International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Deep Learning-Based Phishing Email Detection for Cybersecurity Applications

1st Rokam Likhita 2nd Ponnamanda Lahari 3rd Gunda Joshith Kumar
Dept. of Computer Application,Aditya University,Surampalem, India
rokamlikhitha@gmail.com ponnamandalahari966@gmail.com joshithurs@gmail.com

4th B N.V Sai Durga 5th Peddinti Rajesh
Dept. of Computer Application,Aditya University,Surampalem, India
saidurgab5@gmail.com rajeshpeddinti98@gmail.com

***Abstract*—Phishing emails are one of the most common types of cybersecurity fraud due to human vulnerability, which enables sensitive data and security of organizations to be lost. Conventional rule-based and signature-driven intrusion detection mechanisms are not able to keep up with the changing sophistication of phishing attacks. The review of phishing email detection methods based on deep learning, the state-of-the-art architecture including the Convolutional Neural Network (CNNs), Recurrent Neural Network (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer based models including the BERT architecture and the Distil BERT one will be used in the current paper. The efficiency of the strategies of deep learning is enormously high as our review of 30 recent publications in the topic indicates an accurateness of 97 to 100 percent. The paper analytically reviews the typical datasets (Enron, Nazario, SpamAssassin), preprocessing techniques, architecture innovations and hybrid models and attention mechanisms, and metrics. Our methodology is generalized to combine multilevel feature extraction, transformer-based contextual comprehension, and attention-based mechanisms. It has been demonstrated through the literature of the experiment that hybrid architecture that involves transformers with recurring layers is of the excellent performance, and the types of models like those of improved DistilBERT and BERT LSTM hybrid reflect a high accuracy of 99-100% with a minimal false positive rates. The current review has outlined the challenges that include the imbalance of datasets, adversarial robustness, and domain generalization and offered the future research movement direction to interpretable AI, federated learning, and real-time deployment systems.**

Keywords: Phishing Detection, Deep Learning, Cybersecurity, BERT, LSTM, CNN, Transformer Models, Email Security, Natural Language Processing, Attention Mechanism

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Individual and organization communication has developed with high rate development of digital communication through email which is one of the most prevalent channels of communication. However, this broad use has made email the important favourite of cyber criminals as well [7]. Among other online assaults, phishing is one of the most common and devastating in its attacks of sensitive information. The phishing type of attack presupposes the employment of artificial emails, which appear to be sent by authoritative agencies, e.g. banks, organizations or popular companies [3]. These e-mail messages are professionally made so that they deceive users into giving confidential information like login details, bank details banking records or personal information [10]. As the level and complexity of phishing attacks continue to change, the challenge of identifying phishing has emerged as a significant problem to the existing security framework.

Rule-based filtering, blacklists and signature-based phishing detection methods are the most popular forms of traditional phishing detection. Although these techniques have become useful in unveiling preexisting threats, the techniques are unsuccessful against the emergent or sophisticated phishing operations [6]. Hackers continue to alter email skins, the words they used as well as links incorporated to beat the traditional email security. This has brought the need to have smarter and versatile detection mechanisms that may manage to detect known and new phishing attacks in future [2].

Recent advancements in deep learning have also offered more possibilities toward the detection of phishing emails. Deep learning models are able to automatically extract helpful patterns on large volumes of email data, and it does not need human feature engineering. Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks as well as Transformer networks have shown successful results in textual and contextual processing of emails. Such models are able to distinguish between valid and unscrupulous messages, as they are capable of reproducing the intricate relationships within the email information [4].

The paper will present the ways on how deep learning techniques can be employed to detect phishing email and explain how modern architectures may be implemented to enhance the precision and trustworthiness of security infrastructures [9]. It is intended to analyze the existing practices and offer a future perspective of advanced deep learning systems in improving security in email messages in the face of new phishing trends.

II. Background Study

A. Background and Motivation

Phishing attacks are an urgent and long-standing threat to cybersecurity imposing a risk to people and companies via deceiving emails intended to steal credentials, financial data, or install malware. Nonetheless, phishing is very efficient as

it can exploit the psychology of a person and keep using new methods to develop its attacks despite decades of investigation and implementation of numerous countermeasures [8]. Conventional detection techniques such as blacklists, heuristic rules and signature based filters have high rates of false positives as well as failure to detect zero-day phishing campaigns.

The swift development of deep learning technologies has provided new opportunities of automated probabilities of phishing. In contrast to traditional machine learning methods that often have to be massively hand-engineered by composing in-depth features, deep learning models are capable of learning hierarchical features of raw email messages in an automatic fashion, which capture multifaceted patterns that differentiate between phishing and legitimate messages [5]. Recent research has shown that deep learning designs, especially Transformer-based designs, and hybrid designs can use over 99 percent accuracy in detection, compared to traditional methods by a significant margin [1].

B. Research Objectives

This paper is aimed at critically reviewing the idea of the phishing email detection system based on deep learning and obtaining the following objectives: Presentation of state of art deep learning structures in phishing email detection. Analyze standard datasets, data pre-processing and evaluation. Suggest a generic system model including the best practices of the recent literature. . Compare performance measures among various architectural strategies. Determine the existing challenges and future research directions.

C. Contributions

The main contributions in this piece of work are:

Review of 30 recent studies on deep learning based phishing detection: CNN, RNN/LSTM, and Transformer architecture. Comparative study of architectural innovations such as attention systems, multilevel feature detection, and hybrid systems. A methodology proposal of best practices in preprocessing, feature extraction and classification. Detailed performance comparison showing the high excellence of hybrid Transformer-LSTM models. Recent gain of knowledge about key issues such as dataset imbalance, adversarial robustness, and explainability.

D. Paper Organization

The rest of this paper will be structured as follows: Section 2 will give a review of the related work and the existing literature on methods of phishing detection. The proposed system model and methodology are discussed in section 3. Section 4 provides details of experimental setup and implementation. In section 5, results and comparative analysis is presented. Section 6 is the conclusion of the paper and the future research directions are described.

III. Related Work

A. Evolution of Phishing Detection Methods

Phishing detection has been through numerous generations of approach. Past systems include blacklists and URL blocking that failed to work with the new attacks and polymorphism attacks. Machine learning models introduced in the 2000s offered hand-scraped features, based on email headings, email body and embedded URLs with fair success, but feature engineering was very dependent on domain experience.

The process of feature detection based on raw data feature learning is now feasible since with the introduction of deep learning, the process of detecting phishing has changed completely. Deep neural networks can detect minor linguistic patterns, structural and contextual irregularities that otherwise cannot be detected by a human expert.

B. Convolutional Neural Network Approaches

Phishing email detection using Convolutional Neural Networks has been applied successfully using text as a sequence and local n-gram patterns. Altwaijry et al. created one-dimensional CNN-based models (1D- CNNPD) to detect phishing and with great enhancement added a recurrent layer. Their Advanced 1D-CNNPD with Leaky ReLU and BiGRU had a 100 percent accuracy and 99.68 percent accuracy on standard data sets such as Phishing Corpus, Spam Assassin.

CNNs are good at recognising spatial structure and local feature on text data. Melon et al. have shown that CNN architectures have 97 per cent accuracy, precision and recall of 0.97, on a bespoke dataset of 10,000 labelled URLs. The fact that CNNs can process text of various scales makes them especially useful in detecting signs of phishing in email text.

C. Recurrent Neural Networks and LSTM

The recurrent architectures, specifically, Long Short-Term Memory networks, have been shown to be very effective in modeling sequential dependencies in email text. A comparative study by Truong et al. showed that LSTM had the highest accuracy of 99.41 per cent than any other recurrent architecture significantly better than vanilla RNNs. Foreign to help analyze email structure and linguistic patterns, the capacity of LSTMs to penetrate long-range dependencies causes them to be excellent.

Variants of LSTMs are also further developed to utilize text bi-directionally in forward and backward directions, which provides contextual information of both previous and the future tokens with improved performance. Dewis et al. created a text-based LSTM hybrid system named Phish Responder that had 99 percent and 0.99 percent accuracy and precision respectively. The effectiveness of LSTM designs proves the significance of sequential modelling in the recognition of the email semantics.

D. Transformer-Based Models

Transformer architectures and in particular BERT and its variants have revolutionised natural language processing and have demonstrated remarkable success in phishing detection.

These models use self-addressing machinery to retrieve the contextual associations within entire email messages, which allows elaborate semantic insights.

Farhan et al. proposed an Enhanced DistilBERT model following BERT representations, bidirectional LSTM, and attention models with 100 percent accuracy, 100 percent precision, 100 percent recall, and 100 percent F1-score and 99.76 percent ROC AUC. The model proved to be more resilient to adversarial noise than the traditional DistilBERT and RoBERTa. Melendez et al. discovered that transformer models by far exceed the performance of traditional machine learning models as RoBERTa with 99.43% accuracy and 98.76% accuracy with the best traditional model (SVM).

Ibrahim et al. applied BERT and RoBerta to the blended Nazario Phishing Corpus using Enron emails, which proves that transformer models can be effective with balanced data. Alhuzali et al. tested various types of transformers, such as distilBERT, BERT, XLNet, RoBERTa, and ALBERT, on a dataset of 119,148 emails and made comparisons of their performance in detail. Uddin et al. created a model, explainable transformer-based DistilBERT, that attained a testing accuracy of 98.48 per cent with a model interpretable with LIME and Transformer Interpret.

E. Hybrid and Attention-Based Architectures

Recent research has enquired into having hybrid frameworks that combine the advantages of over one deep learning paradigm. According to the proposals of Fang et al., THEMIS is a decelerated Recurrent Convolutional Neural Net (RCNN) whereby multilevel vectors use multirevel and attention solutions respectively. THEMIS classifies both email headers and bodies on a character and word-level with Bidirectional LSTM to capture contexts to reach an accuracy of 99.848% under an exceptionally low false positive rate of 0.043% on imbalanced data in the IWSPA-AP 2018 shared task.

A suggestion of Transformer encoder + BiLSTM networks was offered by Vasudevan et al. to detect phishing on a case-by-case basis. The architecture is applied in order to utilize the power of BiLSTMs on sequential dependencies, and multi-head attention lengths on extraction of contextual features to achieve an overall 99 percent accuracy on datasets characterized by generative and naturally generated phishing email. Atawneh et al. also found that a BERT-LSTM hybrid achieves the highest accuracy, 99.61 percent and the highest precision, 99.87 percent recall, 99.55 per cent, and F1-score, 99.55, with the rest of the setups of deep learning models using publicly available datasets such as Enron datasets.

Attention processes have also played significant roles in the process of ensuring that relevant features of email information have been paid attention. Chen et al. proposed a deep attention collaborative model of secure educational email services having nonnegative matrix factorization and deep attention as the types of social information analysis and personalized detection. Gupta et al. adopted the application of multi-stage architecture with a combination of BERT to extract features and CNN to classify the features in enterprise systems.

F. Specialized Approaches

Several researches have explored professional ways of addressing phishing vulnerability-specific challenges. Gogoi et al. used pretained transformer models and trained on transfer learning to obtain accurateness, recall, and F1-score = 0.99. The scheme applies the pretraining experience at large scale to performance on small data sets of phishing.

The use of federated learning has turned into a privacy-control method of collaborative model learning. The Federated Phish Bowl system incorporates BiLSTM that runs in a decentralized system where the accuracy is 83 percent and the protection of sensitive email data is ensured. A second study demonstrated a comparison on federated learning with THEMIS and BERT models and revealed that it acted in similar ways as centralized learning with the existence of resilience to asymmetry in data distribution.

Explainable AI has been gaining deepwater to set confidence on phishing detection systems. Having created a web-based AI platform that combines Explainable AI methods, Al-Subaiey et al. attained 99.1 percent accuracy on SVM and TF-IDF and offered interpretable predictions. Comparisons of TabNet, NODE and FT-Transformer models to SHAP and LIME on interpretability (asal) showed that FT-Transformer had an accuracy rate of 97.8 per cent; a feature of the model that returned spelling errors as well as urgency associated words were reported to be valuable features.

TABLE I
PERFORMANCE COMPARISON OF DEEP LEARNING ARCHITECTURES

Architecture	Dataset(s)	Accuracy	Precision
Enhanced DistilBERT	Enron + Nazario + CEAS	100%	100%
1D-CNNPD + Bi-GRU	Phishing Corpus, SpamAssassin	99.68%	100%
THEMIS (RCNN + Attention)	IWSPA-AP 2018	99.848%	99.664%
Transformer + BiLSTM	Mixed authentic/generative	99%	–
BERT + LSTM	Enron + public datasets	99.61%	99.87%
LSTM	Phishing/legitimate emails	99.41%	–
RoBERTa	Large email dataset	99.43%	–
PHISH NET (CNN)	Labeled phishing/legitimate	98.57%	–
CNN	PhishTank + Common Crawl	97.3%	0.97

IV. PROPOSED METHODOLOGY

A. System Architecture Overview

Based on the comprehensive literature review, we propose a generalized deep learning-based phishing email detection system that synthesizes best practices from state-of-the-art approaches. The system architecture consists of five main components: data preprocessing, multilevel feature extraction, contextual encoding, attention-based fusion, and classification. Figure 1 illustrates the complete system model.

logits. Multiple dense layers with ReLU activation enable non-linear decision boundaries. Output Layer: A sigmoid activation function produces probability scores for binary classification (phishing vs. legitimate). For multi-class scenarios (phishing, spam, legitimate), softmax activation is used. Loss Function: Binary cross-entropy loss is commonly used for training:

“‘ L = -[y log(yˆ) + (1-y) log(1-yˆ)] “‘

where y is the true label and yˆ is the predicted probability [2], [16]. Custom loss functions can incorporate false positive penalties to minimize user disruption.

F. Hybrid Architecture Design

The best solutions are a combination of several architectural elements: Transformer-LSTM Hybrid: This architecture encodes context by use of transformer encoder then levels through BiLSTM layers to sequence the models. Transformer can absorb global context and LSTM can remove irregularities in sequence. This mixture has excellent performance in the area of complementary strength. CNN-RNN Hybrid: CNN layers are used to extract local features and these feature elements are processed in RNN layers to learn sequential dependencies. It is a very efficient architecture that is useful in handling character-level and word-level representations. RCNN with Attention: Recurrent Convolutional Neural Networks utilize convolutional feature generation and recurrent processing, with the added benefit of attention. An example of this method is THEMIS, which had 99.848 percent multilevel, multivector accuracy with attention.

G. Training Strategy

Appropriate training must be done in accordance with various factors: Transfer Learning: Phishing data is trained on pretrained language models (BERT, DistilBERT, RoBERTa) which are trained on large blocks of data, exploiting large-scale pretraining. The process will be known as fine-tuning that will bring the general comprehension of the language into the specificities of the phishing emails. Regularization: The drop out layers have the effect of preventing overfitting by randomly shutting off neurons during the training process. The normalization of the layer stabilizes deep architecture training. Late term overtraining will be prevented by premature validation-based early stopping. Optimization: The standard one applied to train deep model is Adam optimizer having a learning rate scheduling. Batches are selected in accordance with the computational resources and data collection size. Gradient clipping uses exploding gradient in graffiti designs.

V. IMPLEMENTATION

A. Datasets

Common Benchmark Datasets-: Various standard datasets are regularly used in the literature to detect phishing emails: Enron Email Dataset: It is a bulk or large size of legitimate corporate mail which has usually been made use of as a source of benign samples. Enron data has real-life examples of proper business communication, which models can learn the features of the original emails. Nazario Phishing Corpus: A well-known collection of spam emails that are utilized as one of the primary sources of compromising samples. Naziario corpus is saturated with the phishing campaigns of every sort of organizations and people. SpamAssassin: The is a publicly available set of spam and phishing email that is actively used to benchmark scoring. The labeled examples in SpamAssassin also store labeled examples containing detailed header information. IWSPA-AP 2018 Dataset: A dataset that was used at the First Security and Privacy Analytics Anti-Phishing competition is with emails by Wikileaks, Enron and SpamAssassin. This data has unbalanced distributions of genuine life depicting production environment. PhishTank and Custom Collections: There are a few studies that use PhishTank URLs, and some samples of phishing are acquired with the help of custom collections to increase the diversity of the dataset.. The custom datasets are often an aggregation of sets of sources to constitute complete training sets.

Dataset Preparation-: The processing of datasets is usually done in the following way: Collection and Merging: Various phishing corpora are aggregated with legitimate email data to make balanced datasets or realistically imbalanced datasets. Preprocessing: Emails are handled through identification of emails packets (email) having the embedded objects, header, and body. The reading is cleansed, standardized, and tokenized according to the chosen structure. Labeling: These are binary labeling (phishing/legitimate) and research that makes use of other groups (spam, fraudulent). Splitting: Data is partitioned into training set, validation set, and test set which is usually in 70-15-15 or an 80-10-10 split. Stratified splitting is used to allocate even distributions of the classes across splits. Augmentation: In techniques of data augmentation (synonym substitution, back-translation, generative compositions, etc.), training is diversified.

B. Implementation Details

Model Architectures-: The following number of architectural configurations have been identified to work as a result of the literature review: Enhanced DistilBERT Configuration: Base model: DistilBERT which has been trained on broad text corpora. Additional blocks: Bidirectional LSTM (128256-release), attention mechanism. Thick layers: 2-3 dropout (0.3-0.5) fully connected layers. Output: Signoid output in classification Binary. Rated performance: 100 percent actuality, accuracy, recall, F 1-score. Bi-GRU 1D-CNNPD: Embedding layer: 300 dimensional word embeddings. CNN layers: The CNN has a few 1D convolutional layers (3, 4 and 5 filters). Activation: Leaky ReLU Recurrent layer: Bidirectional GRU (128 units). Pooling: Max pooling Dense layers: This is a dense layer that has dropout. Reports found: 100, 99.68 percent accuracy. THEMIS RCNN Configuration: Multilevel inputs: Character and word level header and body embeddings. Recurrent cell: LSTM (256 units) 2-way. Mechanism of attention: header-body cross- attention. Dense layers Multiple full connected layers. Reported performance: 99.848% accuracy, 0.043% FPR Transformer-BiLSTM Hybrid Configuration: Transformer encoder Multi-heads self-attention (8 heads). Recurrent layer:

256 unit 2-way LSTM. Thick layers: Fully connected layers and dropout. Reported performance: 99% accuracy

C. Training Configuration

Common training configurations across studies include: Hyperparameters: Learning rate transformation to transformer fine tuning: 1e-5-5e-5 to CNN/RNN models. Size of batch: 16-64 depending on the size of the model and available Gordon memory. Epoch: 10-50 and early stopping (validation loss). Optimizer: Adam weight decay (1 e -4 to 1 e -2) Regularization: Separation: 0.3 to 0.5 thick layers. Normalization of architecture of transformers. Max gradient non-critical clipping (max norm 1.0) of recurrent models. Loss Functions: Binary classification binary cross-entropy. Inequality between classes: Custom weighted loss. The focus on hard instances by going off track. Computational Resources-: The computational facility employed normally is the following in investigations: Hardware: NVIDIA GPUs (Tesla V100, RTX 3090, A100) with 16-40GB memory PyTorch or TensorFlow Hugging Face Transformers library based regimes of BERT. Training Time 2-12 hours with respect to the datasets and challenges of the model.

D. Evaluation Metrics

Primary Metrics-: The literature consistently reports the following evaluation metrics: Accuracy: Overall correctness of predictions, calculated as (TP + TN) / (TP + TN + FP + FN).

Precision: Proportion of predicted phishing emails that are quantum phishing, which is in TP / (TP + FP)]. False alarms are minimized by high precision. Recall (Sensitivity): Percentage of genuine phishing emails that were reported correctly, TP / (TP + FN). Massive recall attracts minimal phishing spam emails to prevent attention. F1-Score: harmonic mean between precision and recall, and it is 2(Precision Recall)/(Precision + Recall). F1-score gives a measure of performance that is balanced. Additional Metrics: False Positive Rate (FPR): Ratio of genuine email falsely identified phishing, FP/ (FP TN). There is a critical requirement of low FPR. False Negative Rate (FNR): A fraction of the number of phishing emails incorrectly identified as legitimate emails to the total number of phishing emails and represented as is FN / (FN + TP). Security of Low FNR is necessary. ROC AUC: Area under the Receiver Operating Characteristic curve; it is a measure of the discriminating ability of the model at different levels of classification. Higher point in AUC is an indicator of better performance. Evaluation Protocols: Cross-Validation: K fold cross-validation (usually 5 or 10 folds) guarantees sound performance estimates. Hold-Out Testing: Various groups of tests determine generalization to unobserved data. Cross-Dataset Evaluation: It involves the evaluation of other datasets to estimate the robustness and transferability of models. Adversarial Testing: Adversarial testing is a type of evaluation that tests the model on adversarial perturbed examples of data, in order to determine the adversarial examples.

E. Baseline Comparisons

Deep learning approaches are customarily compared to a number of baselines: Conventional Machine Learning Support Vector Machines (SVM), Decision Trees, Naive Bayes, and random forests are used as baselines. These are methods that produce feature engineering manually and they are usually 90-98Rule-Based Systems: Blacklists and heuristic Blacklists and heuristic Rule-based email filters offer baseline performance, usually with a large false positive rate. As Ablation Model: Over simpler Deep Learning Models Basic CNN or RNN models without attention or other hybrid modules are used as ablation baselines. Fine-tuning-free Pretrained Models: The consistency of the adaptation value is demonstrated by off-the-shelf transformer models without fine-tuning to domains.

VI. Results and Discussion

A. Performance Analysis

Overall Performance Comparison: The comprehensive literature review reveals that deep learning approaches consistently outperform traditional methods for phishing email detection. Table 2 presents a detailed comparison of performance metrics across different architectural families:

TABLE II
Comparative Analysis of Architecture Families

Architecture Family	Best Accuracy	Precision	Recall	F1-Score	Key Studies
Transformer-only	99.43–100%	99–100%	97–100%	98.48–100%	[1],[3],[4]
Hybrid Transformer–LSTM	99–100%	99.87–100%	99.23–100%	99.55–100%	[1],[6]
RCNN + Attention	99.848%	99.664%	99%	99.331%	[5]
CNN + RNN	97–99.68%	97–100%	94–99.32%	95–99.66%	[2],[8]
LSTM/BiLSTM	99–99.41%	99%	94–99%	95–99%	[7],[13]
Traditional ML	90–98.76%	93–99%	90–96%	92–97%	[8],[10]

These findings prove that transformer-based and hybrid architectures outperform better, and some of them have a near-perfect accuracy.

Transformer Model Superiority: Models based on transformers show outstanding results in various studies. The Enhanced DistilBERT by Farhan et al. scored a 100 percent accuracy, precision, recall, and F1-score with a ROC AUC of 99.76 per cent on a mixture of Enron, Nazario, CEAS, and other sources. False-positive minimization and resistance to adversarial noise, which are specific to the custom loss of the model, make it especially effective in the production deployment.

In a direct comparison of transformer models and usual machine learning models, Mele´ndez et al. established that RoBERTa scored 99.43 percent accuracy as opposed to 98.76 percent in the most efficient traditional model (SVM). This great difference illustrates the importance of advanced text-processing tools and context knowledge offered by transformers.

Alhuzali et al. compared proxies of different transformer variants (distilBERT, BERT, XLNet, RoBERTa, ALBERT) using 119,148 emails, which evidence suggests that transformer architectures scale well to large datasets. The numerous phishing repositories used in the study facilitate various phishing schemes getting recorded.

Hybrid Architecture Advantages: Hybrid models of transformers and recurrent layers give the optimum compromise in contextual knowledge and sequential modeling. Atawneh et al. have shown that BERT-LSTM hybrid was better than single CNN, RNN, LSTM, and BERT models with 99.61, 99.87, 99.23 and 99.55 accuracy, precision, recall, and F1-score respectively. The hybrid model uses contextual embedding of BERT and LSTM to capture sequential dependencies unique to emails.

The Transformer-BiLSTM by Vasudevan et al. attains 99 percent accuracy by using multi-head attention to extract contextual features followed by BiLSTM to extract sequential dependencies. The architecture is effective in gaining contextual and sequential patterns, which can prove to be better in detecting phishing.

Another successful hybrid model of Fang et al. system, which is another attempt to combine a better RCNN with multilevel vectors and attention mechanisms, is called THEMIS. THEMIS, by filtering email headers and messages at both character and word respects, obtained an accuracy of 99.848 percent and a very low false positive rate of 0.043 percent on unbalanced data. This low FPR is especially significant in case in production systems false alarm overrun users.

The performance of CNN and RNNs is compared, as shown below: Convolutional and recurrent architectures are slightly less accurate compared to transformers, but perform excellently and have computational benefits. Advanced 1D-CNNPD with Leaky ReLU and Bi-GRU by Altwaijry et al. on Phishing Corpus and SpamAssassin gave 100 percent accuracy and 99.68 percent precision [2]. Permission of CNN to recurrent layers enhanced performance relatively high than the foundation 1D- CNNPD model.

In a comparative study, Truong et al. established that LSTM ranked the highest accuracy of 99.41 percent of recurrent architectures. The high performance of LSTM compared to the vanilla RNN proves the sensitivity of the long-term memory in the representation of email structure and language patterns.

Melon et al. demonstrated that CNNs are getting 97.3 percent and 0.97 precision and recall on a custom dataset of 10,000 labelled URLs. Although performing worse than the transformers, CNNs have a shorter inference time and reduced computation demands and hence they can be tailored to resource-constrained settings.

B. Key Findings

Architectural Innovations: The architectural inventions have been very effective in some of them.

Soft multilevel Feature extraction: At different levels (character, word, header, body), a number of phishing indicators can be accessed when employing the emails. The most accurate result of the multilevel approach of THEMIS was the 99.848 when taking into account the small scale patterns of the orthography, and the large scale semantic content.

Attention Mechanisms: The concept of attention helps models to concentrate on the most pertinent features in order to make a classification. Header-body cross-attention detects any differences between metadata and content which is a common factor of phishing [5].

Hybrid Architectures: The hybridizing architectures (transformer + LSTM, CNN + RNN) combine complementary models, but apply the benefits of each. Single-modeling will never work compared to the efforts in hybrid models.

Transfer Learning: The performance of the fine-tuning of the language models trained on phishing data is higher than that of training with a blank set. The trained models acquire the universal language knowledge which is effectively applied to the phishing.

Dataset Considerations: The alleged information sets are a significant influence on the model performance.

Balancing Strategies: The combination of phishing corpora and valid email sets yields balanced datasets which boost the training of models. Nevertheless, more realistic representations of production distributions are realistic, unbalanced ratios with as much weighting of the losses as possible.

Dataset Diversity: Various sources of phishing (Nazario, SpamAssassin, PhishTank) are being used to increase pattern diversity and increase generalization. Trained model that is trained on other datasets works better than a model trained on the other datasets.

Dataset Size: The performance of the study is likely to improve as the dataset size is larger with 10,000-119,148 emails being utilized in the studies. However, with good transfer learning, one can even fare well with little data in the real domain.

False Positive Rate: There are no considerations in this category of testing. The false positive value is of great significance to production deployment since excessive false alarms will make users lose confidence in the system and failure to adopt the system. With a FPR of 0.043% and an accuracy of 99.848% THEMIS had an extremely low FPR. A loss function used by the Enhanced DistilBERT by Farhan et al. was designed to reach the lowest number of false positives. Geetha et al. state that this CNN-RNN algorithm reported a false positive error of 3.8 percent, which is still quite low by the conventional rule-based systems.

As depicted in the literature, attention and multilevel feature extraction mechanisms will decrease the rate of false positive aspect because it will contribute to the more advanced decision aspect. With the models that independently study the contents of the header and the body, it is possible to distinguish better between the emails that are legitimate and contain suspicious look contents and phishing.

C. Challenges and Limitations

Adversarial Robustness: Phishing is one of the initial nuisances since they are incessantly finding more methods to bypass detection systems. Specifically, comparison of resilience of their Enhanced DistilBERT model with adversarial noise was as well carried out by Farhan et al. and indicated that it possessed better resilience than regular models. However, adversarial robustness is not explicitly studied in the majority of works.

Adversarial examples may be generated by making slight modifications to the phishing emails and keeping the motive of being malicious without detection. Future research must be designed in such a way that it generates models that are resistant to adversarial perturbation as well as involve adversarial training.

The issue of dataset imbalance and realism: Many studies use artificially balanced data, which does not reflect the situation in the real life where phishing mails are not prevalent. Training with balance datasets can be done, however, not always useful in consideration of performance production.

That was overcome by Fang et al., who compared THEMIS and realistic unbalanced data using IWSPA-AP 2018 in which the model produced the maximum and lowest accuracy (99.848 percent) and FPR (0.043 percent) in realistic data, respectively. More literature must be done on the ways of addressing extreme imbalance of the classes in the encoding of low false positives.

Generalization Across Domains: Models that are trained using particular phishing corpora might lack extrapolation even to novel types of attack or different organizational situations. Mahendru et al. confirmed that DeBERTa V3 generalized much better with appropriately matched data and LLM excelled at generalizing on artificial and synthetic data.

Generalization can only be evaluated through cross-dataset evaluation. Models that have been tested in multi-data sets are more predictors of solidness. The transfer learning technique can improve cross-domain performance, as well as domain adaptation.

Computational Requirements: Transformer-based models are very accurate but run extensively in terms of computational resources during training and inference. This limits their application in resource constrained environments such as mobile devices or edge computers.

DistilBERT possesses lighter architectures, which offer a compromise between performance and efficiency. Mahendru et al. found that DeBERTa V3 takes less time to make predictions than big language models, which can identify real-time better. Quantization and pruning are examples of model compression techniques that can determine the calculations with fewer computations and with accuracy.

Explainability and Interpretability: Deep learning models and transformers are black box models, the interpretation of which can be minimally interpretable. This lack of transparency will make users have no belief in it and it will be difficult to understand why some emails will be termed phishing.

There are several papers, which have addressed this difficulty by considering the Explainable AI methodologies. Uddin et al. employed LIME and Transformer Interpret to decode the predictions of the DistilBERT model and would be able to present the transparency of the decision-making process of the model. Al-Subaiey et al. used XAI in an online application to create credibility of users. Asal et al. used SHAP and LIME to determine the most important features (spelling errors, urgency terms) to make predictions.

The second research direction that is going to apply in the future is based on explainability so that security analysts can appreciate and justify model choices.

Privacy Concerns: Email content is personal or organizational sensitive information and therefore it is a privacy issue to the centralized detection systems. The traditional procedures require central servers to analyze the mailed emails, which can reveal confidential information.

Federated learning is a more privacy-conserving training that is conducted without the spreading of raw email samples in spite of collaborative training. The Federated Phish Bowl system recorded that federated learning is accurate 83 per cent with sensitive data protection. The other study found out that in federated learning with THEMIS and BERT model, asymmetric data distribution resilience is as good as that in centralized learning.

The federated learning gives the chance to implement multi-organizational cooperation in phishing detection, maintaining confidentiality and within the context of enterprises, it is particularly feasible.

D. Practical Implications

Deployment Considerations: The implementation of deep learning-based phishing detection will probably be successful only in case several realistic concerns are taken into account. Real-Time Performance: the process of the detection system cannot be time-intensive to process emails because this will interfere with the users workflow. The lesser models like DistilBert and a fined CNN models are associated with reduced inference times necessary to be deployed as a real-time system. Correction To and Integration with the Existing Infrastructure: The detection systems are to be correlated and integrated with email servers, security information and event management (SIEM) systems, and incident response processes. Al-Subaiey et al. developed an online system which was found to be practically applied.

User Interface and Feedback: The systems should describe types of phishing in a simple way, and should enable users to report on false-positives/negatives. The feedbacks of the users can be taken on board to constantly enhance the performance of the models.

Ongoing Learning: Phishing technology changes fast meaning models need to change according to new attack patterns. The transfer learning and fine-tuning on new samples of phishing aids continuous adaptation.

Cost-Benefit Analysis: The benefits of deep learning modes of detection are massive in comparison to conventional ones. Minimized False Positives: The low numbers of False Positives (0.043-3.8 percent) are lower in comparison to those of the rule-based systems (that is usually relative to 10 percent),, which reduces frustration in the users, and increases their productivity.

Higher Detection Rates: Higher Detection rates 99-100.% are far higher than those of the standard methods (90-95.%) which prevent more successful phishing attacks.

Automated Feature Learning: REMOVED: Removal of manual feature engineering reduces the cost of development and maintenance.

Adaptability: Model can be modified against certain data of the organization to suit certain attacks.

Nevertheless, deployment cost contains computational infrastructure, model training and maintenance, and integration among the existing systems. These expenses have to be contrasted by the organizations with the potential losses in the event of successful phishing attacks.

VII. Conclusion and Future Work

A. Summary of Contributions

The paper provides an extensive overview of how deep learning has been applied to detect phishing email, evaluating 30 new studies in total using CNN, RNN/LSTM, and Transformer architectures. We have classified and evaluated state-of-the-art deep learning frameworks, where it was shown that hybrids of Transformer-LSTM models can be optimized to the highest level with a accuracy of 99-100Methodology Synthesis: We took the best practices used by the literature and offered a generalized system model with multilevel feature extraction, contextual encoding using transformers, attention-based fusion and hybrid architectures.We gave detailed performance comparisons that deep learning models are by far a better solution than the traditional ones with transformer models resulting to a 99.43-100 percent accuracy versus 90-98.76 percent of the traditional machine learning.We singled out such essential challenges as adversarial robustness, imbalance in the dataset, domain generalization, computation needs, explanation, and privacy issues. Along with deployment considerations, cost-benefit analysis and practical implications to implementation in the real-world were discussed.

VIII. Referance

[1] K. Thakur et al., "A Systematic Review on Deep-Learning-Based Phishing Detection," *Electronics*, vol. 12, no. 21, 2023.
[2] S. Salloum et al., "Phishing Email Detection Using Natural Language Processing Techniques," *Procedia Computer Science*, 2021.
[3] P. Bountakas et al., "A Comparison of NLP and Machine Learning Methods for Phishing Email Detection," in *Proc. ACM ARES*, 2021.
[4] V. Sharma et al., "Phishing Detection in Email Using Deep Learning," *International Journal of Modern Science and Research Technology*, 2025.
[5] B. Tesfom et al., "Phishing Detection Using Deep Learning and Machine Learning Algorithms: Comparative Analysis," in *Proc. IEEE DASC*, 2023.
[6] A. Karim et al., "Hybrid Machine Learning-Based Phishing Detection System," *IEEE Access*, vol. 11, pp. 36805–36822, 2023.
[7] A. Aljofey et al., "An Effective Phishing Detection Model Based on Character-Level CNN," *Electronics*, vol. 9, 2020.
[8] M. Somesha et al., "Classification of Phishing Email Using Word Embedding and Machine Learning Techniques," *Journal of Cyber Security and Mobility*, 2022.
[9] H. Shirazi et al., "Adversarial Autoencoder Data Synthesis for Phishing Detection," *IEEE Transactions on Services Computing*, 2023.
[10] A. Kumar et al., "Detection of E-Mail Phishing Attacks Using Machine Learning and Deep Learning," *IJCA*, 2022.