Jurnal Testing dan Implementasi Sistem Informasi
Vol 4(1) 2026 : 87-101

PHISHING DETECTION MODEL ON SOCIAL MEDIA ENHANCED WITH CNN AND BERT

Nurliana Nasution[1*], Wenni Syafitri[2], Feldiansyah[3]
Magister Ilmu Komputer, Universitas Lancang Kuning, Indonesia[123]
nurliananst@unilak.ac.id[1], wenni20@gmail.com[2], feldiansyah@unilak.ac.id[3]

Received: 26 March 2026, Revised: 27 May 2026, Accepted: 28 May 2026
**Corresponding Author*

ABSTRACT
Phishing on social media has become an increasingly serious cyber threat because attackers exploit persuasive language, conversational context, and dynamic interaction patterns to deceive users. This study proposes a hybrid CNN-BERT model for detecting phishing content in Indonesian social media text by combining BERT's contextual semantic representation with CNN's ability to capture locally relevant textual patterns. The dataset was preprocessed to remove noise, normalize writing variations, and prepare the text for deep learning analysis; class proportions were also examined to support fairer evaluation. Model performance was assessed under multiple data-splitting scenarios and cross-validation to examine robustness and consistency. The experimental results indicate that the proposed hybrid model achieves strong and stable performance across accuracy, precision, recall, and F1-score, and outperforms the baseline model when the BERT backbone is frozen. However, when BERT is fully fine-tuned, the performance gain from the CNN layer becomes marginal, suggesting that strong contextual representations are already highly effective for this task. These findings indicate that integrating CNN and BERT is effective for phishing detection on social media, although domain adaptation challenges, overfitting risk, and real-world deployment constraints remain important considerations. The novelty of this work lies in systematically comparing frozen versus fully fine-tuned IndoBERT backbones with and without a CNN head for Indonesian short-message phishing detection.
Keywords : *CNN-BERT, IndoBERT, Phishing detection, Natural language processing, Social Media Phishing.*

1. Introduction

Phishing remains one of the most persistent and adaptive forms of cybercrime because attackers continuously refine their persuasive strategies, linguistic styles, and delivery channels to evade both human judgment and automated security controls(Koto et al., 2020; Wicaksana et al., 2026). In recent years, phishing has expanded beyond conventional email and increasingly appears in SMS, mobile messaging, and social media environments, where messages are often short, conversational, and designed to exploit urgency, trust, and platform familiarity(Ghourabi, 2021; Saidat et al., 2024; Wei et al., 2025). This evolution has made text-based phishing detection more difficult, particularly in multilingual and informal communication settings where malicious intent is embedded in compact and context-sensitive language(Alhuzali et al., 2025; Alotaibi, 2026; Syafitri et al., 2026).

Despite extensive progress in phishing detection research, traditional machine learning (ML) approaches exhibit notable limitations in the dynamic landscape of social media. First, conventional models depend heavily on manual feature engineering of URLs, web content, and third-party metrics, typically achieving 80–97% accuracy. For instance, (Kurnianda et al., 2025) employed ensemble learning using 25 handcrafted features and reported an accuracy of 97.19%, while (Lamas Piñeiro & Wong Portillo, 2022) achieved 80% precision with Classification Trees. Second, these models struggle to adapt to evolving phishing tactics involving deepfakes and generative AI. Third, URL-based methods alone are inadequate on social media, where attacks exploit multimodal signals text content, user metadata, and behavioral interactions.

A growing body of research has attempted to address phishing detection using machine learning and deep learning techniques, but the literature still shows several conceptual and methodological limitations(Yuan et al., 2024). Earlier studies largely depended on handcrafted lexical, structural, and URL-based features, which can perform adequately in relatively stable

environments but often degrade when attackers alter wording, imitate legitimate communication patterns, or move across platforms with different discourse conventions. More recent work has shifted toward neural architectures that reduce reliance on manual feature engineering, yet a considerable portion of this research remains centered on English-language email corpora, phishing URLs, or enterprise settings rather than short Indonesian messages that more closely resemble mobile and social-media communication(Gupta et al., 2024; Uddin et al., 2026; Yasinta Roesmiatun & Zahra, 2025). As a result, the current literature still lacks a sufficiently critical synthesis of how phishing detection methods generalize across multilingual, cross-platform, and short-text contexts.

Advances in deep learning have introduced automatic feature extraction capabilities, mitigating the dependency on manual engineering. (N. H. Hassan & Fakharudin, 2023) compared Convolutional Neural Networks (CNN) and Artificial Neural Networks (ANN) for phishing website classification, finding that a modified ANN performed marginally better but required complex preprocessing. (Brindha et al., 2023) enhanced a GRU-based model with an Intelligent Cuckoo Search optimizer, achieving 99.72% accuracy. Nevertheless, these approaches primarily focus on URL and email data, overlooking the textual and multimodal nature of social media phishing.

This gap is especially important for Indonesian-language phishing detection. Although Bahasa Indonesia is spoken by a very large population, Indonesian remains comparatively underrepresented in natural language processing resources and benchmarked cybersecurity text studies(Koto et al., 2020). The introduction of IndoLEM and IndoBERT marked a major advance by providing benchmark datasets and a pre-trained Indonesian language model that achieved state-of-the-art performance across multiple Indonesian NLP tasks. However, the availability of a strong Indonesian transformer model does not automatically resolve the phishing detection problem, because phishing messages often exhibit highly localized textual cues, such as compressed commands, suspicious number patterns, shortened links, and persuasive call-to-action expressions, which may not be fully captured by a single global semantic representation. Consequently, the research problem is not only whether transformer-based models can classify Indonesian phishing text accurately, but also whether their contextual strengths can be complemented by mechanisms that better exploit local discriminative patterns.

Transformer-based architectures have emerged as state-of-the-art solutions, introducing bidirectional attention mechanisms that enable deep semantic understanding. Models such as BERT can capture long-range dependencies, facilitating the detection of subtle linguistic manipulations in text and URL patterns. (Maneriker et al., 2021) fine-tuned BERT in a model called URLTran, achieving a true positive rate (TPR) of 86.80% at a false positive rate (FPR) of 0.01%, outperforming a baseline CNN by 21.9%. (Gupta et al., 2024) integrated BERT embeddings into CNN for email phishing detection, demonstrating superior overall performance. Similarly, (Otieno et al., 2023) confirmed that BERT effectively detects phishing URLs without manual feature engineering, signaling a paradigm shift in cyber threat detection.

Transformer-based language models are a natural foundation for this task because they have transformed text classification by learning deep bidirectional contextual representations that can be fine-tuned with minimal architectural modification(Z. L. Hassan et al., 2026). BERT, in particular, demonstrated that pre-trained bidirectional transformers can achieve state-of-the-art results across a wide range of NLP tasks, thereby establishing a foundational paradigm for downstream classification problems. For Indonesian-language applications, IndoBERT is especially relevant because it is trained on Indonesian corpora and has been empirically validated on Indonesian benchmarks, making it more appropriate than general multilingual models when the target task depends on language-specific lexical and syntactic patterns . Nevertheless, while transformer architectures are highly effective at capturing contextual semantics, they are not necessarily optimized to emphasize short local patterns that can be highly informative in phishing messages, particularly in short-text environments such as SMS and direct messaging(Ulfath et al., 2021).

This limitation helps justify the integration of convolutional neural networks with transformer encoders. CNN-based text classifiers are well known for their capacity to detect salient local n-grams and phrase-level structures, and hybrid BERT-CNN architectures have

shown promising performance in phishing and related text classification tasks. In phishing detection, this hybridization is theoretically attractive because IndoBERT can model broader semantic context while CNN can highlight token-level or phrase-level regularities, such as imperative prompts, transactional keywords, phone numbers, and suspicious lexical sequences, that are frequently associated with deceptive messages(Ghourabi, 2021). Recent evidence from phishing-related studies supports this rationale: BERT-based smishing detection has shown high accuracy in multilingual SMS settings, and BERT-CNN combinations have also demonstrated strong performance in enterprise phishing email classification. Even so, the literature still offers limited insight into whether CNN augmentation provides meaningful added value once a transformer backbone is fully fine-tuned, particularly in Indonesian-language phishing detection scenarios(Tenney et al., 2019).

The evolution of phishing detection reflects a clear trajectory from traditional machine learning to deep learning and transformer-based methods. (Kurnianda et al., 2025) achieved 97.19% accuracy by employing ensemble models such as Random Forest, XGBoost, and LightGBM with 25 features derived from URLs, content, and third-party sources. Alqahtani et al. (2022) further optimized deep autoencoders using Adaptive Affinity Aggregation (AAA) and Invasive Weed Optimization (IWO), achieving 99.28% accuracy and showcasing the potential of bio-inspired optimization techniques.

CNN architectures extend feature extraction to spatial structures. (N. H. Hassan & Fakharudin, 2023) found that while modified ANNs slightly surpassed CNNs in accuracy, CNNs require less complex preprocessing. (Brindha et al., 2023) achieved 99.72% accuracy using a GRU combined with Intelligent Cuckoo Search optimization, highlighting the synergy between metaheuristic algorithms and deep learning. Nonetheless, methods reliant solely on URLs or emails fall short in the social media context, where diverse modalities text, user behavior, and multimedia contribute to phishing signals. (Naswir et al., 2022) demonstrated that incorporating both email attributes and human behavioral data with PyCaret reached 98% accuracy, underlining the importance of multimodal integration. Meanwhile, (Lamas Piñeiro & Wong Portillo, 2022), using Random Forest, Classification Trees, and SVM on 112 URL-based features, obtained only 80% precision, reaffirming the inadequacy of URL-only models in dynamic environments.

Transformers, addressing semantic understanding through bidirectional encoding, have proven superior in phishing detection. (Maneriker et al., 2021) enhanced BERT via fine-tuning in URLTran, yielding an 86.80% TPR with 0.01% FPR 21.9% higher than CNN performance. (Gupta et al., 2024) demonstrated that integrating BERT embeddings into CNN further improved email phishing detection, surpassing single-model baselines. (Otieno et al., 2023) validated BERT's efficiency through phishing URL detection without feature engineering, evidencing the strength of domain pre-training.

Recent research extends beyond textual analysis, encouraging graph-based and multimodal approaches. (Fu et al., 2024) integrated Graph Convolutional Networks (GCN), LSTM, and SMOTE for Ethereum phishing detection, achieving 97.22% accuracy. Their findings emphasize the importance of temporal and relational feature modeling, principles that align closely with social media's complex interaction dynamics. Across these strands of work, three structural limitations remain particularly relevant for Indonesian short-message phishing detection. First, CNN- and RNN-based models that operate on character- or word-level n-grams excel at capturing local patterns but are typically trained on English-language email or URL corpora, limiting their ability to model sociolinguistic variation in SMS or social-media messages written in low-resource languages(Brindha et al., 2023; Naswir et al., 2022; Tanbhir et al., 2024). Second, transformer-based detectors such as BERT and its multilingual variants provide strong contextual representations yet are most often evaluated on a small number of English datasets, which raises questions about how well they transfer to multilingual smishing contexts and informal mobile communication (Ghourabi, 2021; Maneriker et al., 2021; Wei et al., 2025). Third, only a few studies explicitly analyze hybrid CNN–BERT architectures under controlled frozen versus fine-tuned regimes or discuss how model performance changes when moving from SMS to broader social-media environments, leaving an important gap regarding generalization and cross-platform robustness (Gupta et al., 2024; Johari et al., 2025; Preeti & Sharma, 2024).

The present study is motivated by this unresolved question and by a second conceptual issue identified in prior formulations of the topic: the mismatch between SMS data and the broader claim of social media phishing detection. This inconsistency must be addressed carefully in the introduction because SMS and social media are not identical domains, even though both involve short, persuasive, user-facing messages. A stronger conceptual position is to frame the dataset as Indonesian short-message phishing data with methodological relevance to social-media phishing, rather than to claim that SMS evidence directly represents the entire social media ecosystem. Such a formulation is more defensible because it acknowledges domain differences while also recognizing that both settings share important linguistic and behavioral characteristics, including brevity, informality, urgency cues, and user-targeted deception. This clarification reduces conceptual overreach and improves alignment between the title, dataset, and methodological claims of the study.

Against this background, the specific research gap can be stated more clearly. Existing phishing detection studies have shown the potential of transformer-based models and hybrid deep learning architectures, but the literature still provides limited evidence on Indonesian-language phishing detection using a domain-appropriate transformer such as IndoBERT, especially under a systematic comparison between frozen and fine-tuned settings with CNN augmentation . The need for such work is amplified by the multilingual nature of phishing campaigns, the relative scarcity of Indonesian cybersecurity text resources, and the practical importance of building models that remain robust in short, noisy, and highly variable communication contexts. Therefore, the present study does not merely apply an established model to a new dataset; it also examines whether hybrid IndoBERT-CNN modeling offers a justified architectural advantage for Indonesian phishing text classification.

Based on these considerations, this study proposes a hybrid IndoBERT-CNN framework for detecting phishing messages in Indonesian short-text communication. The study is designed to evaluate four settings that distinguish between frozen and fine-tuned transformer backbones and between transformer-only and hybrid CNN-enhanced classifiers, thereby enabling a more precise analysis of where the performance gains actually originate . This design is important because improvements observed in hybrid architectures may stem either from convolutional feature extraction itself or from the broader effect of full transformer adaptation; without explicit comparison, these effects are difficult to separate. By structuring the problem in this way, the study contributes both empirically and conceptually to Indonesian phishing detection research.

Accordingly, this study makes three principal contributions. First, it addresses the underexplored problem of Indonesian-language phishing detection using IndoBERT as a language-specific transformer backbone that is more appropriate for Bahasa Indonesia than generic multilingual representations. Second, it provides a stronger methodological rationale for combining IndoBERT with CNN by linking contextual semantic modeling with local pattern extraction in short phishing messages. Third, it resolves the conceptual inconsistency between SMS-based evidence and social-media-oriented claims by positioning the dataset as a short-message phishing domain whose findings are relevant to, but not identical with, broader social media phishing detection. Taken together, these contributions position IndoBERT–CNN as a rigorously evaluated, language-specific alternative to prior phishing detectors that either relied on generic multilingual transformers or did not explicitly disentangle the roles of contextual and local pattern representations in Indonesian short-text phishing scenarios (Ghourabi, 2021; Gupta et al., 2024; Pires et al., 2019; Tanbhir et al., 2024; Ulfath et al., 2021).

In summary, ensemble models and CNNs demonstrate proficiency in spatial and statistical pattern recognition but struggle with semantic understanding. Conversely, transformers excel in semantic comprehension yet often overlook structural or platform-specific cues. Hybrid architectures that integrate CNN and BERT combine these strengths structural hierarchy and contextual understanding offering a robust foundation for identifying multimodal phishing signals on social media(Roumeliotis et al., 2024; Tanbhir et al., 2024). Few prior studies, such as (Gupta et al., 2024), have explored this direction. Their hybrid CNN–BERT model achieved 97.5% accuracy, confirming its potential. Nonetheless, further improvements remain possible through parameter optimization and evaluation across multiple social media datasets to assess generalization and efficiency.

2. Research Methods

The experiments are conducted on a supervised binary text classification task for Indonesian SMS phishing detection, using a dataset of 15950 messages labeled as “phishing” (label 1) and “non-phishing” (label 0). The dataset was annotated by cybersecurity experts to ensure label reliability and domain validity, while no external data sources were used because the corpus is proprietary and restricted. The class distribution is moderately imbalanced, with 10 316 phishing instances (64.68 %) and 5 634 non-phishing instances (35.32 %), which implies that recall on the phishing class and overall decision thresholds are particularly important for practical deployment. All messages are preprocessed through a standardized Indonesian text cleaning pipeline that includes lowercasing, URL removal, punctuation and digit normalization, emoji removal, whitespace normalization, and Indonesian stemming and stopword removal, producing a “clean_text” field used as model input. Example preprocessed messages show that typical SMS artifacts (see Figure 1) such as leetspeak digits, punctuation variants, and promotional tokens are normalized into stemmed Indonesian word forms, thereby emphasizing morphological roots and reducing noise from orthographic variation, which is consistent with common Indonesian NLP preprocessing practices(Ramdhan et al., 2022). Using cybersecurity experts instead of generic crowdworkers for annotation aligns with recommendations from recent smishing and SMS spam studies, which emphasize that expert judgement is crucial for correctly distinguishing aggressive marketing, benign alerts, and genuinely malicious phishing attempts in short-message channels (Ghourabi, 2021; Johari et al., 2025; Saidat et al., 2024). At the same time, relying on a proprietary corpus collected from operational messaging flows means that the dataset is likely to capture realistic noise and attack diversity but may also embody sampling bias with respect to telecom operator, time period, and user demographics, so the reported results should be interpreted as characterizing a specific Indonesian SMS phishing environment rather than the entire landscape of Indonesian digital communication. The experiments were run on Google Colab using an NVIDIA A100 GPU, which provided the computational capacity required for fine-tuning transformer-based models efficiently.

	text	clean_text
0	4nda Telah Trdft4r Di BPJS Mnd-dpatkan Bantuan...	nda trdftr bpjs mnddpatkan bantu rpjt kode loc...
1	Selamat tlah M-dapatkan!!! SUBSIDI - PEMERINTA...	selamat tlah mdapatkan subsidi perintah rpjt p...
2	INFO PINJ4M4N D4N4 ONLINE PROSES MUD4H/CEP4T P...	info pinjmn dn online proses mudhcept pinjmn m...
3	KAMI D4RI TEAM'BAPAUFAMILY' MENYAMPAIKAN ANDA ...	dri teambapaufamily sampai pilih raih juta cod...
4	Anda dapat Rp.125.jt dari S-h-o-pe3 P1N PEMEN4...	rpjt shope pn pemenng blw nfo klik

Figure 1. Cleaned dataset

For tokenization and contextual embedding, the study employs IndoBERT-base-p1 from the IndoBenchmark suite via the HuggingFace transformers library. Texts are tokenized with BertTokenizer.from_pretrained("indobenchmark/indobert-base-p1") with padding and truncation to a maximum length of 128 tokens, and the tokenized inputs (input_ids, attention_mask) are fed into BertModel as the shared backbone for all variants. Two BERT-only classifiers are instantiated through a BERTOnly module that applies the BERT pooled output (pooler_output) followed by dropout ($p = 0.5$) and a linear layer from 768 to 2 logits, with an optional finetune flag controlling whether BERT parameters are updated or frozen. When finetune=False, all BERT parameters have requires_grad set to False, so only the final classifier layer is trained; when finetune=True, the full stack including BERT is optimized jointly. This configuration isolates the contribution of fine-tuning the language model versus using IndoBERT as a fixed feature extractor.

The hybrid CNN–BERT architecture (CNNBERT module) shares the same IndoBERT backbone but uses the last hidden state instead of the pooled output, thereby exposing full token-level representations to the convolutional head. After obtaining the last_hidden_state tensor, the model permutes it to (batch, channels, sequence_length) and applies a 1D convolution with 768 input channels, 256 output channels, kernel size 3, and padding 1, followed by ReLU activation

and an adaptive max pooling over the temporal dimension to a single vector of size 256. This pooled feature is regularized with dropout ($p = 0.5$) and then passed to a linear classification layer ($256 \rightarrow 2$), yielding four variants overall: BERT Frozen, BERT FineTune, CNN-BERT Frozen, and CNN-BERT FineTune. As with BERTOnly, the finetune flag controls whether the IndoBERT backbone is frozen or updated jointly with the CNN head, which allows a direct comparison of frozen versus finetuned pre-trained features across both architectural families(Soni et al., 2023).

The training procedure uses a unified training loop built on Adam optimization with learning rate 2×10^{-5} and cross-entropy loss. For each experiment, the model is trained for five epochs with mini-batches of size 16 using a DataLoader over an SMSDataset wrapper, which stores tokenized encodings and integer labels. The choice of five training epochs was guided by convergence behavior observed during training, where validation performance stabilized without indicating substantial overfitting(Kim et al., 2021). At each epoch, the loop accumulates total training loss and computes average training loss per batch, while validating on a held-out set with the same batching configuration and recording mean validation loss. The experiments adopt a stratified train–test split strategy for different training proportions (0.5, 0.4, 0.3, 0.2, 0.1), ensuring that class proportions are preserved across splits and that model behavior can be observed as the amount of labeled training data varies. In addition, the notebook defines a KFold cross-validation setup (5-fold, shuffle=True, random_state=42) and performs cross-model paired t-tests on F1-scores to examine whether observed differences between CNN–BERT FineTune and BERT FineTune are statistically significant across fold-level results. The use of multiple train-test split ratios was intended to evaluate model robustness under varying supervision levels and to reduce dependence on a single partition(Sivakumar et al., 2024). The learning rate of 2×10^{-5}, batch size of 16, and five training epochs follow common fine-tuning regimes for BERT-like models on medium-scale text classification problems, which have been shown to offer a good balance between convergence speed and overfitting control in both general NLP and security-oriented phishing detection tasks (Kim et al., 2021; Qin & Zhang, 2024; Uddin et al., 2026).This design yields a consistent experimental framework where IndoBERT is either frozen or finetuned, with classifier heads that are either purely linear or augmented by convolution and temporal pooling, evaluated under both single splits and k-fold validation(Abdillah & Insani, 2025).

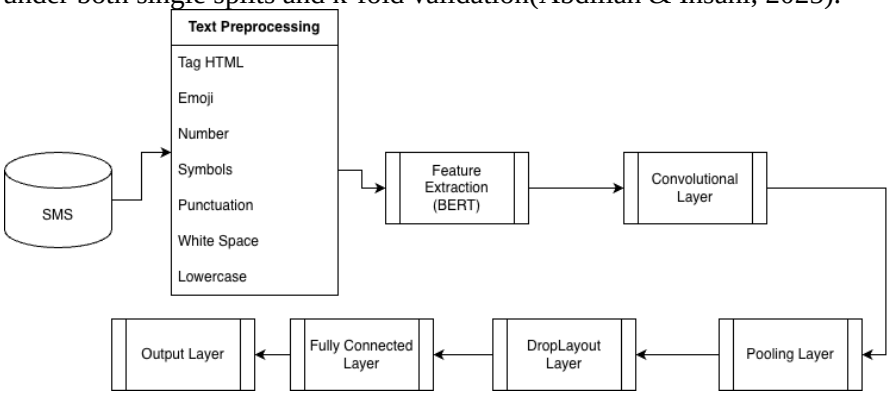

Figure 2. Proposed Social Media Phishing Detection Model enhanced with CNN and BERT

Although the dataset consists of SMS messages, the proposed framework is applicable to other short-text deception environments, including social media, due to the shared linguistic characteristics of brevity, informality, and persuasive intent(Johari et al., 2025).

3. Results and Discussions

Across the experiments, the training configuration and validation strategy jointly determine how each model variant leverages contextual information and label supervision. When BERT is frozen, the classifier either linear or CNN-based adapts to a static representation space, and learning capacity resides primarily in the shallow head; in contrast, full fine-tuning allows both the backbone and head to adapt to the phishing detection task, potentially improving sensitivity to dataset-specific lexical and syntactic patterns. The inclusion of a convolutional layer in CNN–BERT exposes local n-gram structures in IndoBERT's token-level embeddings to a learned

aggregation mechanism, which can be particularly relevant for capturing short, discriminative phrases common in phishing SMS (such as monetary amounts, action prompts, and entity names). The stratified splits, combined with consistent training hyperparameters, mean that comparisons across models and splits can be attributed to differences in architecture and fine-tuning rather than to changes in optimization settings(Christian et al., 2025).

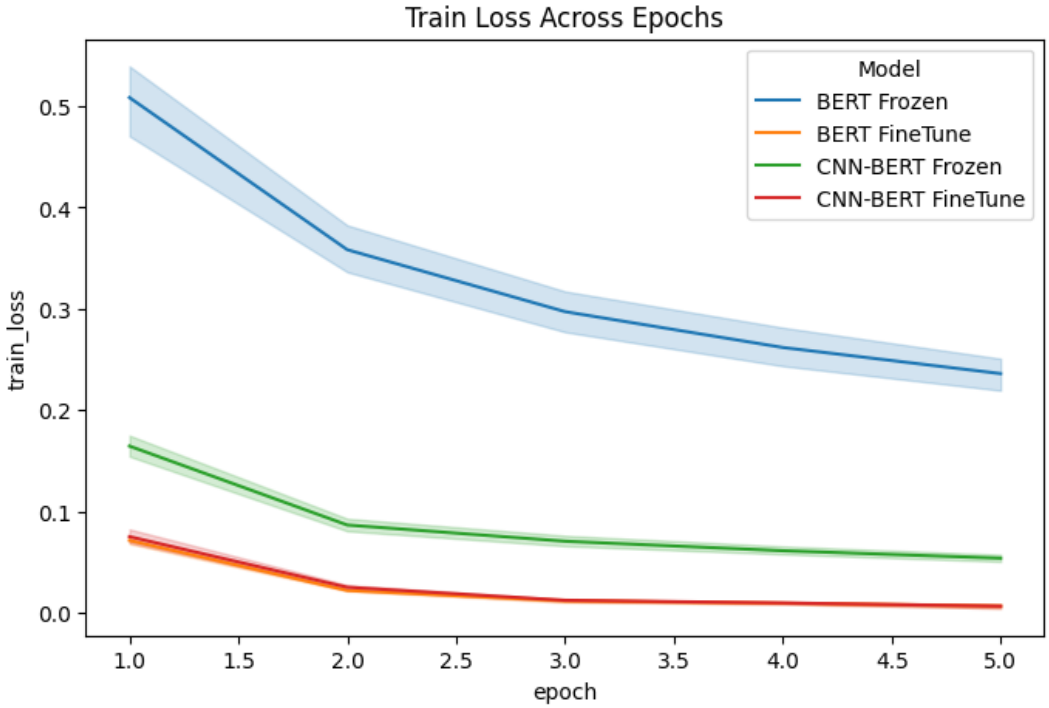

Figure 3. Train and Validation Loss

The training and validation curves for each model variant, plotted separately for each split, exhibit characteristic convergence patterns consistent with deep text classifiers trained on moderately sized datasets (see Figure 3). In the BERT Frozen configuration at a 0.5 split, the training loss decreases steadily over the five epochs, while validation loss also declines but tends to plateau earlier, indicating that the linear classifier is able to adapt to the fixed IndoBERT features without severe overfitting within the five-epoch regime. The corresponding accuracy and F1 curves rise quickly in the initial epochs and then stabilize, suggesting that most of the discriminative capacity achievable with frozen IndoBERT features is reached early in training. In contrast, with BERT FineTune at the same split, the training loss falls more sharply and to lower values, and validation loss continues to decrease with a tighter alignment between training and validation trajectories, reflecting the greater capacity and flexibility of the fully trainable model. The CNN-BERT Frozen curves show training loss declining and validation loss decreasing somewhat more slowly than in BERT Frozen, but achieving lower final validation loss consistent with higher test metrics, indicating that the CNN head is effectively leveraging token-level representations even when IndoBERT is frozen. The CNN-BERT FineTune loss curves converge rapidly with low variance across epochs, with both training and validation losses reaching low plateaus, reflecting stable optimization behavior despite the increased number of trainable parameters, which aligns with reported stability of BERT+CNN hybrids in other text classification tasks(Biswas et al., 2025).

In terms of stability, none of the loss curves exhibit oscillatory or diverging behavior within the five epochs across the reported splits, indicating that the selected learning rate and batch size are appropriate for fine-tuning IndoBERT in this setting. Validation accuracy curves closely track training accuracy curves, with only minor gaps, suggesting limited overfitting under the chosen number of epochs and regularization (dropout). For smaller training splits (e.g., 0.2 and 0.1), the accuracy curves still show monotone increases with epochs but converge to slightly lower plateaus, which is expected given reduced training data; nonetheless, the convergence remains

stable, and the models do not exhibit rapid degradation in validation performance. Where confusion matrices are plotted for example, for 0.5 splits under BERT Frozen and CNN-BERT Frozen most mass lies along the diagonal for both classes, with higher counts in the phishing (label 1) cell reflecting the underlying data imbalance. Off-diagonal entries show that BERT Frozen has a noticeable number of false negatives and false positives, whereas CNN-BERT Frozen reduces both types of errors, yielding more concentrated diagonal cells; CNN-BERT FineTune further reduces misclassifications for both classes, with particularly high true positive counts for phishing messages. This behavior is consistent with broader findings that hybrid deep models can reduce misclassification rates in spam and phishing detection relative to simpler baselines(Adel Al-Zebari, 2025). Given the confusion matrices and the residual false positives and false negatives, the remaining errors are likely dominated by two types of failure modes that are also reported in smishing and SMS spam study. First, legitimate promotional or notification messages that reuse lexical patterns common in phishing templates for example, urgent calls to action combined with monetary amounts or reward claims are prone to being flagged as phishing, reflecting the intrinsic difficulty of separating persuasive but benign communication from truly malicious content (Ghourabi, 2021; Johari et al., 2025; Ramdhan et al., 2022). Second, a small fraction of highly abbreviated or obfuscated phishing messages with minimal lexical context tend to be misclassified as legitimate, illustrating the limits of purely text-based models when attackers deliberately minimize informative tokens (Rahmapuri et al., 2025). From an explainability perspective, the comparative behavior of BERT-only and CNN–BERT variants suggests that IndoBERT's self-attention layers already capture most task-relevant semantics, while the CNN head primarily sharpens decision boundaries by emphasizing short n-gram patterns, a role consistent with analyses of hybrid architectures in smishing and spam detection (Roumeliotis et al., 2024; Tanbhir et al., 2024; Ulfath et al., 2021). In deployment, these observations imply that the proposed models are best used as high-recall back-end filters in SMS gateways or social-media moderation pipelines, complemented by rule-based constraints or business logic for borderline campaigns and periodic recalibration to account for dataset drift and emerging phishing templates (Fu et al., 2024; Preeti & Sharma, 2024).

	Split	Model	Accuracy	Precision	Recall	F1	Train_Time_sec
0	0.5	BERT Frozen	0.917868	0.899415	0.982939	0.939324	134.306022
1	0.5	BERT FineTune	0.988339	0.991843	0.990112	0.990977	270.641140
2	0.5	CNN-BERT Frozen	0.975549	0.973659	0.988949	0.981245	136.293655
3	0.5	CNN-BERT FineTune	0.987712	0.990310	0.990694	0.990502	271.996868
4	0.4	BERT Frozen	0.934639	0.923112	0.980611	0.950993	135.134666

Figure 4. Main result

The strong performance of the proposed models suggests that IndoBERT already captures substantial contextual cues in Indonesian phishing messages, while the CNN layer provides additional gains mainly in the frozen setting by emphasizing short, discriminative local patterns such as imperative phrases, numeric cues, and lexical irregularities that are common in phishing text. This also explains why the benefit of adding CNN becomes marginal after full fine-tuning: once IndoBERT is updated on the task, its contextual representations can already absorb many of the local and semantic features that the convolutional head is intended to highlight, leading to similar convergence behavior between BERT Fine-Tune and CNN-BERT Fine-Tune. From a practical standpoint, the high accuracy and F1-scores should be interpreted alongside the class distribution and deployment objective, because phishing detection systems typically prioritize recall while still controlling false positives to avoid user fatigue and unnecessary blocking. Accordingly, the reported gains are not only numerically strong but also operationally meaningful, especially in a safety-oriented screening context where missing malicious messages is more costly than issuing a limited number of false alarms(H B & H L, 2025; Kaur & Kaur, 2023; Qin & Zhang, 2024).

Quantitatively, the main summary table over split ratios and models (see Figure 4) shows that at split = 0.5, BERT Frozen achieves an accuracy of 0.9179, precision 0.8994, recall 0.9829, and F1-score 0.9393, with training time of about 134.3 seconds. The high recall (0.9829) and lower precision (0.8994) indicate that BERT Frozen tends to classify most phishing messages correctly but produces a non-negligible number of false positives, which may correspond to aggressive detection behavior suitable for safety-critical phishing screening but with some cost in specificity. Under the same split, BERT FineTune attains an accuracy of 0.9883, precision 0.9918, recall 0.9901, and F1-score 0.9910, with training time approximately doubling to 270.6 seconds. This configuration yields both high recall and high precision, indicating that fine-tuning IndoBERT substantially improves discrimination between phishing and non-phishing SMS and reduces both false positives and false negatives, in line with prior evidence that task-specific fine-tuning unlocks substantial performance gains for Indonesian classification tasks(Imaduddin et al., 2023).

The CNN–BERT Frozen model at split = 0.5 attains accuracy of 0.9755, precision 0.9737, recall 0.9889, and F1-score 0.9812, with a training time of 136.3 seconds, comparable to BERT Frozen. Relative to BERT Frozen, CNN–BERT Frozen improves accuracy by about 5.8 percentage points and F1-score by about 4.2 points, mainly through simultaneous gains in precision (0.9737 vs 0.8994) and slightly improved recall (0.9889 vs 0.9829), showing that the CNN head effectively refines the decision boundary in the fixed feature space. CNN-BERT FineTune at split = 0.5 reaches accuracy of 0.9877, precision 0.9903, recall 0.9907, and F1-score 0.9905, with training time of about 272.0 seconds, closely matching BERT FineTune in accuracy and F1 but with slight differences in precision and recall balance. In particular, BERT FineTune shows marginally higher precision and F1, while CNN-BERT FineTune shows slightly higher recall, indicating that the convolutional pooling may favor a slightly more recall-oriented decision surface, though the differences are small(Soni et al., 2023).

The table also reports BERT Frozen performance for split = 0.4, with accuracy 0.9346, precision 0.9231, recall 0.9806, and F1-score 0.9510, at a training time of about 135.1 seconds. Compared to BERT Frozen at 0.5, reducing the training proportion to 0.4 slightly increases precision (0.9231 vs 0.8994) and F1-score (0.9510 vs 0.9393), while recall remains high (0.9806 vs 0.9829), suggesting that this model is relatively robust to moderate reductions in training data and that its linear classifier head can generalize well in this size regime. For smaller splits down to 0.1, confusion matrices and classification reports indicate that both BERT-based and CNN–BERT-based models maintain high recall and balanced precision, although absolute metrics are slightly lower than at 0.5 splits. For instance, at split 0.1, BERT FineTune yields accuracy 0.9887 with phishing-class precision 0.9932 and recall 0.9893, while CNN-BERT Frozen gives accuracy 0.9774 with precision 0.9788 and recall 0.9864, showing that fine-tuned models retain near-Maximal performance even under substantially reduced training data. The stable metrics across splits imply that the models capture robust patterns characteristic of phishing SMS such as lexical cues, call-to-action phrasing, and numeric patterns rather than overfitting idiosyncratic messages from larger training sets, echoing findings from other Indonesian SMS spam and phishing detection studies(Rahmapuri et al., 2025).

```
Paired t-test:
t = 0.6931248983446825
p = 0.5263730611322903
```

Figure 5. Result Paired Test

Beyond single splits, the 5-fold cross-validation experiments provide a more granular view of generalization behavior (Preeti & Sharma, 2024). For BERT FineTune and CNN–BERT FineTune, F1-scores across folds are collected into arrays and compared via a paired t-test (see Figure 5). The resulting t-statistic is 0.6931 with p-value 0.5264, indicating no statistically significant difference in F1-score between the two fine-tuned models at conventional significance

thresholds when evaluated across folds. Thus, although CNN–BERT FineTune and BERT FineTune are numerically very close on the 0.5 split, the cross-validation analysis confirms that their performance is statistically indistinguishable in terms of average F1-scores under the present experimental conditions. Both models achieve very high F1-scores across folds, confirming that fine-tuned IndoBERT architectures provide strong, stable generalization for phishing detection on this dataset(Koto et al., 2020).

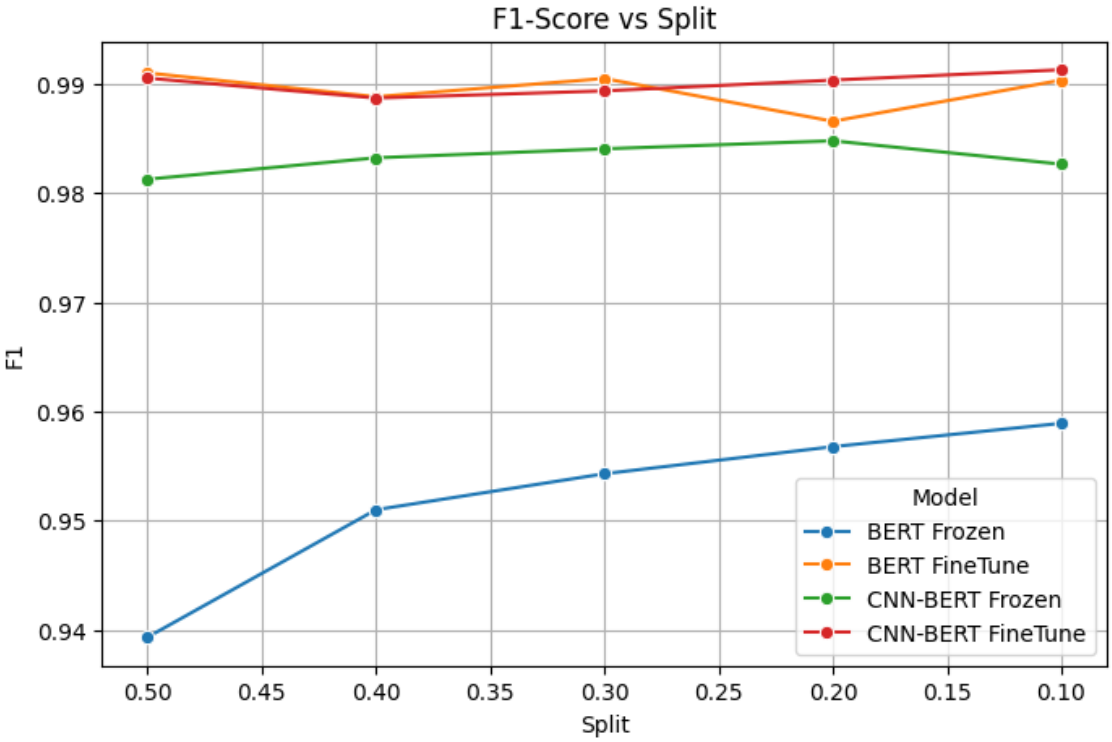

Figure 6. F1 Score vs Split

From the perspective of phishing detection, accuracy summarizes overall correctness, but precision, recall, and F1-score provide more task-relevant insight. At 0.5 splits, BERT Frozen's recall of 0.9829 indicates that nearly all phishing messages are correctly flagged, but precision of 0.8994 implies that approximately 10 % of messages predicted as phishing are actually legitimate, which may lead to user inconvenience or alert fatigue. CNN–BERT Frozen, by contrast, raises precision to 0.9737 while preserving high recall (0.9889), reducing false positives markedly while still capturing almost all phishing cases. The fine-tuned models move into a regime where both precision and recall exceed 0.99 for BERT FineTune and are very close to 0.99 for CNN-BERT FineTune, meaning that both mislabelled phishing and mislabelled legitimate SMS are very rare in the evaluated test sets. Across splits and folds, F1-scores for fine-tuned models consistently exceed 0.98 (see Figure 6), showing that these architectures achieve highly reliable performance when both detection sensitivity and specificity are considered jointly, comparable to the best-performing deep learning and transformer-based approaches reported for SMS spam and phishing detection(Rameem Zahra et al., 2022).

Synthesizing the figure and table results, the hybrid CNN–BERT architecture displays distinct behavior relative to BERT-only models that is consistent with its design. When IndoBERT is frozen, CNN–BERT substantially improves over BERT Frozen in accuracy, precision, recall, and F1, indicating that the convolutional layer and adaptive max pooling leverage token-level contextual embeddings to derive more discriminative global features than the simple pooled representation. This is reflected in the confusion matrices, where CNN–BERT Frozen reduces both false positives and false negatives compared with BERT Frozen, leading to a more concentrated diagonal structure. When IndoBERT is fine-tuned, however, the marginal benefit of adding a CNN head becomes small: CNN-BERT FineTune and BERT FineTune achieve nearly identical performance, with cross-validation showing no statistically significant

difference in F1, and training curves exhibiting similar convergence rates and stability. Thus, fine-tuning IndoBERT appears to absorb most of the task-specific adaptation, diminishing the incremental impact of the CNN head relative to the already strong contextual representations provided by the fully trainable transformer, a trend that is consistent with reports that fine-tuned transformer baselines are often hard to surpass with more complex hybrids(Pratama & Rjito, 2021).

Freezing BERT affects learning by constraining the model to operate in a fixed semantic space, which limits the capacity of the classifier to adapt to dataset-specific cue distributions and domain-specific vocabulary. Under this constraint, the linear BERT-only classifier still achieves high recall but shows reduced precision, suggesting that while many phishing cues are already encoded in IndoBERT, the static representation space lacks some discriminative features necessary to separate hard borderline cases. The CNN–BERT Frozen model alleviates this limitation by applying a learnable local pattern detector across token embeddings; the convolutional filters and max pooling yield task-specific features that emphasize local n-grams and co-occurrence structures salient for phishing detection, such as repeated numeric patterns, URLs, or imperative phrases. Consequently, CNN–BERT Frozen achieves a more balanced precision–recall profile than BERT Frozen while operating under the same frozen backbone and similar training time(Soni et al., 2023).

When BERT is fine-tuned, the backbone itself reshapes its embedding space to align with phishing versus non-phishing distinctions, enabling both BERT FineTune and CNN–BERT FineTune to reach very high performance. In this regime, the transformer can internalize local patterns and higher-order regularities throughout its multi-layer self-attention structure, and the marginal representational gain from adding a CNN head is small compared with the effect of full fine-tuning. The close alignment of training and validation curves for both fine-tuned models across splits indicates that fine-tuned IndoBERT generalizes well without substantial overfitting over five epochs, even on smaller training subsets. The cross-validated t-test results underscore that, averaged across folds, the hybrid CNN–BERT and the BERT-only classifier exhibit comparable F1-scores, highlighting that model capacity, data size, and pretraining alignment jointly saturate performance for this dataset.

Although the statistical test indicates that the difference between the fine-tuned models is not significant in terms of F1-score, this should not be read as evidence that the architectures are identical; rather, it suggests that performance has approached a saturation zone under the current dataset size and training regime. This is consistent with recent findings that fine-tuning strategies and model freezing can strongly influence generalization, but further gains may require adversarial training, regularization, or broader data diversity rather than architectural complexity alone.

Generalization behavior across splits and folds is consistent with the characteristics of the dataset and architecture. As the training fraction decreases from 0.5 to 0.1, accuracy, precision, recall, and F1 remain high, with only modest declines, especially for fine-tuned variants, indicating that IndoBERT's pre-training on Indonesian text provides robust prior knowledge that compensates for reduced supervision. The confusion matrices across splits show that predictions remain well concentrated on the diagonal, with only slight increases in misclassifications at lower splits, consistent with the expected effect of reduced data. In k-fold cross-validation, variance in F1-scores across folds is relatively small, and both BERT FineTune and CNN–BERT FineTune exhibit uniformly strong performance, further confirming stable generalization across different train–test partitions. Together, these observations suggest that the combination of IndoBERT pretraining, appropriate preprocessing, and modest fine-tuning is sufficient to achieve near-ceiling detection performance on this SMS phishing dataset under multiple data splitting strategies.

4. Conclusion

The experimental results demonstrate that IndoBERT-based classifiers with stratified train–test splits and 5-fold cross-validation achieve high performance for Indonesian SMS phishing detection across multiple data regimes. Using IndoBERT as a frozen feature extractor with a linear head yields high recall but relatively lower precision, whereas augmenting the frozen

backbone with a CNN head improves all metrics, indicating that convolutional pooling over token-level embeddings enhances discrimination under a frozen backbone. When IndoBERT is fine-tuned, both BERT-only and CNN–BERT architectures converge stably and attain very high accuracy, precision, recall, and F1-scores, with cross-validation showing no statistically significant difference in F1 between the two fine-tuned variants. Across all configurations, models generalize well across splits and folds, and the combination of IndoBERT contextual representations, task-specific fine-tuning, and CNN-based local feature extraction in the frozen case supports highly reliable phishing detection in this domain.

Beyond the strong performance metrics, the principal scientific contribution of this study lies in demonstrating that a domain-adapted IndoBERT-CNN framework can effectively separate contextual representation learning from local pattern extraction in Indonesian phishing messages, thereby clarifying when CNN augmentation is beneficial and when full fine-tuning already saturates performance. This is methodologically relevant for phishing detection research because recent literature increasingly emphasizes the value of hybrid architectures, dataset-specific adaptation, and the need to move beyond accuracy-only reporting toward more interpretable and deployment-oriented evaluation. At the same time, the present results should be interpreted within several limitations: the corpus is private and domain-specific, which may introduce sampling bias and restrict external reproducibility; the short-text nature of SMS limits the amount of contextual evidence available for classification; and the model remains a black-box system without explicit explanation of individual predictions, which constrains trust and forensic usability in operational settings. These constraints imply that future work should prioritize three concrete directions: adversarially robust training and drift-aware re-evaluation to improve resistance to evolving phishing tactics, cross-platform transfer experiments to test generalization from SMS to social-media and messaging environments, and explainable AI methods such as attention-based rationales or post-hoc feature attribution to make real-time phishing screening more transparent and actionable.

Taken together, the results provide a practically relevant reference point for Indonesian short-message phishing detection by showing that a carefully fine-tuned IndoBERT backbone can already achieve near-saturation performance, while CNN augmentation is particularly beneficial when transformer parameters must remain frozen, for example in resource-constrained or privacy-sensitive deployments. For security practitioners, these findings suggest that transformer-based detectors can be embedded as back-end filters in SMS gateways or social-media moderation pipelines to prioritize high-risk messages for secondary checks, provided that organizations regularly assess dataset drift and recalibrate decision thresholds to manage the precision–recall trade-off in their specific risk context. Building on this work, future research can profitably combine IndoBERT–CNN architectures with explainability techniques and cross-platform evaluation protocols, thereby linking high-performance detection with model transparency and robustness across SMS, chat, and social-media environments.

References

Abdillah, R., & Insani, F. (2025). SMS Phishing Detection Model with Hyperparameter Optimization in Machine Learning. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi*, *11*(1), 35. https://doi.org/10.24014/coreit.v11i1.35547

Adel Al-Zebari. (2025). Deep Learning Hybrid Approach for Accurate SMS Spam Identification. *Journal of Information Systems Engineering and Management*, *10*(10s), 619–635. https://doi.org/10.52783/jisem.v10i10s.1426

Alhuzali, A., Alloqmani, A., Aljabri, M., & Alharbi, F. (2025). In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning and Deep Learning Models Across Multiple Datasets. *Applied Sciences*, *15*(6), 3396. https://doi.org/10.3390/app15063396

Alotaibi, A. (2026). Self-consistency and Graph-based Filtering to Enhance Synthetic Arabic SMS Generation for Smishing Detection. *Journal of Applied Data Sciences*, *7*(1), 357–383. https://doi.org/10.47738/jads.v7i1.1033

Biswas, D., Byun, T.-Y., & Gil, J.-M. (2025). Research Paper Classification Based on CNN and BiLSTM Models Utilizing Word Embedding Methods. *Human-Centric Computing and Information Sciences*, *15*(24).

Brindha, R., Nandagopal, S., Azath, H., Sathana, V., Prasad Joshi, G., & Won Kim, S. (2023). Intelligent Deep Learning Based Cybersecurity Phishing Email Detection and Classification. *Computers, Materials & Continua*, *74*(3), 5901–5914. https://doi.org/10.32604/cmc.2023.030784

Christian, W., Adamlu, D., Yu, A., & Suhartono, D. (2025). *Leveraging IndoBERT and DistilBERT for Indonesian Emotion Classification in E-Commerce Reviews* (arXiv:2509.14611). arXiv. https://doi.org/10.48550/arXiv.2509.14611

Fu, B., Wang, Y., & Feng, T. (2024). CT-GCN+: A high-performance cryptocurrency transaction graph convolutional model for phishing node classification. *Cybersecurity*, *7*(1), 3. https://doi.org/10.1186/s42400-023-00194-5

Ghourabi, A. (2021). SM-Detector: A security model based on BERT to detect SMiShing messages in mobile environments. *Concurrency and Computation: Practice and Experience*, *33*(24), e6452. https://doi.org/10.1002/cpe.6452

Gupta, B. B., Gaurav, A., Arya, V., Attar, R. W., Bansal, S., Alhomoud, A., & Chui, K. T. (2024). Advanced BERT and CNN-Based Computational Model for Phishing Detection in Enterprise Systems. *Computer Modeling in Engineering & Sciences*, *141*(3), 2165–2183. https://doi.org/10.32604/cmes.2024.056473

H B, G., & H L, G. (2025). Detection of Phishing Activities Using Deep Learning Approaches. *2025 17th International Conference on COMmunication Systems and NETworks (COMSNETS)*, 808–810. https://doi.org/10.1109/COMSNETS63942.2025.10885614

Hassan, N. H., & Fakharudin, A. S. (2023). Web Phishing Classification Model using Artificial Neural Network and Deep Learning Neural Network. *International Journal of Advanced Computer Science and Applications*, *14*(7). https://doi.org/10.14569/IJACSA.2023.0140759

Hassan, Z. L., Sani, N. F. M., Abdullah, M. D. H., & Mustapha, N. (2026). Transformer Attention-Driven Concept Extraction for Efficient Smishing Detection. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62028760

Imaduddin, H., A'la, F. Y., & Nugroho, Y. S. (2023). Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach. *International Journal of Advanced Computer Science and Applications*, *14*(8). https://doi.org/10.14569/IJACSA.2023.0140813

Johari, M. F., Chiew, K. L., Hosen, A. R., Yong, K. S. C., Khan, A. S., Abbasi, I. A., & Grzonka, D. (2025). Key insights into recommended SMS spam detection datasets. *Scientific Reports*, *15*(1), 8162. https://doi.org/10.1038/s41598-025-92223-1

Kaur, K., & Kaur, P. (2023). BERT-CNN: Improving BERT for Requirements Classification using CNN. *Procedia Computer Science*, *218*, 2604–2611. https://doi.org/10.1016/j.procs.2023.01.234

Kim, Y., Ohn, I., & Kim, D. (2021). Fast convergence rates of deep neural networks for classification. *Neural Networks*, *138*, 179–197. https://doi.org/10.1016/j.neunet.2021.02.012

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Kurnianda, P. R., Zakiyyah, A. Y., Ibrahim, M. A., & Pradana, R. C. (2025). *Feature-Driven Ensemble Learning for Robust Phishing Website Classification* (No. 11). ICIC International 学会. https://doi.org/10.24507/icicel.19.11.1175

Lamas Piñeiro, J., & Wong Portillo, L. (2022). Web architecture for URL-based phishing detection based on Random Forest, Classification Trees, and Support Vector Machine. *Inteligencia Artificial*, *25*(69), 107–121. https://doi.org/10.4114/intartif.vol25iss69pp107-121

Maneriker, P., Stokes, J. W., Lazo, E. G., Carutasu, D., Tajaddodianfar, F., & Gururajan, A. (2021). URLTran: Improving Phishing URL Detection Using Transformers. *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*, 197–204. https://doi.org/10.1109/MILCOM52596.2021.9653028

Naswir, A. F., Zakaria, L. Q., & Saad, S. (2022). Determining the Best Email and Human Behavior Features on Phishing Email Classification. *International Journal of Advanced Computer Science and Applications*, *13*(8). https://doi.org/10.14569/IJACSA.2022.0130821

Otieno, D. O., Abri, F., Namin, A. S., & Jones, K. S. (2023). Detecting Phishing URLs using the BERT Transformer Model. *2023 IEEE International Conference on Big Data (BigData)*, 2483–2492. https://doi.org/10.1109/BigData59044.2023.10386782

Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4996–5001. https://doi.org/10.18653/v1/P19-1493

Pratama, T., & Rjito, S. (2021). IndoXLNet: Pre-Trained Language Model for Bahasa Indonesia. *International Journal of Engineering Trends and Technology*, *70*(5), 367–381. https://doi.org/10.14445/22315381/IJETT-V70I5P240

Preeti, P., & Sharma, P. (2024). Enhancing phishing URL detection through comprehensive feature selection: A comparative analysis across diverse datasets. *Indonesian Journal of Electrical Engineering and Computer Science*, *36*(2), 1182. https://doi.org/10.11591/ijeecs.v36.i2.pp1182-1188

Qin, S., & Zhang, M. (2024). Boosting generalization of fine-tuning BERT for fake news detection. *Information Processing & Management*, *61*(4), 103745. https://doi.org/10.1016/j.ipm.2024.103745

Rahmapuri, A., Sholikah, R. W., & Firdausi, H. (2025). SMSHIELD: A Real-Time Smishing Detection in Mobile Messaging Systems using Deep Learning Approach. *2025 IEEE 11th Information Technology International Seminar (ITIS)*, 1–6. https://doi.org/10.1109/ITIS67966.2025.11309139

Ramdhan, D., Lucky, H., Kemala, P., & Chowanda, A. (2022). *Short Message Service (SMS) Spam Filtering Using Deep Learning in Bahasa Indonesia* (No. 10). ICIC International 学会. https://doi.org/10.24507/icicelb.13.10.1093

Rameem Zahra, S., Ahsan Chishti, M., Iqbal Baba, A., & Wu, F. (2022). Detecting Covid-19 chaos driven phishing/malicious URL attacks by a fuzzy logic and data mining based intelligence system. *Egyptian Informatics Journal*, *23*(2), 197–214. https://doi.org/10.1016/j.eij.2021.12.003

Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). Next-Generation Spam Filtering: Comparative Fine-Tuning of LLMs, NLPs, and CNN Models for Email Spam Classification. *Electronics*, *13*(11), 2034. https://doi.org/10.3390/electronics13112034

Saidat, M. R. A., Yerima, S. Y., & Shaalan, K. (2024). Advancements of SMS Spam Detection: A Comprehensive Survey of NLP and ML Techniques. *Procedia Computer Science*, *244*, 248–259. https://doi.org/10.1016/j.procs.2024.10.198

Sivakumar, M., Parthasarathy, S., & Padmapriya, T. (2024). Trade-off between training and testing ratio in machine learning for medical image processing. *PeerJ Computer Science*, *10*, e2245. https://doi.org/10.7717/peerj-cs.2245

Soni, S., Chouhan, S. S., & Rathore, S. S. (2023). TextConvoNet: A convolutional neural network based architecture for text classification. *Applied Intelligence*, *53*(11), 14249–14268. https://doi.org/10.1007/s10489-022-04221-9

Syafitri, W., Pane, E. P., & Purwanto, E. (2026). Enhanced social media phishing detection model using LSTM and BERT. *Science, Technology, and Communication Journal*, *6*(2), 165–174. https://doi.org/10.59190/stc.v6i2.360

Tanbhir, G., Shahriyar, Md. F., Shahed, K., Chy, A. M. R., & Adnan, M. A. (2024). Hybrid Machine Learning Model for Detecting Bangla Smishing Text Using BERT and Character-Level CNN. *2024 13th International Conference on Electrical and Computer Engineering (ICECE)*, 57–62. https://doi.org/10.1109/ICECE64886.2024.11024872

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601. https://doi.org/10.18653/v1/P19-1452

Uddin, M. A., Mahiuddin, M., & Sarker, I. H. (2026). An explainable transformer-based model for phishing email detection: A large language model approach. *Computer Networks*, *277*, 112061. https://doi.org/10.1016/j.comnet.2026.112061

Ulfath, R. E., Alqahtani, H., Hammoudeh, M., & Sarker, I. H. (2021). Hybrid CNN-GRU Framework with Integrated Pre-trained Language Transformer for SMS Phishing Detection. *The 5th International Conference on Future Networks & Distributed Systems*, 244–251. https://doi.org/10.1145/3508072.3508109

Wei, Y., Nakayama, M., & Sekiya, Y. (2025). Enhancing Generalization in Phishing URL Detection via a Fine-Tuned BERT-Based Multimodal Approach. *IEEE Access*, *13*, 131197–131216. https://doi.org/10.1109/ACCESS.2025.3591843

Wicaksana, H. S., Ependi, U., & Muzakir, A. (2026). URL-Based Phishing Detection Using a BERT-LSTM Model. *Journal of Information Systems and Informatics*, *8*(1), 1344–1367. https://doi.org/10.63158/journalisi.v8i1.1543

Yasinta Roesmiatun, P., & Zahra, A. (2025). Enhancing detection of zero-day phishing email attacks in the Indonesian language using deep learning algorithms. *Bulletin of Electrical Engineering and Informatics*, *14*(1), 505–512. https://doi.org/10.11591/eei.v14i1.8759

Yuan, Y., Hao, Q., Apruzzese, G., Conti, M., & Wang, G. (2024). "Are Adversarial Phishing Webpages a Threat in Reality?" Understanding the Users' Perception of Adversarial Webpages. *Proceedings of the ACM Web Conference 2024*, 1712–1723. https://doi.org/10.1145/3589334.3645502