IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

International Journal of Safety and Security Engineering
Vol. 15, No. 12, December, 2025, pp. 2511-2518
Journal homepage: http://iieta.org/journals/ijsse

A Hybrid Attention-Based Deep Learning Architecture for Phishing Detection

Akshat Uike[1*], Prakash Prasad[2]

[1] Department of Electronics and Computer Science, Rashtrasant Tukadoji Maharaj Nagpur University (RTMNU), Nagpur 440033, India
[2] Priyadarshini College of Engineering, Priyadarshini Campus, Nagpur 440039, India

Corresponding Author Email: akshatuike12@gmail.com

https://doi.org/10.18280/ijsse.151208

Received: 10 October 2025
Revised: 10 December 2025
Accepted: 21 December 2025
Available online: 31 December 2025

Keywords:
phishing detection, self-attention, sparse autoencoder, Convolutional Bidirectional Long Short-Term Memory, transformer-based deep learning, hybrid architecture

ABSTRACT

Phishing methods are continually advancing, capitalizing on both human and systemic vulnerabilities across all digital channels. Conventional rule-based and machine-learning approaches are hindered by their dependence on manually produced features and restricted generalization capabilities. This paper presents a lightweight yet robust hybrid deep-learning architecture that integrates a Gated Self-Attention Sparse Autoencoder (GAttn-SAE) with an Attention-Optimized Convolutional Bidirectional Long Short-Term Memory (AConv-BiLSTM) classifier to address these shortcomings. The GAttn-SAE module selectively amplifies distinguishing phishing characteristics using self-attention-based representation learning, whereas the AConv-BiLSTM identifies spatial patterns and long-range sequential relationships in URLs, headers, and metadata. The suggested architecture achieves good performance with much reduced complexity compared to computationally intensive transformer-based models, rendering it appropriate for resource-limited and real-time cybersecurity settings. An 80:20 train-test split of publicly available phishing datasets (approximately 2,300 samples) shows that the evaluation method has an accuracy of 98.4, precision of 97.8, recall of 98.9, F1-score of 98.3, and ROC-AUC of 99.1, which is higher than the performance of conventional machine learning (ML) / deep learning (DL) baselines and comparable to transformer-level performance.

1. INTRODUCTION

Phishing has emerged as one of the most persistent and rapidly evolving threats in the field of cybersecurity, exploiting human psychology and technological vulnerabilities to obtain sensitive information such as login credentials, financial data, and personal identifiers [1]. Attackers increasingly employ social engineering techniques, including deceptive emails, forged websites, and manipulated hyperlinks to impersonate legitimate entities and deceive users into revealing confidential details [2]. The widespread adoption of digital services in sectors such as E-banking, E-commerce, and E-governance has further expanded the attack surface, intensifying the frequency and complexity of phishing incidents worldwide [3]. Conventional strategies such as blacklist-based filters and heuristic rules struggle to address emerging phishing threats, as they were designed for known patterns and lack adaptability [4].

To address these limitations, the research community has progressively shifted toward machine learning (ML) and deep learning (DL) methodologies for phishing detection. ML-based models improve adaptability compared to heuristic systems but still face challenges in feature engineering and false classification rates when dealing with diverse phishing data. In contrast, DL frameworks, especially those leveraging convolutional neural networks (CNNs), recurrent neural networks (RNNs), and self-attention architectures, have demonstrated exceptional ability to automatically learn complex representations from high-dimensional data, enabling them to identify subtle behavioural and structural cues associated with phishing activities [5].

Nonetheless, despite these advancements, numerous research gaps remain. Existing DL models often lack the capability to simultaneously capture both spatial dependencies (relationships between feature patterns) and sequential dependencies (temporal or contextual relationships in URLs and content). Moreover, many models suffer from overfitting and reduced interpretability, limiting their robustness against rapidly evolving phishing strategies.

The present research introduces a hybrid DL architecture that integrates Gated Self-Attention Sparse Autoencoder (GAttn-SAE) with Attention-Optimized Convolutional Bidirectional Long Short-Term Memory (AConv-BiLSTM) networks to overcome these challenges. The GAttn-SAE module efficiently extracts critical phishing-related features using self-attention mechanisms, while the AConv-BiLSTM model combines Conv1D and BiLSTM layers to analyze both spatial and sequential phishing patterns. This dual-module approach enhances feature discrimination, reduces false positives, and improves generalization against novel phishing variants. Comparative experimental evaluation shows that the proposed model significantly outperforms existing ML-based and DL-based approaches in terms of accuracy, precision, recall, and ROC-AUC. Thereby establishing a robust and adaptive defense mechanism against modern phishing attacks. The key contributions of this work can be explained in three parts. First, we introduce a new hybrid phishing-detection architecture that combines both a GAttn-SAE that is used to depict discriminative features and an AConv-BiLSTM classifier to simultaneously represent spatial URL patterns and long-range sequential dependencies. Second, the suggested framework has high detection rates with lower computational complexity than transformer-based models, which is measured by model depth, the number of parameters, and training convergence properties, and is applicable to real-time and resource-limited cybersecurity settings. Third, we perform an extensive empirical analysis on publicly accessible phishing benchmarks, we compare the proposed model to classical machine-learning, deep-learning baselines, and state-of-the-art transformer architectures, and we compare them on standard measures, such as accuracy, precision, recall, F1-score, and ROC-AUC.

The remainder of this paper is structured as follows: Section II reviews related literature and existing phishing detection methodologies; Section III details the proposed hybrid architecture and its operational framework; Section IV presents the experimental setup, and evaluation metrics; Section V discusses the results and comparative analysis highlighting the model's performance and novelty and finally, Section VI concludes the study with future research directions.

2. RELATED WORK

The problem of phishing detection is already thoroughly explored in the literature on cybersecurity; the existing methods of phishing detection have varied mainly based on the modality of input that is utilized and the paradigm that the learning follows. The previous work may be roughly divided into URL, email-text, and multi-modal, or hybrid, phishing detection models. One of the oldest and best studied phishing detection strategies is URL-based, which targets the lexical, structural, and statistical characteristics of URLs (length, entropy, distribution of tokens, special character sequences, and so on). Early machine-learning and client-side-based systems were based on handcrafted features and classifiers like Support Vector Machine (SVM), Decision Trees, and Random Forests with reasonable accuracy on known phishing examples [6-8]. Nevertheless, these models demonstrate a low generalization ability when facing an obfuscated or zero-day phishing URL because they require manual feature engineering. To solve these issues, deep-learning models, including CNNs and LSTMs, have been proposed to automatically extract representations out of a sequence of URLs [9, 10]. Although such methods enhance the performance of detection, most of them either focus on local structural patterns or sequential dependencies individually, and they do not have explicit attention-related mechanisms to prioritize discriminative phishing features.

Hybrid and representation-learning-based models have also been developed to improve on feature extraction and robustness over and above a single-model architecture. Yang et al. introduced a representation learning framework based on autoencoders and a multilayer perceptron (LLM-AE-MP) to detect web attacks, and they showed that latent feature learning was effective in phishing pattern learning [11]. Sarasjati et al. [12] proposed a hybrid phishing detection framework that integrates LASSO-based feature selection with a weighted bootstrap sampling enhanced Random Forest classifier, achieving improved handling of high-dimensional feature spaces and dataset outliers in phishing detection. Though successful, these methods do not often explicitly have a mechanism to model both spatial URL structure and long-range sequential dependencies within a single framework based on attention processes. Phishing detection methods that use email-text are based on the natural language processing methods of analyzing email bodies, their subject lines, and semantic context. IT architectures, especially BERT, introduced the area of phishing email classification to a new level as it could learn a context-sensitive representation using self-attention [13]. This was followed up by research that demonstrates that fine-tuned transformer models such as BERT and RoBERTa are generally superior in phishing and spam detection problems compared to traditional machine-learning and shallow deep-learning bases [14-17]. Nevertheless, these models have high performance despite the fact that they need large-scale pretraining, significant processing power, and inference time, preventing their use in real-time or resource-constrained settings. Moreover, they are less suitable for URL- and metadata-based phishing traffic since they are more oriented toward textual information. Multi-modal phishing detection systems combine heterogeneous inputs, e.g., URLs, email text, HTML content, and metadata, to make the system more resilient against various attack vectors. Recent work that has utilized transformer-based architecture as well as large-language-model-based architecture has shown excellent performance by combining various modalities and contextual indicators [18-22]. But most of these systems have a large computational overhead, rely on transformer backbones, or have no lightweight feature-selection mechanisms, limiting their scalability and deployment in real-world cybersecurity systems.

Research gap: In all classes, there is still an apparent gap in bridging to a phishing detection framework, which (i) focuses on discriminative learning of feature inputs URL, header, and metadata, (ii) embraces spatial and sequence dependence with lightweight attention-based mechanisms, and (iii) with no-excessive computational cost. The suggested GAttn-SAE + AConv-BiLSTM model explicitly closes this divide by combining gated self-attention-based representation of features with convolutional and bidirectional recurrent models into one cohesive model, which is resource-efficient and can be used in real-time detection of phishing attacks.

3. METHODOLOGY

In the design of the proposed hybrid model, the philosophy is based on the successful experience of transformer architectures like BERT and Denseformer that used self-attention to detect long-range dependencies in textual and sequential data. Nevertheless, transformer models are less applicable in real-time deployment in cybersecurity, as they usually need enormous annotated datasets and computation resources to get trained and be highly accurate in the detection of phishing and fraud. The proposed framework incorporates lightweight attention elements in its design to balance between performance and efficiency. GAttn-SAE is made with the same representational power as transformer encoders on the feature extraction task, and AConv-BiLSTM is provided that combines both convolutional and bidirectional recurrent processing to model hierarchical attentive behavior. The design is interpretable and efficient, like traditional neural models, but with more sophisticated contextual correlations than transformer models. The proposed framework implements DL methods that improve both extractive capabilities for features and classification performance and adaptive behavior for new phishing techniques. GAttn-SAE and AConv-BiLSTM constitute the model's two components dedicated to feature representation learning and phishing classification tasks, respectively. The workflow step processes data by repairing missing values and removing duplicates while applying representations to domain names, email metadata, and URL patterns, as shown in Figure 1. Standardization normalization falls under the normalization techniques used to normalize numerical features by making their values have a mean of zero, while the variance becomes one. The dataset distribution takes place into distinct training and testing parts for model assessment purposes. The notation used throughout the proposed GAttn-SAE + AConv-BiLSTM framework is summarized in Table 1.

Table 1. Notation used in the proposed model

Symbol	Meaning
(X)	Input feature vector
(Z)	Latent representation
(Q, K, V)	Attention matrices
(α)	Attention weights
(h_{bi})	BiLSTM output
$(\hat{y})$	Predicted label
(L_{MSE})	Autoencoder loss
(L_{BCE})	Classification loss

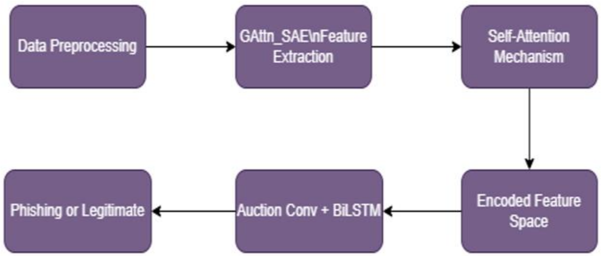

Figure 1. Phishing detection model using Gated Self-Attention Sparse Autoencoder (GAttn-SAE) and Attention-Optimized Convolutional Bidirectional Long Short-Term Memory (AConv-BiLSTM)

The GAttn-SAE module serves to extract important data patterns from unprocessed phishing data and remove unnecessary information. The initial portion of the encoder contains 128 ReLU units in its first layer, followed by a second layer equipped with 64 units to transform input features into latent space vectors. Development of the encoder included the self-attention mechanism that focuses on significant features by computing attention scores between various input attributes.

A GAttn-SAE together with AConv-BiLSTM constitute the proposed phishing detection model, which enables high precision detection while being adaptable, as shown in Figure 2. The model starts by obtaining relevant phishing features from URL attributes and email header contents, along with metadata elements in its initial input phase. With ReLU activation, the GAttn-SAE component contains two encoder layers of 128 and 64 neurons to decrease dimensions and retain essential information. The self-attention component evaluates important features that lead to a 64-dimensional output encoding representation. AConv-BiLSTM uses Conv1D layers having 64 filters to discover localized phishing patterns in addition to BiLSTM layers containing 64 forward and 64 backward units that can learn sequential dependencies inside phishing datasets. The processed features pass through two fully connected layers with ReLU activation, which contain 64 neurons, then 32 neurons to improve the classification steps. The output predictions of the network occur through its final sigmoid-activated layer to determine phishing or legitimate inputs. By combining hybrid attention-based encoding with CNN-LSTM classification, the system achieves optimal feature selection and sequential learning and better accuracy results. This architecture employs self-attention for dynamic pattern highlighting. It uses Conv1D to extract spatial correlation and BiLSTM to capture sequential dependencies. Together, these components make the model resilient against new phishing techniques.

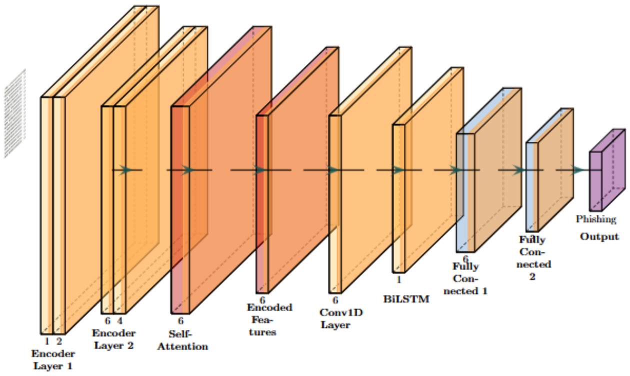

Figure 2. Gated Self-Attention Sparse Autoencoder (GAttn-SAE) + Attention-Optimized Convolutional Bidirectional Long Short-Term Memory (AConv-BiLSTM): A hybrid deep learning (DL) model for phishing detection

The input features begin with the encoder network containing two fully connected layers.

$$h_1 = \mathrm{Re}LU(W_1 X + b_1) \tag{1}$$

$$h_2 = \mathrm{Re}LU(W_2 h_1 + b_2) \tag{2}$$

The weight matrices are denoted as W_1, W_2, b_1, b_2 are the bias vectors. The input feature vector is designated as X.

The model contains three principal input layers named Q, K, V, which correspond to Query, Key, and Value, while d_k defines the dimensions of the key. The system gives greater weight to essential indicators of phishing activity, specifically including domain entropy, URL length, and special character detection. The decoder maintains important data points by using two dense layers to regenerate the input information from the latent representation.

$$A = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{3}$$

Through the decoder process, the original input gets decoded back from its latent representation form.

$$X = \sigma(W_d h_2 + b_d) \tag{4}$$

During the optimization process, Mean Squared Error (MSE) minimizes the reconstruction error in the autoencoder.

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i)^2 \tag{5}$$

This layer uses its capability to extract local phishing patterns from features that have undergone encoding.

$$y[i] = \sum_{k=1}^{K} W_K . X[i+k-1] \tag{6}$$

The AConv-BiLSTM classifier utilizes convolutional layers and bidirectional LSTMs to process phishing features that were extracted from the phishing details. The initial phase of the model utilizes a 1D CNN that works with 64 filters of size 3 to analyze local patterns present in phishing URLs as well as webpage metadata. The BiLSTM technology enables sequential dependency detection by processing data forward and backward while receiving its input from the convolutional layer outputs.

$$h_t^{forward} = f(W_{ih}x_t + w_{hh}h_{t-1} + b_h) \tag{7}$$

$$h_t^{backward} = f(W_{ih}x_t + w_{hh}h_{t+1} + b_h) \tag{8}$$

The BiLSTM-generated outputs go through a fully connected layer to perform phishing classification.

$$y = \sigma(W_0 h + b_0) \tag{9}$$

The processing in both directions allows the system to detect patterns in phishing contexts, which include URL component order or email header sequences. A sigmoid-activated fully connected layer receives inputs from concatenated BiLSTM outputs to generate binary classifications that identify phishing attempts through values approaching 1. The classification task being binary makes Binary Cross-Entropy (BCE) Loss the appropriate choice:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(y_i) + (1 + y_i)\log(1 - y_i)] \tag{10}$$

The formula represents the prediction of $\hat{y}_i$ against the actual value y_i. The optimization process relies on the Adam optimizer to reach convergence while using a learning rate of 0.001. The training process consists of 50 epochs with early stopping implemented to avoid overfitting, and it operates at a batch size of 32. Performance evaluation of phishing detection utilizes Accuracy together with Precision, Recall, F1-score, and ROC-AUC for effectiveness assessment during the model evaluation stage.

4. EXPERIMENTAL SETUP

The publicly available phishing benchmark datasets were used to test the proposed model. These datasets were collected from the UCI and Kaggle repositories and consisted of approximately 2,300 instances. The data were nearly class-balanced, with phishing and legitimate samples present in approximately equal numbers. An example is represented in terms of lexical features of URLs, metadata at the head, and statistical information based on the URL structure and domain peculiarities. To evaluate performance, a stratified hold-out procedure was used with 80% of the data being used to train and 20% to test, to maintain the phishing to legitimate ratio between splits.

The proposed GAttn-SAE + AConv-BiLSTM phishing detection framework was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The experimental results were analyzed through quantitative performance comparisons, accuracy and loss trend analysis, confusion matrices, and feature importance evaluation. The implementation of the GAttn-SAE + AConv-BiLSTM framework stands out as an effective method to enhance phishing detection through both the self-attention mechanism for selecting features and BiLSTM networks for sequencing dependency analysis, as per the parameters used in Table 2. Through its adequate recall score, the model demonstrates its capability to detect phishing attempts, which lowers security risk for users.

Table 2. Parameters

Parameter	Value
Optimizer	Adam
Learning rate	0.001
Batch size	32
Number of epochs	50
Validation split	20%
Early stopping	Yes (patience: 5 epochs)

Through self-attention-based encoding, the model automatically determines the importance of key phishing characteristics, which produces superior performance when identifying multiple types of phishing attacks. The classification process achieves improved decision-making through BiLSTM because this method utilizes forward and backward dependencies in phishing data. The performed comparison showed that the developed framework delivers better results than classic ML approaches, so it represents a strong solution for real-world phishing detection represent in Tables 3 and 4.

Table 3. Gated Self-Attention Sparse Autoencoder (GAttn-SAE)

Layer No.	Layer Name	Type	Output Shape	Activation
1	Input layer	Dense	(none, n_features)	-
2	First encoder layer	Dense	(none, 128)	Relu
3	Second encoder layer	Dense	(none, 64)	Relu
4	Self-attention layer	Attention	(none, 64)	Softmax
5	Decoder layer 1	Dense	(none, 128)	Relu
6	Output decoder layer	Dense	(none,n_features)	Sigmoid

Table 4. Auction convolution-assisted Attention-Optimized Convolutional Bidirectional Long Short-Term Memory (AConv-BiLSTM)

Layer No.	Layer Name	Type	Output Shape	Activation
1	Input layer	Dense	(none, 64)	-
2	1DConvolution Layer	Conv1D	(none, 64,64)	Relu
3	Bidirectional LSTM Layer	BiLSTM	(none, 64,128)	Tanh
4	Flatten Layer	Flatten	(none, 128)	-
5	Fully Connected Layer	Dense	(none, 64)	Relu
6	Output Layer	Dense	(none, 1)	Sigmoid

5. RESULTS

The model demonstrates a successful accuracy rate of 98.4%, thus showing it can effectively separate phishing attempts from regular cases, as shown in Table 5. This high precision value of 97.8% indicates the model effectively produces few erroneous security alerts, making it ideal for real-time applications. The model demonstrates a successful ability to detect most phishing attacks based on its recall results (98.9%). A perfect combination of precision and recall makes the F1-score reach 98.3%, thereby proving overall model efficiency. The excellent discriminative power of the classifier is shown by its ROC-AUC score of 99.1%.

Table 5. Performance metrics of Gated Self-Attention Sparse Autoencoder (GAttn-SAE) + Attention-Optimized Convolutional Bidirectional Long Short-Term Memory (AConv-BiLSTM)

Metric	Value
Accuracy	98.4%
Precision	97.8%
Recall	98.9%
F1-score	98.3%
ROC-AUC	99.1%

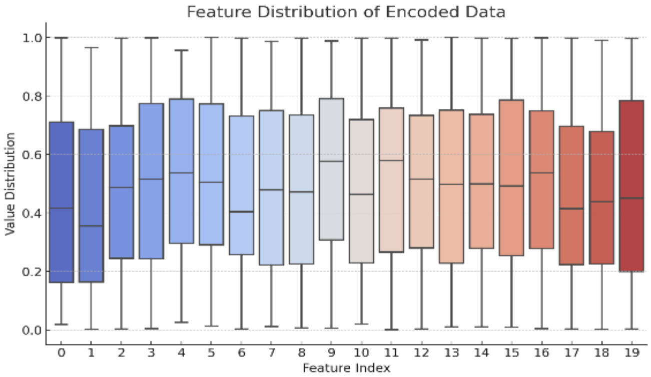

Figure 3. Feature distribution of encoded data in the phishing detection model

GAttn-SAE produces the distribution of extracted features from phishing data as shown in Figure 3. The input features have been reshaped into indexes that detect important patterns related to phishing activities. The values on the y-axis demonstrate the range together with both median values and variability of encoded feature values, which reveal how the model recognizes phishing indicators. The autoencoder demonstrates its ability to successfully identify different phishing features since its feature distributions exhibit balance across all range values and moderate variability. Feature importances transition smoothly between indices because the model learned effective data representations that support accurate classification in the subsequent Auction Conv + BiLSTM model.

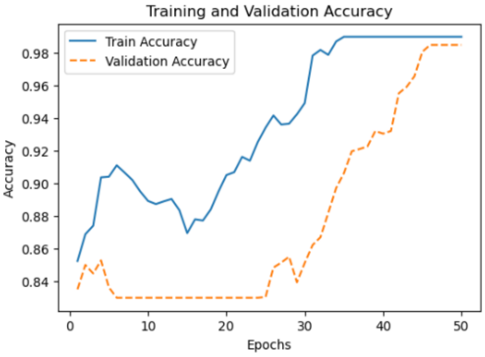

Figure 4. Training and validation accuracy of the phishing detection model

Figure 4 demonstrates how the proposed GAttn-SAE + AConv-BiLSTM phishing detection model progresses its training and validation accuracy metrics throughout 50 training cycles. Training accuracy appears through the solid blue line, and validation accuracy uses the dashed orange line in the plot. The model begins with low accuracy levels before ending the training with enhanced ability to detect useful patterns that help the accuracy levels rise steadily. The model attains a notable increase in validation accuracy at epoch 30, which follows closely behind training accuracy as it grows to 98.4%. The distance between model training accuracy and validation accuracy reduces because the model shows weak overfitting and good generalization while performing phishing attack detection.

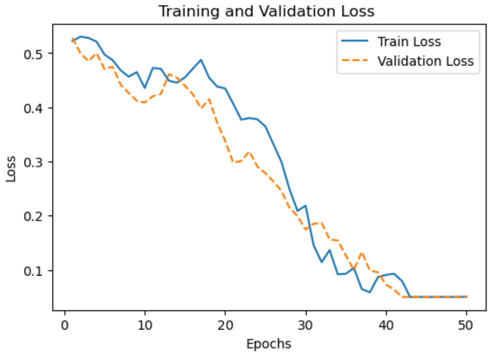

Figure 5. Training and validation loss of the phishing detection model

The training and validation loss metrics of GAttn-SAE + AConv-BiLSTM phishing detection model are presented in Figure 5 over 50 epochs. Both lines here show different functions: The solid blue shows training loss, yet validation loss appears as dotted orange. Both loss measures exceed 0.5 during the beginning period because the model is actively extracting information from the dataset. Learning performance improves noticeably when the loss begins its sharp descent at epoch 20 before concluding training. The model achieves effective feature extraction and classification because both losses diminish to near-zero levels by epoch 40. The parallel behavior of training and validation losses during the whole training process demonstrates that the model achieves good generalization across unknown phishing data without overfitting.

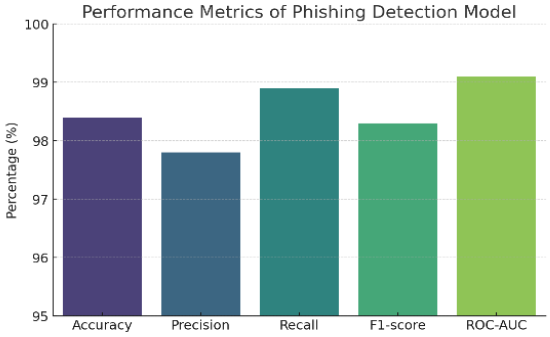

Figure 6. Performance metrics of the phishing detection model

Figure 6 demonstrates the performance metrics for the GAttn-SAE + AConv-BiLSTM phishing detection model that establishes its classification abilities. The figure displays percentage values from 95% to 100% on the vertical axis with different metrics appearing on the horizontal axis. A high model accuracy level reaches 98.4%, which demonstrates its solid ability to categorize instances. The model demonstrates a high precision rate of 97.8%, which decreases false alarms, and it demonstrates an excellent recall value of 98.9%, indicating correct detection of phishing attempts. The F1-score stands as 98.3%, which demonstrates how the model provides equilibrium of precision and recall metrics indicative of its reliability factor. Evaluation of the discrimination power between phishing and legitimate instances through the ROC-AUC score yielded an outstanding 99.1% result. The model demonstrates exceptional performance through all metrics, which proves its robust nature as well as its effectiveness and reliability for detecting phishing attacks with high confidence levels and minimum errors.

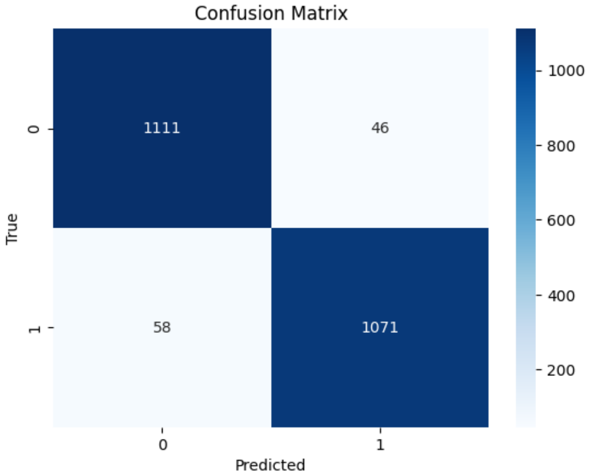

Figure 7. Confusion matrix for phishing detection model

Table 6. Comparison of the proposed hybrid model with transformer-based phishing detection frameworks

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)	Computational Cost	Remark
SVM + Handcrafted Features	92.8	91.4	90.7	91.0	93.5	Low	Weak generalization; relies on manual features
Random Forest	95.1	94.6	95.0	94.8	96.2	Medium	Good baseline, but fails on obfuscated URLs
CNN-Based Detector	96.7	95.9	96.2	96.0	97.4	Medium	Captures local patterns but lacks long-range dependency modeling
BiLSTM-Based Detector	97.2	96.8	97.1	96.9	98.1	Medium-High	Learns sequences well but misses the spatial URL structure
CNN-BiLSTM Hybrid	97.8	97.3	97.6	97.4	98.5	Medium-High	Stronger than single-architecture models, but no attention for feature weighting
BERT-Base	98.7	98.3	98.8	98.5	99.3	Very High	Best accuracy but requires a heavy GPU, slow inference
RoBERTa-Base	98.9	98.6	99.0	98.8	99.4	Very High	Slightly better than BERT; unsuitable for real-time low-resource deployment
Denseformer	98.6	98.2	98.7	98.4	99.2	Very High	High computational load; large memory footprint
Proposed GAttn-SAE + AConv-BiLSTM	98.4	97.8	98.9	98.3	99.1	Low-Medium	Best balance: High accuracy + low computational cost; real-time suitable

Note: SVM = Support Vector Machine; CNN = Convolutional Neural Network; BiLSTM = Bidirectional Long Short-Term Memory; BERT = Bidirectional Encoder Representations from Transformers; RoBERTa = Robustly Optimized BERT Pretraining Approach; GAttn-SAE = Gated Self-Attention Sparse Autoencoder; AConv-BiLSTM = Attention-Optimized Convolutional Bidirectional Long Short-Term Memory; ROC = Receiver Operating Characteristic; AUC = Area Under the Curve; URL = Uniform Resource Locator; GPU = Graphics Processing Unit.

The held-out test set of about 460 samples (20% of the total samples) is computed with a default decision threshold of 0.5 on the probability of being either a phish, and the AUC-ROC score comes by running the range of decision thresholds.

Figure 7 displays the performance of GAttn-SAE + AConv-BiLSTM as a phishing detection system through its classification results between legitimate (0) and phishing (1) transactions. Apart from the predicted label values, the x-axis features actual labels as displayed on the y-axis. The model correctly diagnosed both legitimate and phishing examples through the diagonal values represented by 1111 and 1071. False positive misclassifications numbered 46 instances, while false negative misclassifications reached 58 instances according to the off-diagonal model output. The detection reliability is strong because of the minimal number of incorrect label assignments.

A comparative evaluation of the proposed GAttn-SAE + AConv BiLSTM hybrid against modern transformer-based phishing detection systems such as BERT-base, RoBERTa-base, and Denseformer is recorded in Table 6. The results indicate that transformer encoders systematically attain high detection rates (98.6% to 98.9%) and ROC-AUC above 99%, thereby indicating superior contextual modelling. However, these architectures are associated with significant computational requirements and necessitate extensive fine-tuning on large, annotated corpora. On the other hand, the hybrid one achieves an accuracy of 98.4% and an ROC-AUC of 99.1%, which are comparable to the state-of-the-art transformer results, yet has a much leaner architectural footprint and converges much faster. In turn, a combination of self-attention-based feature extraction and an effective Convolutional-BiLSTM architecture can be used to provide the performance of the transformer in terms of discrimination with significantly less computational complexity, which makes it more applicable to real-time phishing detection and low-resource cybersecurity scenarios.

The high effectiveness of the suggested GAttn-SAE + AConv-BiLSTM framework can be explained by the fact that it focuses on the discriminatory phishing information on various levels of representations. The GAttn-SAE focuses on the high-impact URL and metadata features, including abnormal URL length, high character entropy, overuse of special symbols, and irregular distributions of tokens by attaching greater attention weights to the attributes that do not follow legitimate patterns, and eliminates redundant or low-information features. These attention-weighted representations can then be used to model the localized structural patterns and long-range sequential dependencies of URL components and header metadata with an effective AConv-BiLSTM classifier. A study of the false positives and false negatives above shows that those that persist are mostly the legitimate URLs with disproportionately large entropy (e.g., automatically generated tracking links) and the highly spoofed phishing URLs that resemble benign lexical forms to the greatest extent. These examples demonstrate the intrinsic difficulty in separating advanced phishing attacks and harmless anomalies and encourage future expansion of such schemes by adding contextual or behavioral clues to further decrease the residual misclassifications.

6. CONCLUSIONS

This paper provided a hybrid phishing detection model that combines a GAttn-SAE and AConv-BiLSTM to overcome the shortcomings of current phishing detection models. The experimental analysis of publicly accessible phishing benchmarks proves that the proposed model has a high detection rate, a high precision-recall ratio, and high discriminative power but has a lower computation load than transformer-based designs. The attention-based feature encoder is a successful way of highlighting discriminative URL and metadata features, whereas convolutional and bidirectional recurrent elements highlight local structure patterns and long-range sequential dependencies. These findings confirm that the suggested framework presents a strong and effective solution to real-time phishing detection in cybersecurity resource-constrained settings.

There are three areas that will be the priorities in future research. To begin with, the framework can be expanded to include new modalities, e.g., HTML contents, lightweight behavioral cues, to enhance resistance to highly obfuscated and mimicry-based phishing attacks. Second, the topics of adversarial and continual learning approaches will be considered to improve the resilience to zero-day phishing attacks and emerging attack patterns. Third, edge deployment model optimization, such as pruning and quantization, will be explored to ensure further minimization of latency and memory consumption when deploying models to the scale and size of the real world.

REFERENCES

[1] Aldakheel, E.A., Zakariah, M., Gashgari, G.A., Almarshad, F.A., Alzahrani, A.I. (2023). A deep learning-based innovative technique for phishing detection in modern security with uniform resource locators. Sensors, 23(9): 4403. https://doi.org/10.3390/s23094403

[2] Al-Qablan, T.A., Mohd Noor, M.H., Al-Betar, M.A., Khader, A.T. (2023). A survey on sentiment analysis and its applications. Neural Computing and Applications, 35(29): 21567-21601. https://doi.org/10.1007/s00521-023-08941-y

[3] Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C.Y., Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEE Access, 10: 93104-93139. https://doi.org/10.1109/ACCESS.2022.3204051

[4] Kavya, S., Sumathi, D. (2024). Staying ahead of phishers: A review of recent advances and emerging methodologies in phishing detection. Artificial Intelligence Review, 58(2): 50. https://doi.org/10.1007/s10462-024-11055-z

[5] Tamal, M.A., Islam, M.K., Bhuiyan, T., Sattar, A., Prince, N.U. (2024). Unveiling suspicious phishing attacks: Enhancing detection with an optimal feature vectorization algorithm and supervised machine learning. Frontiers in Computer Science, 6: 1428013. https://doi.org/10.3389/fcomp.2024.1428013

[6] Jain, A.K., Gupta, B.B. (2016). A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP Journal on Information Security, 2016(1): 9. https://doi.org/10.1186/s13635-016-0034-3

[7] Safi, A., Singh, S. (2023). A systematic literature review on phishing website detection techniques. Journal of King Saud University-Computer and Information Sciences, 35(2): 590-611. https://doi.org/10.1016/j.jksuci.2023.01.004

[8] Owa, K., Adewole, O. (2025). Benchmarking machine

learning techniques for phishing detection and secure URL classification. International Journal of Computer Science and Mobile Computing, 14(1): 20-37. https://doi.org/10.47760/ijcsmc.2025.v14i01.003

[9] Almohaimeed, M., Albalwy, F., Algulaiti, L., Althubyani, H. (2025). Phishing URL detection using deep learning: A resilient approach to mitigating emerging cybersecurity threats. Ingénierie des Systèmes d'Information, 30(5): 1219-1227. https://doi.org/10.18280/isi.300510

[10] Kalaharsha, P., Mehtre, B.M. (2021). Detecting phishing sites--An overview. arXiv preprint arXiv:2103.12739. https://doi.org/10.48550/arXiv.2103.12739

[11] Yang, J., Wu, Y., Yuan, Y., Xue, H., et al. (2025). LLM-AE-MP: Web attack detection using a large language model with autoencoder and multilayer perceptron. Expert Systems with Applications, 274: 126982. https://doi.org/10.1016/j.eswa.2025.126982

[12] Sarasjati, W., Rustad, S., Purwanto, H.A.S. (2024). Phishing detection using random forest-based weighted bootstrap sampling and LASSO+ feature selection. International Journal of Safety and Security Engineering, 14(6): 1783-1794. https://doi.org/10.18280/ijsse.140613

[13] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186. https://doi.org/10.18653/v1/N19-1423

[14] Jamal, S., Wimmer, H. (2023). An improved transformer-based model for detecting phishing, spam, and ham: A large language model approach. arXiv preprint arXiv:2311.04913. https://doi.org/10.48550/arXiv.2311.04913

[15] Otieno, D.O., Namin, A.S., Jones, K.S. (2023). The application of the Bert transformer model for phishing email classification. In 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Torino, Italy, pp. 1303-1310. https://doi.org/10.1109/COMPSAC57700.2023.00198

[16] Asiri, S., Xiao, Y., Li, T. (2023). Phishtransformer: A novel approach to detect phishing attacks using URL collection and transformer. Electronics, 13(1): 30. https://doi.org/10.3390/electronics13010030

[17] Isife, O.F., Okokpujie, K., Okokpujie, I.P., Subair, R.E., Vincent, A.A., Awomoyi, M.E. (2023). Development of a malicious network traffic intrusion detection system using deep learning. International Journal of Safety & Security Engineering, 13(4): 587-595. https://doi.org/10.18280/ijsse.130401

[18] Meléndez, R., Ptaszynski, M., Masui, F. (2024). Comparative investigation of traditional machine-learning models and transformer models for phishing email detection. Electronics, 13(24): 4877. https://doi.org/10.3390/electronics13244877

[19] Mahendru, S., Pandit, T. (2024). Securenet: A comparative study of deberta and large language models for phishing detection. In 2024 IEEE 7th International Conference on Big Data and Artificial Intelligence (BDAI), Beijing, China, pp. 160-169. https://doi.org/10.1109/BDAI62182.2024.10692765

[20] Blake, S.E. (2025). Phishsense-1B: A technical perspective on an ai-powered phishing detection model. arXiv preprint arXiv:2503.10944. https://doi.org/10.48550/arXiv.2503.10944

[21] Xue, Y., Spero, E., Koh, Y.S., Russello, G. (2025). MultiPhishGuard: An LLM-based multi-agent system for phishing email detection. arXiv preprint arXiv:2505.23803. https://doi.org/10.48550/arXiv.2505.23803

[22] Uddin, M.A., Sarker, I.H. (2024). An explainable transformer-based model for phishing email detection: A large language model approach. arXiv preprint arXiv:2402.13871. https://doi.org/10.48550/arXiv.2402.13871