None, 300, 32)	28832	embedding[0][0]
batch_normalization (BatchNorma	(None, 300, 32)	128	conv1d[0][0]
leaky_re_lu (LeakyReLU)	(None, 300, 32)	0	batch_normalization[0][0]
conv1d_1 (Conv1D)	(None, 300, 32)	3104	leaky_re_lu[0][0]
batch_normalization_1 (BatchNor	(None, 300, 32)	128	conv1d_1[0][0]
leaky_re_lu_1 (LeakyReLU)	(None, 300, 32)	0	batch_normalization_1[0][0]
dropout (Dropout)	(None, 300, 32)	0	leaky_re_lu_1[0][0]
max_pooling1d (MaxPooling1D)	(None, 150, 32)	0	dropout[0][0]
conv1d_2 (Conv1D)	(None, 150, 64)	6208	max_pooling1d[0][0]
batch_normalization_2 (BatchNor	(None, 150, 64)	256	conv1d_2[0][0]
leaky_re_lu_2 (LeakyReLU)	(None, 150, 64)	0	batch_normalization_2[0][0]
conv1d_3 (Conv1D)	(None, 150, 64)	12352	leaky_re_lu_2[0][0]
batch_normalization_3 (BatchNor	(None, 150, 64)	256	conv1d_3[0][0]
leaky_re_lu_3 (LeakyReLU)	(None, 150, 64)	0	batch_normalization_3[0][0]
conv1d_4 (Conv1D)	(None, 150, 64)	12352	leaky_re_lu_3[0][0]
batch_normalization_4 (BatchNor	(None, 150, 64)	256	conv1d_4[0][0]
leaky_re_lu_4 (LeakyReLU)	(None, 150, 64)	0	batch_normalization_4[0][0]
embedding_1 (Embedding)	(None, 300, 300)	30000000	input_2[0][0]
dropout_1 (Dropout)	(None, 150, 64)	0	leaky_re_lu_4[0][0]
bidirectional (Bidirectional)	(None, 300, 128)	140544	embedding_1[0][0]
max_pooling1d_1 (MaxPooling1D)	(None, 75, 64)	0	dropout_1[0][0]
bidirectional_1 (Bidirectional)	(None, 300, 128)	74496	bidirectional[0][0]
input_1 (InputLayer)	[(None, 768)]	0	
flatten (Flatten)	(None, 4800)	0	max_pooling1d_1[0][0]
gru_2 (GRU)	(None, 64)	37248	bidirectional_1[0][0]
concatenate (Concatenate)	(None, 5632)	0	nput_1[0][0] flatten[0][0] gru_2[0][0]
dense (Dense)	(None, 1024)	5768192	concatenate[0][0]
batch_normalization_5 (BatchNor	(None, 1024)	4096	dense[0][0]
leaky_re_lu_5 (LeakyReLU)	(None, 1024)	0	batch_normalization_5[0][0]
dropout_2 (Dropout)	(None, 1024)	0	leaky_re_lu_5[0][0]
dense_1 (Dense)	(None, 1024)	1049600	dropout_2[0][0]
batch_normalization_6 (BatchNor	(None, 1024))	4096	dense_1[0][0]
leaky_re_lu_6 (LeakyReLU)	(None, 1024)	0	batch_normalization_6[0][0]
dropout_3 (Dropout)	(None, 1024)	0	leaky_re_lu_6[0][0]
dense_2 (Dense)	(None, 512)	524800	dropout_3[0][0]
batch_normalization_7 (BatchNor	(None, 512)	2048	dense_2[0][0]
leaky_re_lu_7 (LeakyReLU)	(None, 512)	0	batch_normalization_7[0][0]
dropout_4 (Dropout)	(None, 512)	0	leaky_re_lu_7[0][0]
dense_3 (Dense)	(None, 512)	262656	dropout_4[0][0]
batch_normalization_8 (BatchNor	(None, 512)	2048	dense_3[0][0]
leaky_re_lu_8 (LeakyReLU)	(None, 512)	0	batch_normalization_8[0][0]
dropout_5 (Dropout)	(None, 512)	0	leaky_re_lu_8[0][0]
dense_4 (Dense)	(None, 2)	1026	dropout_5[0][0]

"cash," "prize," and "call", as well as website links, contact numbers, and claim codes.

4.2 Evaluation Metrics

It is essential to analyze the performance of a machine learning model using well-defined evaluation metrics in order to justify its efficacy. The following metrics are used to evaluate our framework's performance:

Table 2: Comparison of evaluation matrix on UCI Spam Dataset

Metrics	CNN	GRU	MLP	SVM	Xgboost	Proposed Model
Precision (Phishing)	98.45%	95.78%	60.40%	65.00%	87.95%	97.16%
Recall (Phishing)	85.24%	91.28%	40.94%	34.90%	48.99%	91.95%
F1-Score (Phishing)	91.37%	93.47%	48.80%	45.42%	62.93%	94.48%
Precision (Legitimate)	97.77%	98.66%	91.32%	90.63%	92.64%	98.77%
Recall (Legitimate)	99.79%	99.38%	95.86%	97.10%	98.97%	99.59%
F1-Score (Legitimate)	98.77%	99.02%	93.54%	93.75%	95.70%	99.18%
Accuracy	97.85%	98.30%	88.52%	88.79%	92.29%	98.57%

Table 3: Performance Evaluation Scores on British Dataset

Metrics	CNN	GRU	MLP	SVM	XgBoost	Proposed Model
Precision (Phishing)	99.98%	99.98%	95.42%	95.77%	99.73%	99.98%
Recall (Phishing)	93.41%	95.77%	63.77%	58.59%	87.29%	96.47%
F1-Score (Phishing)	96.59%	97.84%	76.45%	72.70%	93.10%	98.20%
Precision (Legitimate)	94.14%	96.15%	73.94%	71.38%	89.26%	96.77%
Recall (Legitimate)	99.98%	99.98%	97.11%	97.56%	99.78%	99.98%
F1-Score (Legitimate)	96.98%	98.04%	83.96%	82.44%	94.23%	98.36%
Accuracy	96.80%	97.94%	80.91%	78.63%	93.71%	98.29%

Figure 2: Wordcloud of UCI Spam SMS Dataset

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Recall: Recall is delineated as the ratio of the number of positive samples that are being correctly predicted as legitimate corresponding to all legitimate SMS in the data. It can define as

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Precision: Precision is delineated as the ratio of the number of positive samples that are being correctly predicted as legitimate corresponding to all SMS predicted as legitimate. It can define as

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Here, True Positive, True Negative, False Positive, and False Negative are all abbreviated as TP, TN, FP, and FN. F1-Score: F1-Score is delineated as the term which balances between precision and recall. It can be defined as

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (4)$$

4.3 Experimental Analysis and Discussion

In the evaluation stage, we have evaluated the efficiency of our model through the Precision, Recall, and F1-Scores for both Phishing SMS and Legitimate SMS class including both UCI SMS Dataset and British English SMS Corpora. In Table 2, we have done a thorough investigation of classifiers' performance on UCI Spam Dataset by evaluating the Precision, Recall, and F1 scores of each class label, naming phishing, and Legitimate.

As the UCI Spam Dataset has a large number of legitimate SMS samples than phishing SMS samples, the individual CNN-based model is performing more biased towards the prediction of legitimate SMS. However, by including pretrained transformer-based unsupervised feature extraction into our proposed framework, we were able to successfully reduce the bias induced by class imbalance, as seen by the high recall value of the phishing SMS class. The robust feature extraction strategy of combining permuted, masked, contiguous, and sequential (local and global) textual features proposed in this study is overcoming the limitations of individual models and successfully avoiding overfitting issues towards one particular class despite having a great imbalance in class distribution. To validate the efficiency with more confidence, in Table 3, we have done a thorough investigation of classifiers' performance on the British English SMS Spam dataset by evaluating the Precision, Recall, and F1 scores of each class label, naming phishing and Legitimate.

Our proposed framework has evinced great generalization capacity due to multi-aspect feature extraction strategy by outperforming individual model in terms of British SMS corpora dataset also. Furthermore, the weighted average scores of F1 Score, Precision and Recall for individual classifiers and our proposed hybrid deep learning framework on UCI Spam dataset presented in Figure 3 justify the robustness of our proposed framework in an imbalanced class

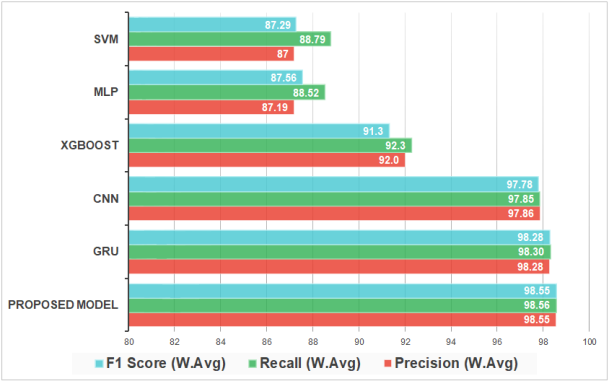

Figure 3: Comparison of Weighted Average Scores for UCI SPAM Dataset

condition perfectly. The scores of evaluation metrics clarify that our proposed hybrid CNN-GRU Framework with Integrated Pre-trained Language Transformer for SMS Phishing Detection outperforms all other classifiers with 98.57% accuracy for UCI spam dataset and 98.29% accuracy for British Dataset. From the performance analysis of the hybrid deep framework proposed in this study in comparison to other models, it can be stated that the integration of state-of-the-art transformer-based pretrained MPNet and Ensemble of lightweight Convolutional Neural Networks with Bi-directional GRU unit complements the performance of the proposed framework for effectively detecting Smishing attacks. The individual deep modules integrated into one hybrid deep framework overcome the limitations of each other to better corroborate the detection of phishing SMS by adopting the domain of AI-Driven cybersecurity.

5 CONCLUSION

Smishing messages are on the rise, and they now account for the majority of cyber-attacks in cyberspace. Even though most researchers are offering advanced methods to slow down the rate of these attacks, they have yet to do more. In this paper, we have offered a hybrid CNN-GRU Framework for Phishing SMS Detection with an Integrated Pre-trained Language Transformer. We have discovered that our suggested integrated framework outperforms all other machine learning and deep learning classifiers after a thorough investigation of the performance. The design of the hybrid framework has immense potential for adoption in large-scale real-world scenarios to defend against cyber-attacks for different forms. In the future, we would like to build GAN-based advanced models for predicting trends of SMS Phishing to defend large-scale attacks against variability ever-growing text-based phishing.

REFERENCES

[1] 2011. Text Message Spam Infographic. https://www.tatango.com/blog/text-message-spam-infographic/

[2] 2012. UCI Machine Learning Repository: SMS Spam Collection Data Set. https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

[3] 2017. Daily SMS Mobile Usage Statistics. https://www.smseagle.eu/2017/03/06/daily-sms-mobile-statistics/

[4] 2018. *Mobile Phishing Report 2018.* Technical Report. https://www.wandera.com/mobile-phishing-report/

[5] 2021. Mobile Phishing Increases More Than 300% as 2020 Chaos Continues | Proofpoint US. https://www.proofpoint.com/us/blog/threat-protection/mobile-phishing-increases-more-300-2020-chaos-continues

[6] Sahar Bosaeed, Iyad Katib, and Rashid Mehmood. 2020. A Fog-Augmented Machine Learning based SMS Spam Detection and Classification System. In *2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC).* 325–330. https://doi.org/10.1109/FMEC49853.2020.9144833

[7] Badr Eddine Boukari, Akshaya Ravi, and Mounira Msahli. 2021. Machine Learning Detection for SMiShing Frauds. In *2021 IEEE 18th Annual Consumer Communications Networking Conference (CCNC).* 1–2. https://doi.org/10.1109/CCNC49032.2021.9369640

[8] E. Burke-Kennedy, J. Brennan, and C. Taylor. 2020. Bank of Ireland does U-turn after refusal to reimburse 'smishing' victims. https://www.irishtimes.com/business/financial-services/bank-of-ireland-does-u-turn-after-refusal-to-reimburse-smishing-victims-1.4326502

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[10] Abdallah Ghourabi, Mahmood A. Mahmood, and Qusay M. Alzubi. 2020. A Hybrid CNN-LSTM Model for SMS Spam Detection in Arabic and English Messages. *Future Internet* 12, 9 (2020). https://doi.org/10.3390/fi12090156

[11] Diksha Goel and Ankit Kumar Jain. 2017. Smishing-classifier: a novel framework for detection of smishing attack in mobile environment. In *International conference on next generation computing technologies.* Springer, 502–512.

[12] Wael Hassan Gomaa. 2020. The Impact of Deep Learning Techniques on SMS Spam Filtering. *International Journal of Advanced Computer Science and Applications* 11, 1 (2020). https://doi.org/10.14569/IJACSA.2020.0110167

[13] Paul A Grassi, James L Fenton, Elaine M Newton, Ray A Perlner, Andrew R Regenscheid, William E Burr, Justin P Richer, Naomi B Lefkovitz, Jamie M Danker, Yee-Yin Choong, et al. 2020. Digital identity guidelines: Authentication and lifecycle management [includes updates as of 03-02-2020]. (2020).

[14] Gauri Jain, Manisha Sharma, and Basant Agarwal. 2019. Optimizing semantic LSTM for spam detection. *International Journal of Information Technology* 11, 2 (01 Jun 2019), 239–250. https://doi.org/10.1007/s41870-018-0157-5

[15] Onur Karasoy and Serkan Ballı. 2021. Spam SMS detection for Turkish language with deep text analysis and deep learning methods. https://link.springer.com/article/10.1007/s13369-021-06187-1

[16] Sumit Kumar, Arup Kumar Pal, SK Hafizul Islam, and Mohammad Hammoudeh. 2021. Secure and efficient image retrieval through invariant features selection in insecure cloud environments. *Neural Computing and Applications* (2021), 1–26.

[17] Xiaoxu Liu, Haoye Lu, and Amiya Nayak. 2021. A Spam Transformer Model for SMS Spam Detection. *IEEE Access* 9 (2021), 80253–80263. https://doi.org/10.1109/ACCESS.2021.3081479

[18] Sandhya Mishra and Devpriya Soni. 2020. Smishing Detector: A security model to detect smishing through SMS content analysis and URL behavior analysis. *Future Generation Computer Systems* 108 (2020), 803–815. https://doi.org/10.1016/j.future.2020.03.021

[19] Next Caller. 2020. *Next Caller's Fraud COVID-19 Report.* Technical Report (Week 2 3). https://nextcaller.com/blog/next-caller-covid-19-fraud-report/

[20] XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences* 63, 10 (01 Oct 2020), 1872–1897. https://doi.org/10.1007/s11431-020-1647-3

[21] Sergio Rojas-Galeano. 2021. Using BERT Encoding to Tackle the Mad-lib Attack in SMS Spam Detection. arXiv:2107.06400 [cs.CL]

[22] Jibran Saleem and Mohammad Hammoudeh. 2018. Defense methods against social engineering attacks. In *Computer and network security essentials.* Springer, 603–618.

[23] Iqbal H Sarker. 2021. CyberLearning: Effectiveness analysis of machine learning security modeling to detect cyber-anomalies and multi-attacks. *Internet of Things* 14 (2021), 100393.

[24] Iqbal H Sarker. 2021. Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective. *SN Computer Science* 2, 3 (2021), 1–16.

[25] Iqbal H Sarker, Md Hasan Furhad, and Raza Nowrozy. 2021. AI-driven cybersecurity: an overview, security intelligence modeling and research directions. *SN Computer Science* 2, 3 (2021), 1–18.

[26] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. arXiv:2004.09297 [cs.CL]

[27] Gunikhan Sonowal. 2020. Detecting Phishing SMS Based on Multiple Correlation Algorithms. *SN Computer Science* 1, 6 (2020), 1–9.

[28] Gunikhan Sonowal and K S Kuppusamy. 2018. SmiDCA: An Anti-Smishing Model with Machine Learning Approach. *Comput. J.* 61, 8 (04 2018), 1143–1157. https://doi.org/10.1093/comjnl/bxy039 arXiv:https://academic.oup.com/comjnl/article-pdf/61/8/1143/25209236/bxy039.pdf

[29] Xu Tan. 2020. MPNet combines strengths of masked and permuted language modeling for language understanding. https://www.microsoft.com/en-us/research/blog/mpnet-combines-strengths-of-masked-and-permuted-

language-modeling-for-language-understanding/

[30] Rubaiath E. Ulfath, Iqbal H. Sarker, Mohammad Jabed Morshed Chowdhury, and Mohammad Hammoudeh. 2022. Detecting Smishing Attacks Using Feature Extraction and Classification Techniques. In *Proceedings of the International Conference on Big Data, IoT, and Machine Learning*. Springer Singapore, Singapore, 677–689.

[31] Feng Wei and Trang Nguyen. 2020. A Lightweight Deep Neural Model for SMS Spam Detection. In *2020 International Symposium on Networks, Computers and Communications (ISNCC)*. 1–6. https://doi.org/10.1109/ISNCC49221.2020.9297350

[32] Tian Xia and Xuemin Chen. 2020. A Discrete Hidden Markov Model for SMS Spam Detection. *Applied Sciences* 10, 14 (2020). https://doi.org/10.3390/app10145011

[33] Shudong Yang, Xueying Yu, and Ying Zhou. 2020. LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example. In *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*. 98–101. https://doi.org/10.1109/IWECAI50956.2020.00027

[34] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf