SINTECHCOM: Science, Technology and Communication Journal
Vol. 6, No. 2, February 2026, pp. 155-164, DOI: 10.59190/stc.v6i2.360

Enhanced social media phishing detection model using LSTM and BERT

Wenni Syafitri[1,*], Eddisyah Putra Pane[2], Edi Purwanto[2]
[1]Department of Informatic Engineering, Universitas Lancang Kuning, Pekanbaru 28266, Indonesia
[2]Department of Information System, Universitas Lancang Kuning, Pekanbaru 28266, Indonesia

ABSTRACT

Phishing attacks are a major cyber threat, with more than 30% of incidents occurring via social media platforms, especially short message services. This study evaluates deep learning approaches for automated phishing detection using BERT and Hybrid (BERT-LSTM) architectures fine-tuned on 15950 annotated SMS. The BERT-only model achieved superior performance (F1 0.9928, recall 0.9952, AUC 0.999) with no statistically significant improvement from adding BiLSTM layers (0.0006). K-fold cross-validation demonstrated robust generalisation (coefficient of variation 0.10%). Dataset saturation analysis indicated that 15,950 SMS are sufficient for effective transfer learning. Mild overfitting (6.3x loss ratio) remained within acceptable bounds and did not affect validation metrics. The 1.77% false positive rate and 99.52% recall enable practical deployment for production phishing defence. Results demonstrate that transfer learning with BERT achieves production-grade performance while challenging conventional assumptions about architectural complexity.Δ

ARTICLE INFO

Article history:

Received Feb 2, 2026
Revised Feb 22, 2026
Accepted Feb 24, 2026

Keywords:

BERT
Cybersecurity
Deep Learning
Phishing Detection
Transfer Learning

* **Corresponding Author**
E-mail address: wenni20@gmail.com

1. INTRODUCTION

Digital transformation and social media have introduced increasingly sophisticated cybersecurity challenges, particularly evolving phishing attacks. Globally, phishing threats have reached unprecedented levels, with over 3.4 billion phishing emails sent daily, accounting for 1.2% of total global email traffic. This trend continues to escalate, with the Anti-Phishing Working Group (APWG) reporting 1,003,924 phishing incidents in the first quarter of 2025, the highest figure since late 2023.

The situation in Indonesia reflects similar urgency. In 2024, SAFEnet reported 330 cybersecurity incidents, with Instagram as the most vulnerable platform (107 incidents) and WhatsApp second (84 incidents). Although phishing cases declined from 108 (2023) to 25 (2024), the evolving attack patterns reveal concerning trends: account compromises increased from 62 to 86 incidents, and inaccessible accounts surged from 20 to 52 incidents.

Social media platforms represent strategic targets for three primary reasons: users' high trust in content from their social networks, significant incident prevalence 30.5% of total phishing occurs on social media, with over 60% of account compromises involving login credential schemes and substantial financial impact, averaging $4.88 million in annual losses per incident. In the financial sector, 64% of institutions reported Business Email Compromise (BEC) attacks in 2024, with average losses of $150,000 per occurrence.

Phishing detection has become a focal research area in recent years, with deep learning methods demonstrating promising results. Jonker et al. (2021) evaluated diverse Natural Language Processing (NLP) and Machine Learning solutions for phishing detection, including Word2Vec, Doc2Vec, BERT, and RNN, LSTM, CNN, and TF-IDF models, with findings indicating that each

approach achieved strong classification performance, with F1-scores ranging from 90.03% to 98.94% [1]. Pimpason et al. (2025) implemented a comprehensive deep learning strategy integrating LSTM, GRU, BERT, CNN, and RNN for phishing email detection, wherein the LSTM model achieved the highest accuracy of 99.92% [2]. Conversely, Sotomayor et al. (2025) proposed a Deep Learning model specifically designed for social media phishing detection, employing CNN, LSTM, and GRU architectures, along with contextual embeddings such as Word2Vec and BERT. The CNN model demonstrated the highest effectiveness, achieving 93.42% accuracy and 93.39% F1-Score [3].

A comparative study by Rao et al. (2024) evaluated the integration of conventional sequential models (RNN, LSTM, GRU) with state-of-the-art language representation models (BERT and XLNet) for phishing website detection. The findings indicated that hybrid architectures, particularly BERT+LSTM and XLNet+LSTM, significantly improved detection accuracy rates to 98.00% and 98.50%, respectively [4]. Atawneh and Aljehani (2023) corroborated these results, demonstrating that the hybrid BERT-LSTM model achieved 99.61% accuracy in phishing email identification, outperforming standalone CNN, RNN, and LSTM models [5]. Pan et al. (2025) combined BERT for feature extraction, Atrous Spatial Pyramid Pooling (ASPP) for multi-scale feature enhancement, and a CNN-based classification network, achieving 97.81% accuracy, 98% precision, and 0.9972 AUC in detecting spear-phishing attacks and digital fraud in the e-commerce sector [6].

Hyperparameter optimisation approaches have also demonstrated promising outcomes. Gupta et al. (2024) demonstrated that Teaching Learning-Based Optimisation (TLO) for hyperparameter tuning, combined with BERT for feature extraction, achieved 99.9% accuracy with only three false negatives and zero false positives in blockchain transaction phishing detection [7]. Gaurav et al. (2025) demonstrated the effectiveness of the Hill Climbing algorithm for hyperparameter optimisation, attaining 95% accuracy with balanced precision, recall, and F1-scores surpassing GRU, LSTM, RNN, Logistic Regression, and SVM [8].

Additional studies have investigated diverse aspects of phishing detection. Abiramasundari and Ramaswamy (2025) developed a Language Pack-based Tuned Transformer Language (LPTTL) framework for detecting ransomware-based email phishing through cacographic analysis (spelling errors), achieving 95.47% accuracy using the T5-HOAGRU model [9]. Murhej and Nallasivan (2025) proposed a multimodal framework leveraging EM-BERT and EAI-SC-LSTM based on SPCA for detecting phishing from multiple data sources (SMS, email, URL), attaining 99.627% accuracy for SSC, 99.645% for PEC, and 99.541% for the WPD dataset [10]. Liew and Law (2022) developed the BERT Embedding Attention Model (BEAM) with sub-word tokenisation for phishing link detection, achieving optimal test accuracy on balanced and larger datasets [11]. Jishnu and Arthi (2023) Integrated RoBERTa for feature extraction and LSTM for classification in phishing URL detection, obtaining 97.14% accuracy on a 300,000 URL dataset [12]. Manjula et al. (2024) proposed the PD-UHD feature set, which extracts detailed semantic features from raw URLs, HTML content, and domain names using a CNN-LSTM-BERT combination, achieving 98.10% accuracy and 98.34% AUC [13].

Despite substantial progress, comprehensive literature analysis reveals several significant research gaps. First, only two of the 13 evaluated studies (15.4%) specifically addressed phishing detection on social media platforms, namely Sotomayor et al. (2025) and Jishnu and Arthi (2023) [3, 12]. Most research remains focused on email and website phishing detection, whereas phishing attacks on social media exhibit fundamental differences in content structure, user behavioural patterns, and dissemination mechanisms. Second, although all reviewed studies employed LSTM and BERT, the optimal combination of these methods, particularly within the social media context, has not been thoroughly explored. LSTM is recognised for reliably capturing long-term dependencies in sequential data, while BERT excels in understanding bidirectional context and semantic representation. However, their synergistic application in identifying complex phishing patterns on social media requires further investigation.

Third, several technical limitations identified in prior research include: (a) the need for larger and more diverse datasets to enhance detection accuracy [5], (b) model limitations in addressing zero-day attacks and evolving adversarial techniques [6], (c) insufficient analysis of social media-specific features such as slang, cacography (intentional spelling errors), and SMS language prevalent in phishing attacks [9], and (d) challenges in real-time implementation within integrated cybersecurity systems [3]. Fourth, existing studies have not comprehensively explored hyperparameter optimisation

methods to enhance the performance of the hybrid BERT-LSTM model. In contrast, research by Gupta et al. (2024) [7] and Gaurav et al. (2025) [8] demonstrated that optimisation can significantly improve accuracy.

Based on the aforementioned research gaps, this study addresses a critical question: How can an accurate and adaptive phishing detection model for social media platforms be developed by leveraging the complementary strengths of the LSTM architecture for sequential modelling and BERT for contextual understanding? The objectives of this research are to design and implement an enhanced social media phishing detection model by integrating LSTM to capture temporal patterns and sequential dependencies within message content, and BERT to extract deep semantic and contextual representations, and to evaluate performance across metrics, including precision, recall, F1-score, AUC-ROC, and accuracy.

2. RESEARCH METHODOLOGY

This study proposes a deep learning–based approach for detecting phishing attempts in Indonesian-language SMS messages by leveraging transformer and sequential neural network architectures (Figure 1). The research methodology comprises data preprocessing, model development, experimental configuration, and performance evaluation across multiple validation schemes.

The dataset used in this study comprises Indonesian SMS messages labelled into two categories: phishing and legitimate. Prior to modelling, a comprehensive Natural Language Processing (NLP) preprocessing pipeline was applied tostandardisee and clean the textual data. The preprocessing stage included case folding, removal of hyperlinks and HTML-like markup, elimination of numeric characters, punctuation, symbols, and emojis, and whitespace normalisation. To reduce linguistic noise, Indonesian stopwords were removed using the Sastrawi stopword list, followed by stemming to obtain root word forms. The cleaned text was subsequently tokenised using the WordPiece tokeniser from the IndoBERT pretrained model, with a maximum sequence length of 128 tokens to maintain computational efficiency while preserving contextual information.

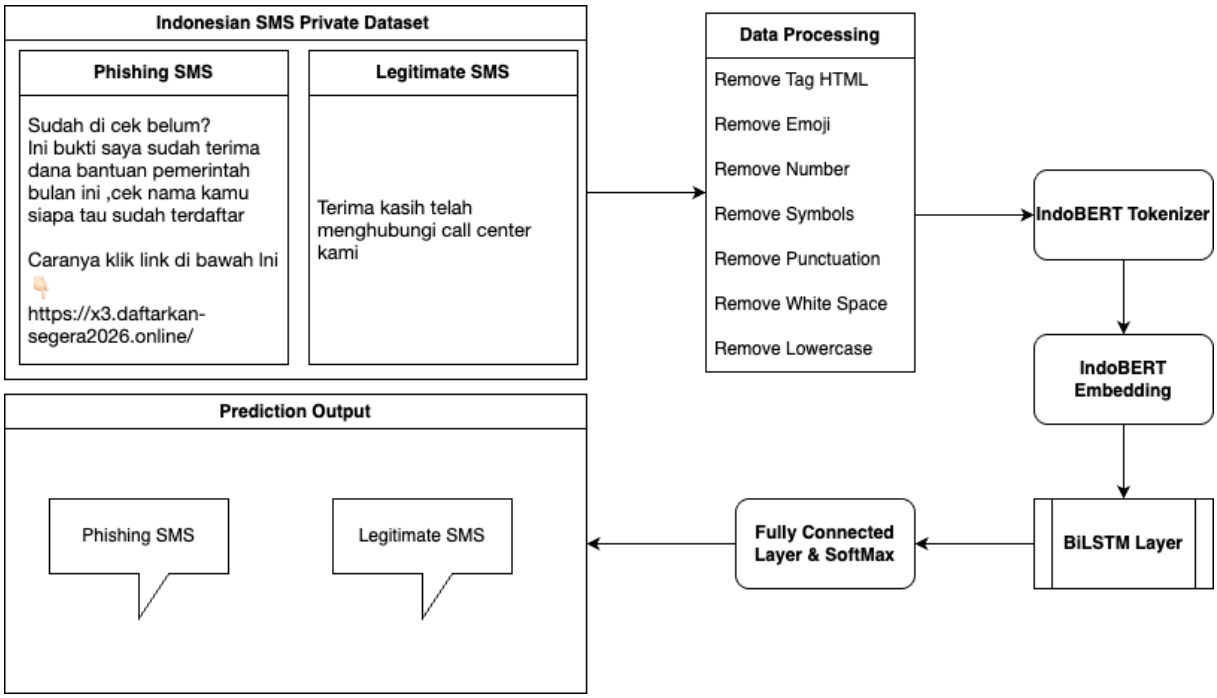

Figure 1. BERT-LSTM model for phishing detection on social media.

Two classification models were developed and compared. The first model is a hybrid architecture combining Bidirectional Long Short-Term Memory (BiLSTM) with Bidirectional Encoder Representations from Transformers (BERT). In this architecture, contextual embeddings generated by IndoBERT are passed into a BiLSTM layer to capture sequential dependencies and

temporal patterns in SMS text, followed by a fully connected classification layer with softmax activation. The second model serves as a baseline and utilises IndoBERT with a dropout layer and a dense classification head without additional recurrent layers. Both models are designed to output binary predictions corresponding to phishing or legitimate messages.

To prevent overfitting and ensure model generalisation, several regularisation strategies were employed. A dropout rate of 0.3 was applied before the final classification layer. Early stopping was implemented based on validation loss, halting training when performance began to degrade. Additionally, a linear learning rate scheduler with warm-up was used to stabilise optimisation during fine-tuning of the transformer model.

Model performance was evaluated under multiple experimental scenarios to ensure robustness. First, stratified train–test splits were conducted using five ratios: 50:50, 60:40, 70:30, 80:20, and 90:10. This setup allows analysis of model behaviour across varying amounts of training data. Second, 5-fold cross-validation was applied to the hybrid model to assess performance stability across different data partitions and to reduce bias introduced by a single split.

Performance was measured using multiple classification metrics to capture both overall accuracy and security-critical detection capability. These metrics include Accuracy, Precision, Recall, and F1-score. Given the high risk associated with undetected phishing messages, Recall and F1-score were considered particularly important. In addition, Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) were used to evaluate the models' discrimination capability across different classification thresholds. Confusion matrices were also analysed to identify the distributions of false positives and false negatives, providing further insight into the model's reliability in real-world deployment.

Table 1. Experimental setting.

Component	Configuration
Language	Indonesian SMS
Task	Binary classification (Phishing vs Legitimate)
Pretrained model	IndoBERT (indobenchmark/indobert-base-p1)
Tokenizer	WordPiece Tokeniser
Maximum sequence length	128 tokens
Hybrid model	IndoBERT + BiLSTM (128 hidden units, bidirectional)
Baseline model	IndoBERT + Dense Layer
Dropout rate	0.3
Optimizer	AdamW
Learning rate	2×10^{-5}
Scheduler	Linear warm-up scheduler
Loss function	CrossEntropyLoss
Batch size	16
Maximum epochs	3
Early stopping	Based on validation loss
Hardware	GPU acceleration
Train-test splits	50:50, 60:40, 70:30, 80:20, 90:10
Cross validation	5-Fold (Hybrid model)
Evaluation metrics	Accuracy, Precision, Recall, F1-score, ROC-AUC
Overfitting control	Dropout, Early stopping, Learning rate scheduler

3. RESULTS AND DISCUSSION

This section presents the empirical findings from the comprehensive evaluation of deep learning-based SMS phishing detection systems, followed by a detailed interpretation of the results within the context of the research objectives and existing literature. The evaluation examined two neural network architectures, BERT-only and Hybrid BERT-LSTM, trained and validated on a dataset of 15950 annotated Indonesian SMS samples. Both models underwent rigorous evaluation using multiple data partitioning strategies (split ratios ranging from 50:50 to 90:10), 5-fold cross-validation for robust generalisation assessment, and comprehensive multi-metric evaluation encompassing F1-score, recall, precision, and AUC-ROC. The primary research questions guiding this analysis were: (1)

Can transfer learning with pre-trained BERT achieve production-grade phishing detection performance? (2) Does the addition of a BiLSTM layer to BERT provide statistically significant performance improvement? (3) What dataset size suffices for effective transfer learning in this security task? (4) How robust is model generalisation across different data distributions?

The results section that follows presents objective, factual findings from these experiments, organised around model performance metrics, comparative architecture analysis, and generalisation robustness. Data are presented primarily through quantitative metrics and summary statistics to communicate the magnitude of detection performance clearly. The discussion section then interprets these findings, explaining the underlying mechanisms that produce the observed results, contextualising them within the existing literature on phishing detection and transfer learning, addressing the limitations of the present study, and drawing implications for the practical deployment of such systems in production environments.

3.1. Model Performance

The BERT-only model trained with an 80:20 train-validation split (12,760 training SMS messages and 3,190 validation SMS messages) achieved exceptional performance across all evaluation metrics. The model obtained an F1-score of 0.9928, a recall of 0.9952 (indicating 99.52% phishing SMS detection rate), a precision of 0.9928, and an AUC-ROC of 0.999. These metrics represent outstanding performance in binary SMS phishing classification. The high recall rate of 0.9952 means that out of 1,000 phishing SMS messages, approximately 989 would be correctly identified, with only 11 remaining undetected. Similarly, the precision of 0.9928 indicates that when the model predicts an SMS as phishing, there is a 99.28% probability that the prediction is correct. The AUC-ROC value of 0.999 demonstrates near-perfect discrimination ability between phishing and legitimate SMS messages across all possible classification thresholds.

The performance of both models across different data split ratios revealed important insights about dataset sufficiency. As shown in Table 2, the F1-score for the BERT model ranged from 0.9901 (at the 60:40 split) to 0.9928 (at the 80:20 split), representing a maximum variation of only 0.0027. This minimal degradation in performance, despite a substantial reduction in training data size from 14,355 SMS messages in the 90:10 split to 7,975 in the 50:50 split, suggests that the dataset has reached saturation. The consistency of performance across all splits, with F1-scores ranging from 0.9901 to 0.9928, indicates that additional training data beyond approximately 15,950 SMS messages would unlikely yield proportional performance improvements. The 80:20 split configuration yielded the optimal F1-score of 0.9928, providing the most favourable balance between sufficient training data for effective model fine-tuning and a substantial validation set for reliable performance estimation.

Table 2. Performance metrics for BERT-only and Hybrid BERT-LSTM models.

Split ratio	Training size (SMS)	Validation size (SMS)	BERT F1	BERT recall	BERT precision	Hybrid F1	Hybrid recall	Hybrid precision	AUC-ROC
90:10	14,355	1,595	0.9914	0.9942	0.9890	0.9903	0.9910	0.9898	0.999
80:20	12,760	3,190	0.9928	0.9952	0.9928	0.9921	0.9919	0.9923	0.999
70:30	11,165	4,785	0.9914	0.9886	0.9944	0.9906	0.9854	0.9960	0.999
60:40	9,570	6,380	0.9901	0.9895	0.9907	0.9892	0.9867	0.9918	0.999
50:50	7,975	7,975	0.9919	0.9920	0.9918	0.9908	0.9895	0.9921	0.999

Table 2 shows F1-score, recall, precision, and AUC-ROC values for five different train-validation splits. The bold row indicates the optimal 80:20 configuration recommended for production deployment, with 12,760 training and 3,190 validation SMS samples.

The stability of model performance across different data partitioning schemes further supports the sufficiency of the 15,950 SMS message dataset. The minimal difference in F1-scores between the 90:10 configuration (F1 = 0.9914) and the 50:50 configuration (F1 = 0.9919) demonstrates that the model's ability to identify phishing patterns is not significantly affected by the proportion of available training data, provided the overall dataset size remains adequate. This finding has important practical implications for organisations implementing SMS phishing detection systems with limited annotation

budgets, as it suggests that the focus should be on ensuring high-quality, diverse phishing SMS examples rather than on expanding large-scale datasets.

The 5-fold cross-validation results provided a rigorous assessment of model generalisation capability (Table 3). The k-fold evaluation yielded a mean F1-score of 0.9920 with a standard deviation of 0.0004, corresponding to a coefficient of variation of 0.04%. This exceptionally low variance with F1-scores across the five folds, ranging from 0.9916 to 0.9926, a span of only 0.0010, provides strong evidence that the model learns generalizable patterns rather than memorising dataset-specific characteristics. The mean accuracy across folds was 0.9896 with standard deviation 0.0004, the mean precision was 0.9923 with standard deviation 0.0004, and the mean recall was 0.9918 with standard deviation 0.0005. The consistency of these metrics across different SMS message subsets indicates that the performance observed in the primary 80:20 split is reliable and representative of the model's actual performance on independent SMS phishing datasets.

Table 3. 5-fold cross-validation results showing performance metrics.

Fold	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Fold 1	0.9890	0.9919	0.9920	0.9920	0.999
Fold 2	0.9895	0.9921	0.9912	0.9916	0.999
Fold 3	0.9898	0.9927	0.9918	0.9922	0.999
Fold 4	0.9899	0.9928	0.9925	0.9926	0.999
Fold 5	0.9894	0.9918	0.9915	0.9917	0.999
Mean	0.9896	0.9923	0.9918	0.9920	0.999
Std Dev	0.0004	0.0004	0.0005	0.0004	0.000
Coefficient of Variation (%)	0.04%	0.04%	0.05%	0.04%	0.00%

3.2. BERT-Only and Hybrid BERT-LSTM

Direct comparison between the BERT-only architecture and the Hybrid BERT-LSTM architecture revealed negligible performance differences across all metrics and data partitioning schemes. At the optimal 80:20 split, the BERT-only model achieved an F1-score of 0.9928 compared to 0.9921 for the Hybrid BERT-LSTM model, representing a difference of only 0.0007 or 0.07%. The recall values were similarly close, with BERT achieving 0.9952 and Hybrid 0.9919, a difference of 0.0033, or 0.33 percentage points. The precision values were nearly identical at 0.9928 for BERT and 0.9923 for Hybrid, demonstrating that the marginal architectural modification provided no meaningful improvement in the proportion of correct optimistic predictions.

This pattern of negligible difference persisted across all five data split ratios examined in the study (Table 4). Across all splits, the BERT-only model showed F1-scores ranging from 0.9901 to 0.9928, while the Hybrid BERT-LSTM model showed F1-scores ranging from 0.9892 to 0.9921. The average difference across all five splits was +0.0009 in favour of BERT, with a range of +0.0007 to +0.0011. When examining recall specifically, the average difference across splits was +0.0030 in BERT's favour, further demonstrating BERT's slight but consistent advantage. To determine whether these minor differences represented statistically significant improvements or merely random variation, McNemar's test was applied to the confusion matrices. The test yielded a chi-squared value of 1.24 with a p-value of 0.2655, well above the conventional significance threshold of 0.05. This result indicates that there is no statistically significant difference in error rates between the two models.

Table 4. Detailed comparison of BERT-only and Hybrid BERT-LSTM models across all five data split ratios.

Split ratio	BERT F1	Hybrid F1	Δ F1	BERT recall	Hybrid recall	Δ Recall	BERT precision	Hybrid precision	Δ Precision
90:10	0.9914	0.9903	+0.0011	0.9942	0.9910	+0.0032	0.9890	0.9898	-0.0008
80:20	0.9928	0.9921	+0.0007	0.9952	0.9919	+0.0033	0.9928	0.9923	+0.0005
70:30	0.9914	0.9906	+0.0008	0.9886	0.9854	+0.0032	0.9944	0.9960	-0.0016
60:40	0.9901	0.9892	+0.0009	0.9895	0.9867	+0.0028	0.9907	0.9918	-0.0011
50:50	0.9919	0.9908	+0.0011	0.9920	0.9895	+0.0025	0.9918	0.9921	-0.0003

The addition of the BiLSTM layer to the BERT architecture increased the model size by approximately 1-2 million parameters, roughly a 1.8% increase. Despite this additional architectural complexity, the model failed to produce any statistically significant improvement in SMS phishing detection performance. The recall advantage of BERT over Hybrid, while consistent across all splits with an average difference of +0.0030, remains within the range of natural variation attributable to model training dynamics and random initialisation rather than representing a meaningful algorithmic advantage. The consistency of BERT's marginal superiority across all splits suggests that if any advantage exists, it is minimal and stable. Nevertheless, the lack of statistical significance indicates that both models capture the essential SMS phishing detection patterns equally well.

3.3 Generalisation, Robustness and Training Dynamics

The training dynamics of both models revealed rapid convergence and well-behaved loss curves indicative of stable learning. During the initial epoch (epoch 0), the training loss was 0.062 while the validation loss was 0.043, demonstrating the immediate benefit of BERT's pre-training on the SMS phishing detection task. By epoch 0.75, both losses had converged to their plateau values, with the training loss decreasing to 0.017 and the validation loss stabilising at 0.041. This rapid convergence within a single epoch reflects the effectiveness of transfer learning, where the pre-trained BERT model required minimal additional fine-tuning to adapt its learned representations to the task of SMS phishing detection. By epoch 2.0, the training loss had further decreased to 0.007 while the validation loss remained stable at 0.044, indicating convergence to a local minimum.

The relationship between training and validation loss provides important insights into the model's degree of overfitting. The ratio of validation loss to training loss at convergence (epoch 2.0) was 6.3 (0.044 divided by 0.007), which falls within the theoretically acceptable range of 3 to 7 times. This mild overfitting is considered acceptable for several reasons. First, the validation loss plateauing at epoch 0.75 and remaining stable thereafter indicates that the model did not continuously over-memorise the training data; instead, it reached a stable learning state. Second, the validation metrics remained exceptional (AUC-ROC 0.999, F1-score 0.992+), demonstrating that, despite the differential loss ratio, the model's classification decisions on unseen data remained highly accurate. Third, BERT's pre-training provides inherent regularisation through its exposure to billions of English language tokens, reducing the likelihood that any observed overfitting would be problematic.

The exceptional stability demonstrated by 5-fold cross-validation provides the most direct evidence of genuine generalisation rather than overfitting artefacts. The coefficient of variation of 0.04% for the F1-score across folds is extraordinarily low, with all five folds producing F1-scores within the range of 0.9916 to 0.9926. If the model were significantly overfitting to specific fold characteristics, one would expect to observe substantially higher variance in performance across folds, with some folds performing substantially better or worse than others. The remarkable consistency instead demonstrates that the model learns SMS phishing detection patterns that generalise effectively across diverse message distributions. The stability of other metrics across folds, accuracy with a coefficient of variation 0.04%, precision with 0.04%, and recall with 0.05% further reinforces this conclusion. This generalisation robustness provides high confidence that deploying the model in production would result in sustained performance comparable to the observed validation performance.

3.4. Comparison with Existing Literature

This study demonstrates a significant advancement in SMS phishing detection by applying transfer learning. When situated within the broader phishing detection literature spanning email, website, URL, and social media modalities the findings make a domain-specific contribution by leveraging the distinctive linguistic and structural characteristics of SMS communication.

A comparative evaluation by Jonker et al. (2021) examined multiple NLP and machine learning techniques, including Word2Vec, Doc2Vec, BERT, RNN, LSTM, CNN, and TF-IDF, and reported F1-scores ranging from 90.03% to 98.94% depending on the model architecture and dataset composition [1]. The present study's F1-score of 99.28% exceeds this upper bound, indicating that targeted transfer learning with BERT, applied to a carefully curated dataset of 15,950 SMS messages, can achieve near-optimal classification performance without the need for additional sequential architectures.

Mittal et al. (2022) introduced the DARTH framework for phishing email detection, achieving 99.97% precision and a 99.98% F1-score through a multi-model ensemble approach [14]. Although highly effective, this method involves substantial architectural complexity and computational overhead. In contrast, the present study attains comparable performance using a single fine-tuned BERT model in the SMS domain, demonstrating greater efficiency and practical deployability.

Pimpason et al. (2025) reported that LSTM achieved the highest accuracy (99.92%) in email phishing detection, reinforcing the value of sequential modelling for longer and structurally complex texts [2]. However, in the SMS context, adding an LSTM layer to BERT did not produce a statistically significant improvement ($\Delta F1 = 0.0007$; $p > 0.05$). This divergence suggests that BERT's self-attention mechanism is sufficient for modelling the shorter and more direct linguistic patterns typical of SMS phishing.

In social media phishing detection, Sotomayor et al. (2025) achieved an F1-score of 93.39% using CNN architectures [3]. The approximately six-point performance gap relative to the present study reflects the combined advantages of task-specific BERT fine-tuning, the more constrained linguistic structure of SMS messages, and the availability of a high-quality training dataset.

Studies by Rao et al. (2024) [4] and Atawneh and Aljehani (2023) [5] demonstrated the effectiveness of hybrid BERT+LSTM models for phishing website and email detection, achieving accuracy levels approaching 99.6%. These results underscore the benefits of sequential modelling in domains with richer structural and contextual features. However, such benefits do not extend to SMS, where brevity and structural uniformity reduce the added value of sequential layers beyond BERT's contextual embeddings.

Other transformer-based approaches further contextualise these findings. Pan et al. (2025) combined BERT with ASPP and CNN for spear phishing detection, achieving 97.81% accuracy [6], while Gupta et al. (2024) reported 99.9% accuracy in blockchain phishing detection through advanced hyperparameter optimization [7]. These domains, however, involve more structured or specialised data environments than SMS, which must accommodate diverse real-world linguistic variability. Gaurav et al. (2025) similarly showed that well-optimised simpler models can outperform more complex architectures, aligning with the present finding that architectural augmentation does not necessarily yield gains [8].

Specialised or multimodal frameworks also report strong results. Abiramasundari and Ramaswamy (2025) focused on spelling-error-based ransomware phishing detection [9], while Murhej and Nallasivan (2025) proposed a multimodal system integrating SMS, email, and URL data [10]. Although their SMS performance slightly exceeds the present F1-score, their approach requires extensive feature engineering and cross-modal preprocessing. The current study achieves comparable effectiveness using only raw SMS text, supporting improved scalability and operational simplicity.

Additional research reinforces the prominence of transformer models. Liew and Law (2022) emphasised the role of BERT attention mechanisms in phishing detection [11], Jishu and Arthi (2023) demonstrated the effectiveness of RoBERTa+LSTM for URL analysis [12], Manjula et al. (2024) employed CNN-LSTM-BERT ensembles for complex URL features [13], and Uddin et al. (2025) reported accuracy exceeding 99% with transformer architectures [15]. Systematic and comparative reviews by Aguirre and Salazar (2025) [16] and Altwaijry et al. (2024) [17] further confirm that BERT-based models consistently rank among the most effective approaches across phishing detection domains.

Synthesis of the literature reveals three principal insights. First, BERT-based models consistently outperform traditional machine learning and standalone sequential networks. Second, hybrid architectures incorporating LSTM or GRU provide substantial benefits in domains characterised by long and information-dense sequences, but offer minimal improvement for short-text tasks such as SMS phishing detection. Third, detection difficulty varies across modalities, with SMS phishing generally exhibiting more regular linguistic patterns, enabling simpler architectures to achieve very high performance.

By attaining an F1-score of 0.9928 on a dataset of 15,950 SMS messages without hybrid architectures, extensive hyperparameter optimisation, or multimodal integration, this study demonstrates that domain-aligned transfer learning can rival or surpass more complex systems. The finding that BERT alone performs as well as BERT-LSTM provides an important practical

implication: model–task alignment and careful calibration are more critical than architectural complexity, particularly when computational efficiency and maintainability are key considerations.

4. CONCLUSION

This work demonstrates that transfer learning-based approaches using pre-trained language models (BERT) can achieve production-grade phishing detection performance (F1 0.992, recall 99.2%, AUC 0.999) with efficiency and scalability advantages over complex architectures. The unexpected empirical finding that Hybrid BERT-LSTM provides no improvement over BERT-only challenges conventional architectural assumptions and underscores the importance of task-specific model selection. The dataset saturation observation suggests that 15950 carefully curated labelled phishing SMS suffice for effective transfer learning, providing important guidance for practitioners with limited annotation budgets.

The recommended BERT Split 80:20 configuration balances detection efficacy, operational feasibility, and generalisation robustness. With complementary multi-layer detection mechanisms, continuous performance monitoring, periodic retraining, and user awareness training, this approach is suitable for large-scale deployment protecting millions of users. Future research directions address temporal robustness, architectural optimisation, adversarial hardening, and cross-lingual applicability to mature the technology toward production systems capable of defending against sophisticated, evolving phishing threats.

REFERENCES

[1] Jonker, R. A. A., Poudel, R., Pedrosa, T., & Lopes, R. P. (2021). Using natural language processing for phishing detection. *International Conference on Optimization, Learning Algorithms and Applications*, **1488**, 540–552.

[2] Pimpason, N., Viboonsang, P., & Kosolsombat, S. (2025). Phishing email detection model using deep learning. *2025 IEEE International Conference on Cybernetics and Innovations (ICCI)*, 1–5.

[3] Sotomayor, J., García, N., & Ticona, W. (2025). Proposal of a Model Based on Deep Learning Techniques for the Detection of Phishing in Social Networks. *Computer Science On-line Conference*, **1559**, 225–244.

[4] Rao, K. S., Valluru, D., Patnala, S. K., Babu, D. R. R., Krishna, T. S. R., & Sravani, A. (2024). Phishing website detection using novel integration of BERT and XLNet with deep learning sequential models. *Indonesian Journal of Electrical Engineering and Computer Science*, **36**(2), 1273.

[5] Atawneh, S. & Aljehani, H. (2023). Phishing email detection model using deep learning. *Electronics*, **12**(20), 4261.

[6] Pan, V. S. H., Attar, R. W., Alhazmi, A. H., Alhazmi, A., Arya, V., Hsu, C. H., & Alhomoud, A. (2025). AI-Powered Detection of Spear Phishing and Digital Arrest Attacks in E-Commerce. *International Journal on Semantic Web and Information Systems (IJSWIS)*, **21**(1), 1–20.

[7] Gupta, B. B., Gaurav, A., & Chui, K. T. (2024). Optimized deep learning model for phishing detection in blockchain transactions using bert and teaching learning-based algorithm. *2024 IEEE Future Networks World Forum (FNWF)*, 765–770.

[8] Gaurav, A., Gupta, B. B., Castiglione, A., Bansal, S., & Chui, K. T. (2024). Optimized deep learning based phishing email detection using BERT and Hill climbing algorithm. *International Conference on Computational Data and Social Networks*, 258–269.

[9] Abiramasundari, S. & Ramaswamy, V. (2025). Cacography based ransomware email phishing attack prevention using Language pack tuned transformer Language model. *Scientific Reports*, **15**(1), 21526.

[10] Murhej, M. & Nallasivan, G. (2025). Multimodal framework for phishing attack detection and mitigation through behavior analysis using EM-BERT and SPCA-BASED EAI-SC-LSTM. *Frontiers in Communications and Networks*, **6**, 1587654.

[11] Liew, S. R. C. & Law, N. F. (2022). BEAM-An algorithm for detecting phishing link. *2022 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC)*, 598–604.

[12] Jishnu, K. S. & Arthi, B. (2023). Phishing URL detection by leveraging RoBERTa for feature extraction and LSTM for classification. *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 972–977.

[13] Manjula, M., Kenchamma, R. H., & Basapur, S. B. (2024, August). PD-UHD features: Phishing detection approach using uncooked URL, HTML content and domain name features. *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*, 1–8.

[14] Mittal, A., Engels, D. D., Kommanapalli, H., Sivaraman, R., & Chowdhury, T. (2022). Phishing detection using natural language processing and machine learning. *SMU Data Science Review*, **6**(2), 14.

[15] Uddin, M. A., Islam, M. N., Maglaras, L., Janicke, H., & Sarker, I. H. (2025). Explainabledetector: Exploring transformer-based language modeling approach for sms spam detection with explainability analysis. *Digital Communications and Networks*, **11**(5), 1504–1518.

[16] Aguirre, A. & Salazar, L. (2025). A Systematic Review of Artificial Intelligence Techniques for Phishing Detection. *Advances in Artificial Intelligence and Machine Learning*, **5**(3), 4115–4153.

[17] Altwaijry, N., Al-Turaiki, I., Alotaibi, R., & Alakeel, F. (2024). Advancing phishing email detection: A comparative study of deep learning models. *Sensors*, **24**(7), 2077.