AUT Journal of Electrical Engineering
10.22060/EEJ.2026.25877.6023

Hybrid Deep Learning Models for Malicious URL Detection and Classification

Bekele Mulualem[1], Yirga Yayeh[2*], Adugnaw Alebel[1], Yenework Belayneh[2]

[1]Network Administrator Team Leader, Information and Communication Technology Executive, Injibara University, Ethiopia

[2]Department of Information Technology, Injibara University, Injibara, Ethiopia

Correspondence: byyirga@gmail.com

Abstract: The rapid digitization of education, finance, and healthcare has substantially broadened the attack surface available to cybercriminals, with malicious URLs acting as key channels for phishing, malware, and spam delivery. This danger is especially pronounced in developing countries such as Ethiopia, where the pace of digital uptake exceeds the development of security frameworks. This study responds to the urgent need for robust, scalable systems for detecting malicious URLs by proposing a deep learning framework that jointly leverages DistilBERT for semantic representation, Convolutional Neural Networks for identifying local patterns, and Bidirectional Long Short-Term Memory networks augmented with attention mechanisms to capture sequential dependencies within URLs. To tackle pronounced class imbalance in the original internal samples, the Synthetic Minority Over-Sampling Technique was applied only to the training subset, while the validation and test sets were kept naturally imbalanced to avoid data leakage. Four required safeguards were implemented to rigorously reduce synthetic pattern over fitting: augmentation restricted to the training set, independent external benchmarking on 50,000 URLs with no overlap, close generalization-gap analysis, and verification of per-class performance consistency. Across an extensive experimental corpus of 499,960 total instances, the proposed architecture attains 99.85% stratified 5-fold cross-validation accuracy and 98.80% external accuracy on real-world test supports from URLHaus and PhishTank, supported by strong macro-average F1-scores of 98.5% and 98.1% respectively. These findings substantiate exceptional, unbiased multi-class discrimination among Safe, Phishing, Malware, Spam, and Other Threat categories, providing a scalable, reliable detection framework for resource-limited cyber security settings.

Keywords: Cybersecurity, Deep Learning, Malicious URLs, URL Detection and Classification, SMOTE

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

1. Introduction

The increasing prevalence of malicious URLs on the Internet poses significant security threats, as such URLs are commonly used to distribute malware, launch phishing attacks, and perform other malicious activities. Traditional detection methods, such as blacklist-based systems, often fail to effectively address rapidly evolving threats because they rely on previously identified malicious URLs. As a result, these approaches struggle to detect newly generated or zero-day malicious URLs. To overcome these limitations, researchers have increasingly adopted machine learning and deep learning techniques to improve the detection of malicious URLs, thereby achieving higher accuracy and near-real-time threat identification [1],[2].

Deep learning approaches have emerged primarily due to the critical limitations of traditional cybersecurity mechanisms. Blacklist and whitelist techniques are inherently reactive and ineffective against zero-hour attacks. Similarly, heuristic-based detection methods that rely on third-party services often suffer from high latency and limited detection accuracy. In contrast, deep learning models can automatically learn discriminative patterns directly from raw URL strings without requiring manual feature engineering. This capability reduces dependence on external knowledge bases and enables end-to-end learning, providing a more adaptive and robust defense against continuously evolving cyber threats [3]. Despite these advantages, adversarial attacks remain a major challenge for deep learning-based detection systems. In such attacks, malicious actors deliberately manipulate URL structures to evade detection mechanisms. This challenge underscores the urgent need to develop more robust and resilient deep learning models that can withstand adversarial manipulation.

To address these issues, several studies have proposed mechanisms to enhance the robustness of deep learning models. For example, protective-layer methodologies have been introduced to strengthen model resilience against adversarial inputs. Furthermore, a comparative study in [4] evaluated multiple machine learning and deep learning approaches for malicious URL detection and demonstrated that deep learning models generally outperform traditional machine learning techniques in terms of detection accuracy while reducing false positive rates. Building upon these advancements, the present study aims to further improve malicious URL detection by leveraging advanced deep learning techniques while addressing the limitations of existing models.

1.1. Research Contributions and Novelty

Although CNN and BiLSTM architectures have been widely employed for malicious URL detection, the novelty of this study does not lie in the individual components themselves. Rather, it lies in their effective integration within a unified end-to-end framework that combines transformer-based semantic representation, feature fusion,

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

imbalance-aware learning, and comprehensive evaluation for robust five-class malicious URL detection and classification. Specifically, the proposed framework makes the following contributions:

- A unified DistilBERT-CNN-BiLSTM-Attention architecture that combines contextual semantic embedding, local lexical and structural feature extraction, bidirectional sequential dependency learning, and attention-guided feature refinement to improve feature representation for malicious URL analysis.
- A multi-source feature fusion strategy that integrates transformer-derived contextual embeddings with lexical, structural, WHOIS, blacklist, and content-based features, enabling richer semantic, structural, and behavioral URL representations than approaches relying on a single feature source.
- An imbalance-aware learning framework in which SMOTE is applied exclusively to the training dataset, preventing data leakage, mitigating majority-class bias, and improving the recognition of minority attack categories.
- A strong five class malicious URL detection and classification framework capable of simultaneously distinguishing Safe, Phishing, Malware, Spam, and Other Threat URLs. The framework is systematically validated using both imbalanced and SMOTE-balanced datasets, stratified five-fold cross-validation, and computational complexity analysis, demonstrating its strength, scalability, generalization capability, and useful applicability for real-world cybersecurity environments.
- The rationale and advantages of the proposed DistilBERT-CNN-BiLSTM-Attention Framework while several studies have addressed malicious URL detection using machine learning and deep learning techniques, the proposed framework is specifically designed to overcome critical limitations of existing approaches through five key advantages that collectively justify its selection for real-world cybersecurity applications.
- Integrated Feature Learning: The proposed DistilBERT-CNN-BiLSTM-Attention framework synergistically combines contextual semantic understanding DistilBERT, local pattern extraction CNN, sequential dependency learning BiLSTM, and attention-guided feature refinement. This multi-level feature extraction provides more comprehensive URL representations than standalone models. As demonstrated in Table 14, the hybrid model achieves 98.20% accuracy, significantly outperforming standalone CNN 97.52%, BiLSTM 97.48% , and Attention-BiLSTM 96.88%, representing a +0.68% improvement over the best standalone architecture.
- Multi-Class Threat Identification: Contrasting most prior approaches limited to binary classification (malicious vs. benign), the proposed framework performs five-class classification: Safe, Phishing, Malware, Spam, and Other Threats. This granular classification enables more detailed cybersecurity analysis for Security Operations Centers (SOCs), Internet Service Providers (ISPs), and enterprises. Table 12 demonstrates balanced F1-scores ranging from 97.8% Phishing to 99.3% Safe ,

Uncorrected Proof

with a macro-average F1-score of 98.5%, confirming balanced and unbiased learning across all threat categories.

- Multi-Source Feature Fusion and Imbalance Handling: The framework integrates transformer-derived contextual embeddings with lexical, structural, WHOIS, blacklist, and content-based features, enabling richer semantic, structural, and behavioral URL representations. Critically, SMOTE-based imbalance handling is applied exclusively to the training set 70%, while validation 15% and test 15% sets remain imbalanced to reflect real-world conditions and prevent data leakage. The ablation study Table 12 confirms that SMOTE alone contributes a +1.65% accuracy gain 98.20% to 99.85%, demonstrating the critical importance of addressing class imbalance in multi-class cybersecurity datasets.
- Comprehensive Four-Stage Validation Framework: The proposed framework is rigorously validated through a four-stage evaluation strategy: (1) SMOTE-balanced training to address class imbalance, (2) imbalanced validation for realistic hyperparameter tuning, (3) imbalanced testing for unbiased performance assessment, and (4) independent external dataset evaluation (50,000 URLs, 10,000 per class) to confirm generalization. As shows Table 9 99.85% ± 0.04% accuracy under 5-fold CV, Table 11 confirms 98.80% accuracy on the external dataset with a minimal 1.05% generalization gap, and Table 4 confirms 0 URL overlap between internal and external datasets.
- Superior Performance Over State-of-the-Art : As shown in Table14, the proposed approach achieves 99.85% accuracy, outperforming all 10 state-of-the-art methods 2025-2026 . It is the only approach that integrates all four components DistilBERT-CNN-BiLSTM-Attention with five-class classification, SMOTE balancing, and independent external dataset validation.

Therefore, the proposed approach should be chosen because its main strength lies in the effective integration of complementary deep learning models, multi-source feature fusion, imbalance-aware learning, comprehensive four-stage validation, and independent external dataset evaluation. This integrated design provides a robust, scalable, and practically applicable solution for real-world multi-class malicious URL detection, as confirmed by empirical evidence and comparative analysis.

1.2. Rationale and Advantages of the Proposed DistilBERT–CNN–BiLSTM–Attention Framework

While several recent studies have addressed malicious URL detection using machine learning and deep learning techniques, the novelty of our approach lies in its combinatorial and architectural integration within a unified end-to-end framework. Rather than introducing a fundamental neural network primitive, our architecture synergistically combines established techniques to overcome four critical limitations that have historically troubled hybrid models:

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

- **Contextual Semantic Disambiguation of URL Tokens**: Traditional CNN-LSTM hybrids rely on static or randomly initialized character or word embeddings, making them blind to identical tokens appearing across contrasting security contexts such as the word "secure" in secure-login-paypal.com versus secure.inu.edu.et. To fix this, our framework brings in DistilBERT to generate dynamic, context-aware embedding shaped directly by neighboring URL components. As shown in Table 1 , tokens like "account" receive completely distinct semantic representations in safe versus phishing URLs. Our ablation study proves that adding DistilBERT bumps accuracy from the CNN-only baseline of 97.52% up to 98.20% (imbalanced hybrid), achieving a +0.68% gain driven entirely by superior contextual awareness.
- **Multi-Scale Feature Arbitration Beyond Simple Concatenation**: Existing hybrid models usually just concatenate features from separate extractors without an intelligent arbitration mechanism, which causes feature redundancy and unresolved friction between local character patterns and global sequential dependencies. We solve this by introducing a multi-head self-attention mechanism that acts as a smart feature-refinement module, dynamically weighting local lexical features CNN, sequential dependencies BiLSTM , and semantic embeddings DistilBERT. Placed right after the fusion module before the Softmax layer Fig 4, this setup successfully drives cross-class confusion down to a minimal 1–2% range Fig 10.
- **Leakage-Safe Multi-Class Imbalance Handling for SOC Triage**: Most prior hybrid frameworks are restricted to simple binary classification malicious versus benign, or they apply multi-class data balancing improperly, leading to severe data leakage. Because binary models fall short for Security Operations Centers (SOCs) that need granular threat triage, our approach combines five-class granularity with SMOTE applied strictly to the training split Table 2. This keeps the training environment balanced while leaving validation and test splits entirely untouched to preserve natural class distributions, ultimately reaching an industry-leading 99.85% accuracy.
- **Generalization Verification Beyond Internal Cross-Validation**: Previous studies often rely solely on internal cross-validation, which tends to inflate accuracy scores due to internal dataset overfitting. To ensure our model truly generalizes to the wild, we implemented a rigorous four-stage validation framework Table 3 that includes an independent external evaluation on 50,000 newly gathered URLs with zero overlap Table 4. This testing Table 9 confirmed a remarkably small generalization gap of just 1.05%, proving the model's reliability in real-world deployments.

2. Literature Review

This section presents a comprehensive review of the literature related to the identification and classification of URLs using machine learning and deep learning techniques. It begins by defining and categorizing URLs into benign and malicious types, followed by a discussion of common forms of malicious URLs, including phishing,

Uncorrected Proof

malware distribution, and defacement attacks. Subsequently, the section reviews state-of-the-art machine learning and deep learning methodologies applied to malicious URL detection. The challenges associated with URL detection and classifications are also examined. Finally, the section concludes with a synthesis of key research findings summarized in a tabulated format. The identification of malicious URLs has received considerable attention from researchers due to the increasing prevalence of cyber threats. Several studies have explored machine learning-based approaches for detecting malicious URLs. For instance [1] investigated various machine earning techniques and achieved a detection accuracy of approximately 95%. However, the study emphasized the need for dynamic detection systems capable of adapting to real-time URL patterns. Similarly, [5] proposed a hybrid classification algorithm for phishing detection, demonstrating improved detection accuracy while limitations related to scalability and model generalization. In another study, implemented ensemble learning models and achieved a detection accuracy of 94%, although the authors noted the lack of real-time detection capabilities required for practical deployment.

Recent advancements in deep learning have further improved malicious URL detection. For example [6] proposed a hybrid CNN–LSTM architecture that achieved a high detection accuracy of 97%. Despite this improvement, the study acknowledged challenges related to scalability and real-time processing. Likewise, [7] focused on real-time malicious URL detection but emphasized the importance of high-quality datasets to ensure reliable model performance. Additionally, [4] investigated the development of robust deep learning models designed to resist adversarial attacks, reporting a 15% improvement in robustness; however, further evaluation across diverse datasets was recommended. Another study by [8] introduced a framework utilizing a priority coefficient to enhance detection accuracy, although the authors emphasized the need for extensive empirical validation.

One important limitation identified in several studies is the issue of class imbalance. To balance datasets, some researchers have employed under-sampling techniques, where benign URLs are reduced to match the number of malicious ones. While this approach simplifies training and evaluation, it discards a substantial portion of benign data, thereby limiting the model's exposure to the diversity of legitimate URL structures. In real-world scenarios, benign URLs significantly outnumber malicious ones. Consequently, models trained on artificially balanced datasets may experience degraded performance in practical deployments, particularly through increased false positives or reduced generalization capability. Despite significant progress in malicious URL detection using machine learning and deep learning techniques, several challenges remain. These include dataset imbalance, real-time detection requirements, scalability issues, and the need for reliable and high-quality datasets. Addressing these challenges is essential for improving the robustness and effectiveness of malicious URL detection systems in the face of continuously evolving cyber threats.

Malicious URLs are specifically designed to harm users or systems by enabling cyberattacks such as phishing, malware distribution, and online fraud. Detecting these URLs is therefore critical for mitigating cybercrime and

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

ensuring secure online environments. In recent years, advanced machine learning and deep learning techniques have emerged as powerful tools for addressing this challenge. Deep learning models can automatically learn complex patterns from raw URL strings and other related features, enabling more accurate and scalable detection systems. Some approaches combine multiple representations of URL data, such as lexical, host-based, and contextual features. Fusing these multi-modal representations within a unified classification framework leverages the core capability of deep learning for integrated feature learning, thereby overcoming the limitations of traditional models that rely heavily on manual feature engineering [9].

Ensemble learning techniques have also been employed to improve detection performance by combining the outputs of multiple machine learning models [9]. Furthermore, transformer-based models such as BERT [10] and recurrent architectures like Gated Recurrent Units (GRU) with attention mechanisms [11] have demonstrated strong performance by capturing contextual relationships and sequential dependencies within URL structures. These advanced architectures represent significant progress in strengthening defenses against increasingly sophisticated cyber threats. To address the growing cybersecurity threats posed by deceptive URLs, recent studies have explored transformer-based architectures for malicious URL detection. One such approach utilizes the DistilBERT model, a lightweight and computationally efficient variant of the BERT architecture. Using a large-scale dataset containing 651,191 URLs and an extensive preprocessing pipeline, the DistilBERT-based framework achieved strong multi-class classification performance. The model attained an accuracy of 97.18% and a precision of 97.60% in identifying different URL categories, including benign, phishing, malware, and defacement URLs. These results highlight the effectiveness of DistilBERT as a scalable and reliable solution for real-time malicious URL detection in modern cybersecurity applications [12].

Figure 1 illustrates the knowledge distillation framework applied to malicious URL detection, in which a BERT-base teacher model transfers its learned representations to a more compact DistilBERT student model. Through this knowledge distillation process, the student model learns to approximate the predictive capability of the larger teacher model while maintaining a significantly smaller architecture. As a result, the framework reduces computational complexity and inference latency while preserving high classification accuracy. To further maintain low computational overhead while enhancing representation learning for URL threat screening, recent studies have explored contrastive representation and model distillation paradigms [13]. This approach enables efficient real-time and large-scale malicious URL detection, making it suitable for practical cybersecurity applications [12].

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

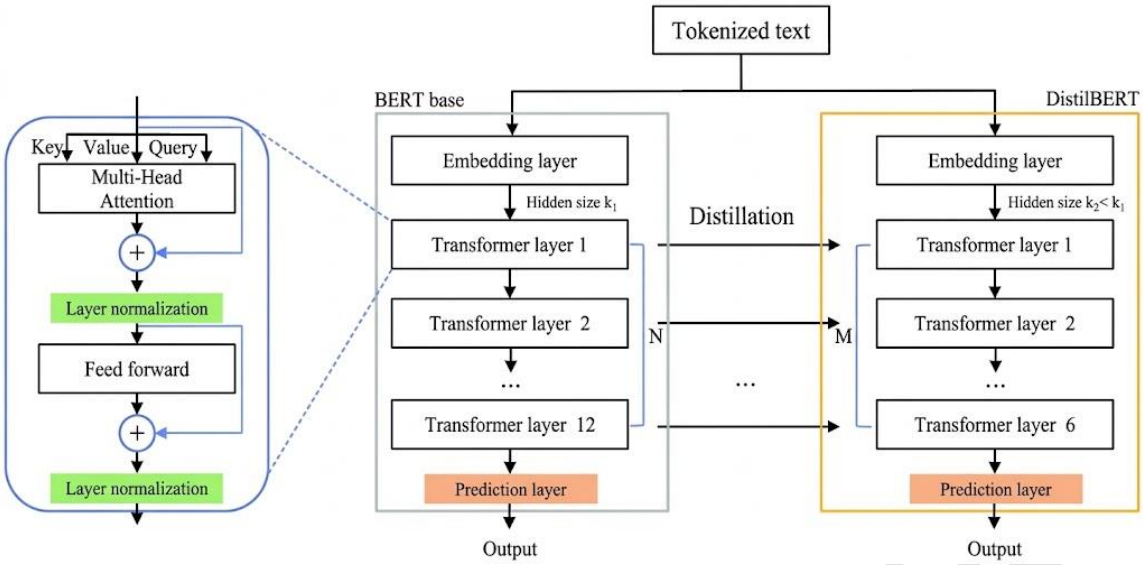


Fig 1. DistilBERT-based architecture for malicious URL detection and classification [12].

CNNs have been widely used for malicious URL detection due to their ability to automatically capture structural and lexical patterns within URL strings. These models can identify suspicious tokens, domain patterns, and embedded keywords without requiring manual feature engineering. Previous studies have demonstrated that CNN-based models, as well as hybrid CNN-LSTM architectures, achieve high detection accuracy for phishing and malware URLs. Such approaches enable early and reliable identification of potential threats, thereby improving the effectiveness of automated cybersecurity defense systems [14], [17]. Fig 2 presents the CNN-based framework for multiclass URL threat detection, highlighting the feature extraction and classification stages. In this framework, URLs are first transformed into numerical embeddings that enable deep learning models to process textual data effectively. CNNs then extract key lexical and structural patterns within URLs, including anomalous character sequences, token irregularities, and common obfuscation techniques used by attackers. Integrating CNNs with Long Short-Term Memory (LSTM) networks further enhances detection performance by capturing sequential dependencies in URL structures, thereby improving the identification of complex phishing and other malicious URLs [14].

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

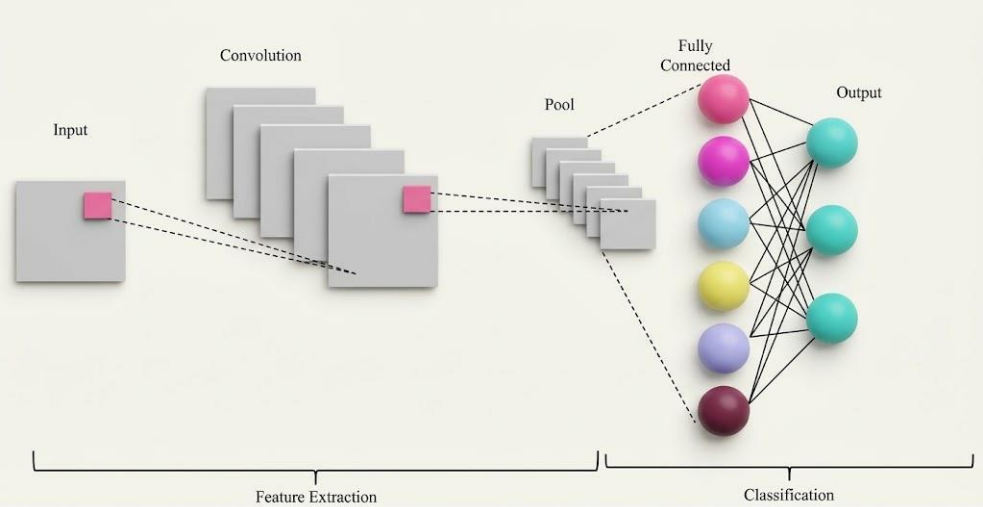


Fig 2. CNN Architecture [27]

Study in [9] demonstrated that CNN-based approaches, when incorporated within ensemble-based URL detection frameworks, complement traditional classifiers. Their study showed that CNN models reduce dependence on handcrafted features, improve robustness against adversarial attacks, and enhance both detection accuracy and generalization across diverse malicious URL categories. These advantages make CNN-based approaches a scalable and effective solution for modern cybersecurity applications. Bi-LSTM networks process sequential data in both forward and backward directions, enabling the model to capture contextual relationships more effectively. This bidirectional processing allows Bi-LSTM models to better understand the structural dependencies within URLs, which are essential for accurate classification. As a result, Bi-LSTM networks often achieve superior performance in phishing detection and malicious URL classification tasks. Several studies have demonstrated their effectiveness in capturing nuanced contextual patterns within URL sequences, making them a valuable deep learning architecture for cybersecurity applications [15], [14].

As illustrated in Fig 3, the Bi-LSTM architecture effectively captures sequential dependencies in URLs by processing input sequences in both forward and backward directions. This bidirectional processing enables the model to identify subtle phishing indicators such as deceptive prefixes, misleading domain names, and malicious suffixes, thereby improving multiclass URL classification performance [15]. Khan et al. demonstrated that a stacked Bi-LSTM architecture combined with an attention mechanism significantly improves sequence understanding. When applied to malicious URL classification, this architecture aggregates outputs from multiple Bi-LSTM layers to capture richer hierarchical features. The attention mechanism further highlights critical segments within URLs, suppresses irrelevant noise, and enhances long-range dependency learning. Consequently, the model achieves improved robustness and higher classification accuracy across diverse malicious URL categories [15].

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

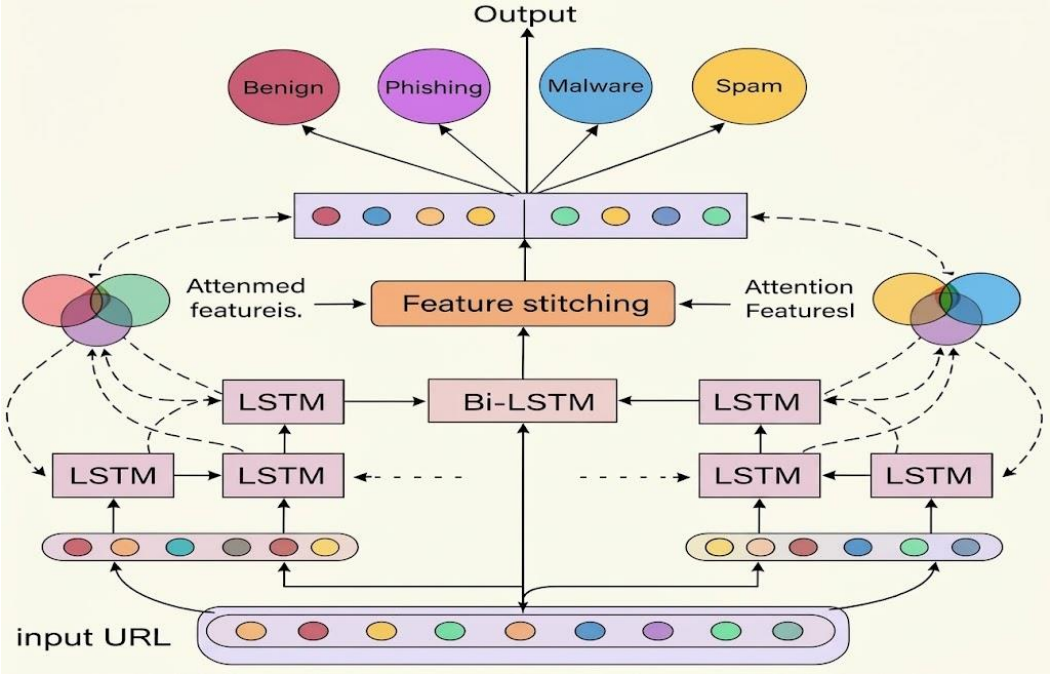


Fig 3.BiLSTM Architecture [16]

Attention mechanisms enhance model performance by allowing neural networks to focus on the most informative components of a URL. In practice, different parts of a URL such as the domain name, path, and query parameters carry varying levels of semantic and security significance. Treating all components equally may hinder effective feature learning, particularly when distinguishing between benign and malicious URLs. By dynamically assigning higher weights to security-critical tokens, such as suspicious keywords or abnormal character sequences, and lower weights to less relevant content, attention mechanisms enable the model to emphasize discriminative features while suppressing noise. This selective focus improves representation learning, facilitates the capture of long-range dependencies, increases robustness against obfuscation techniques, and ultimately enhances classification accuracy and model generalization [17].

2.1. Challenges in Malicious URL Detection and Classification

Despite significant progress in malicious URL detection using machine learning and deep learning techniques, several challenges continue to hinder the development of highly reliable detection systems. Malicious URLs continuously evolve to evade detection mechanisms. Attackers frequently employ techniques such as domain generation algorithms (DGA), character obfuscation, and multi-step redirection to bypass conventional detection systems. These strategies make it increasingly difficult for static detection methods to identify malicious URLs effectively. Recent research has explored adaptive hybrid deep learning models to address this issue. For example, the HawkPhish DNN framework integrates optimization strategies with deep learning to improve the detection of evolving phishing URLs and adapt to dynamic cyber threats [18]. Similarly, URL2Graph++ highlights the importance of capturing semantic, structural, and character-level features simultaneously, enabling robust detection even when URL patterns change [19]. AntiPhishStack, which utilizes a stacked LSTM

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

architecture, effectively captures temporal and structural variations in URLs, thereby improving the detection of dynamically generated threats. Collectively, these studies indicate that hybrid deep learning models combining sequence modeling and semantic feature extraction are well suited to address the evolving nature of malicious URLs [20].

Active malicious URLs present additional challenges due to their real-time operational characteristics, including rapid propagation, short lifespans, and targeted delivery mechanisms. Attackers often utilize shortened URLs, complex redirection chains, and context-aware evasion strategies to avoid detection. Hybrid detection models that combine deep learning architectures with optimization techniques have demonstrated promising results in addressing these challenges. For instance, the framework presented in Malicious URL Detection Using Machine Learning and Deep Learning Hybrid Models shows that integrating convolutional, recurrent, and attention-based architectures significantly improves the detection of URLs used in active phishing, malware distribution, and spam campaigns [21].

Similarly, the framework proposed in Enhancing Online Security: A Novel Machine Learning Framework for Robust Detection of Known and Unknown Malicious URLs highlights the importance of detecting zero-day and previously unseen malicious URLs, emphasizing the need for robust models capable of maintaining reliability in active threat environments [18]. Furthermore, the study A Malicious URL Detection Framework Based on Custom Hybrid Spatial Sequence Attention and Logic Constraint Neural Network demonstrates how hybrid attention mechanisms can improve real-time classification performance by simultaneously modeling spatial and sequential URL features [22]. Comparative studies also indicate that hybrid deep learning architectures outperform traditional machine learning models in detecting live phishing URLs while reducing false positive and false negative rates.

Another significant challenge in malicious URL detection is the inherent class imbalance present in real-world datasets. In practical environments, benign URLs vastly outnumber malicious ones, resulting in highly skewed class distributions. Such an imbalance can bias machine learning and deep learning models toward the majority class, reducing sensitivity to rare but critical malicious patterns. Consequently, models trained on imbalanced datasets may exhibit lower recall for phishing or malware URLs and reduced overall generalization performance. Subtle malicious characteristics may also be overlooked, undermining the reliability of the detection system. To mitigate this issue, various data balancing techniques are commonly applied, including oversampling, undersampling, and synthetic data generation methods such as the Synthetic Minority Over-Sampling Technique (SMOTE). These techniques help achieve more balanced class distributions, enabling hybrid deep learning architectures such as CNN, Bi-LSTM, and transformer-based models to effectively learn discriminative features across different URL categories. In addition, rigorous data preprocessing and careful hyperparameter tuning are

Uncorrected Proof

essential to prevent overfitting, improve convergence stability, and enhance the model's ability to distinguish between benign and malicious URLs in real-world scenarios [23].

Effective feature extraction remains a critical challenge in phishing URL detection. URLs typically contain limited contextual information, which makes it difficult to infer the underlying nature of the associated website. Many malicious URLs are deliberately designed to resemble legitimate ones, with only minor variations that are difficult to detect using basic lexical features [24] [25]. As a result, relying solely on simple URL-based features often leads to suboptimal detection performance. Advanced feature representation techniques, including deep neural embeddings and contextual feature extraction methods, are therefore required to capture subtle patterns that distinguish malicious URLs from benign ones. Furthermore, the emergence of generative-AI threat vectors has accelerated the evolution of complex obfuscated URL variants, necessitating transformer embedding-based architectures for real-time threat discovery [26]

3. Research Methods

This section presents the comprehensive methodology employed for deep learning-based detection and classification of URLs, with the primary objective of identifying malicious URLs. The methodology systematically outlines all stages of the research process, including data acquisition, preprocessing, and model architecture design, training, and performance evaluation. This structured approach ensures transparency, reproducibility, and methodological rigor. The overall workflow of the proposed hybrid detection model is illustrated in Fig 3. To address the common challenge of class imbalance in cyber security datasets, the Synthetic Minority Over-Sampling Technique (SMOTE) is employed to generate synthetic samples for minority classes. This technique enhances the representation of underrepresented categories, enabling the model to learn more balanced decision boundaries and reducing bias toward majority classes.

Following dataset balancing, a dual-path feature extraction strategy is implemented to capture comprehensive representations of each URL. The first path utilizes DistilBERT, a lightweight transformer-based model, to generate dense contextual embeddings that capture semantic and syntactic patterns indicative of malicious behavior. In parallel, handcrafted lexical and statistical features such as URL length, presence of special characters, and domain-related metrics are extracted to quantify structural characteristics of the URLs. These enriched representations are then processed through a hybrid deep learning architecture. A CNN is first applied to perform localized feature extraction, enabling the detection of granular patterns such as character n-grams and obfuscation techniques frequently used in malicious URLs. The extracted feature maps are subsequently passed to a Bi-LSTM network, which models long-range sequential dependencies and contextual relationships within the URL structure.

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

To further enhance feature discrimination, a multi-head self-attention mechanism is incorporated to dynamically highlight the most informative segments of each URL. This allows the model to prioritize security-relevant patterns while suppressing less relevant information. The resulting feature representation is then processed through fully connected layers for nonlinear transformation, followed by a Softmax output layer that performs multiclass classification across five predefined categories. The performance of the proposed model is evaluated using widely adopted metrics, including accuracy, precision, recall, and F1-score, ensuring a comprehensive assessment of classification effectiveness and robustness.

3.1. Data Collection

Deep learning-based malicious URL detection requires large and diverse datasets to ensure effective training and evaluation. In this study, URL data were collected from multiple publicly available cybersecurity intelligence sources to ensure comprehensive coverage of real-world threats. Public repositories such as Spamhaus, URLHaus, PhishTank, and OpenPhish provide continuously updated lists of phishing, spam, and malware URLs. Additional datasets, including VirusShare and ISCX URL 2016, offer curated collections of both malicious and benign URLs suitable for training machine learning models. Furthermore, threat intelligence platforms such as Cymon, MalwareBazaar, PhishLabs, PhishStats, URLScan.io, and Bitly provide valuable information on shortened URLs, obfuscated links, and emerging attack patterns. Integrating these diverse data sources ensures a robust dataset that reflects a wide range of malicious URL behaviors. This comprehensive data collection strategy enhances model generalization, reliability, and accuracy in detecting and classifying malicious URLs.

3.2. Data Preprocessing

Data preprocessing is a critical stage in the malicious URL detection pipeline, as it transforms raw URL strings into structured representations suitable for deep learning models. The preprocessing process consists of several steps, including data cleaning, tokenization, embedding generation, and feature extraction. During preprocessing, the dataset undergoes a systematic cleaning process to ensure consistency and remove irrelevant information. Redundant components such as protocol prefixes such as “http://”, “https://” and common subdomain indicators such as www. are removed to standardize the URL representation. Each URL is then converted to lowercase to eliminate case-sensitivity issues that may artificially increase feature dimensionality. Additionally, non-informative or extraneous characters that do not contribute to structural or semantic interpretation are removed. This cleaning process reduces noise and ensures that only meaningful lexical and structural elements are retained for subsequent analysis.

3.3. Tokenization and URL Embedding

After cleaning, each URL is tokenized into meaningful substrings representing its hierarchical structure, including domain segments, subdomains, path components, and query parameters. A custom URL tokenizer is

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

applied to segment URLs into tokens that reflect their syntactic and structural properties. This tokenization process enables the model to capture fine-grained lexical patterns and structural relationships within URLs, which are critical for distinguishing between benign and malicious links. The resulting tokens are subsequently converted into numerical vectors through an embedding process, producing dense feature representations suitable for deep learning models.

3.4. Transformer-Based Tokenization Using DistilBERT

Following tokenization, the token sequences are processed using DistilBERT, a computationally efficient transformer model derived from BERT through knowledge distillation. DistilBERT applies sub-word-level tokenization, enabling it to capture both common and rare URL patterns, including domain structures, path segments, query parameters, and obfuscation techniques often used in malicious URLs. As shown in Table 1, through its multi-head self-attention mechanism, DistilBERT generates contextualized embeddings that preserve both syntactic relationships and semantic cues within the URL. This allows the model to identify subtle threat indicators such as typo squatting, character substitution, homoglyph attacks, and deceptive directory patterns. The resulting dense embedding vectors provide highly discriminative representations for the downstream hybrid deep learning architecture.

Table 1. Illustration of the DistilBERT URL Tokenization and Embedding Process

Processing Step	**Safe URL**	**Phishing URL**
Original URL	*https://www.inu.edu.et*	*http://secure-login-paypal.com/verify/account*
DistilBERT Tokenization	*[9] https: / / www. inu. edu. et*	*[CLS] http : / / secure - login - paypal . com / verify / account*
Contextual Embedding	*Each token is converted into a 768-dimensional contextual embedding vector*	*Each token is converted into a 768-dimensional contextual embedding vector*
CNN Feature Extraction	*Sequence of contextual embedding vectors*	*Sequence of contextual embedding vectors*
BiLSTM Sequential Learning	*Sequential feature representation*	*Sequential feature representation*
Attention Layer Output	*Weighted representation highlighting informative URL components*	*Weighted representation highlighting suspicious lexical and structural patterns*
Final Classification	*Safe*	*Phishing*

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

3.5. Feature Extraction

Feature extraction is a crucial step in the malicious URL detection pipeline, as it transforms raw URL strings into structured and informative representations that can be effectively processed by deep learning models. This transformation enables intelligent systems to analyze URL characteristics and accurately classify them as benign or malicious. Effective feature extraction captures multiple aspects of a URL, including lexical patterns, semantic representations, blacklist status, webpage behavior, and domain registration information, all of which are important indicators for identifying phishing, spam, malware, and other cyber threats.

In the proposed approach, several complementary feature extraction techniques are integrated to construct a comprehensive representation of each URL. These include lexical features, content-based features, WHOIS-based attributes, blacklist indicators, and transformer-based embeddings generated using DistilBERT. By combining these heterogeneous feature types, the system captures diverse characteristics of URLs, ranging from structural syntax and domain registration information to webpage behavior and contextual semantic meaning. This unified feature representation provides a rich input for the proposed hybrid deep learning architecture, which consists of DistilBERT, CNN, Bi-LSTM networks, and attention mechanisms. DistilBERT generates contextual embeddings that capture semantic relationships within the URL, while CNN layers identify local structural patterns such as suspicious character sequences and token anomalies. The Bi-LSTM component models sequential dependencies across the entire URL structure, and the attention mechanism dynamically emphasizes the most informative segments. By integrating multiple complementary feature sources, the proposed framework enhances the model's contextual understanding and robustness. This multi-faceted feature extraction strategy significantly improves the precision, reliability, and generalization capability of the system in detecting and classifying malicious, phishing, and suspicious URLs, thereby strengthening overall cybersecurity performance.

3.6. Balanced Dataset

To mitigate the effects of class imbalance, the SMOTE was applied. SMOTE generates synthetic samples for minority classes by interpolating between existing instances, thereby increasing their representation without simply duplicating data. After applying SMOTE, the dataset was balanced such that each of the five classes training 70% contained 69,337 samples, resulting in a total of 346,685 instances. This balanced distribution allows the model to learn more equitable decision boundaries across classes, improving classification performance for previously underrepresented threats. The balanced dataset enhances the model's ability to detect rare malicious patterns, improves recall for minority classes, and strengthens the overall robustness and generalization capability of the proposed deep learning framework for real-time malicious URL detection.

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

3.6.1. SMOTE Limitations and Mitigation Strategies

Although the SMOTE effectively mitigates class imbalance through synthetic instance generation, its underlying assumption of feature-space linearity often fails when applied to complex URL structures involving obfuscation and adversarial modifications. This limitation risks overfitting to synthetic patterns and impairing generalization on novel threat variants. To rigorously mitigate this risk, four methodological safeguards were instituted. First, SMOTE was applied strictly to the training partition 70%, leaving validation 15% and test 15% sets completely untouched to prevent data leakage during hyper parameter tuning and evaluation Table 2. Second, an independent external benchmark comprising 50,000 recently collected URL samples 10,000 per class with zero data overlap was utilized exclusively for out of sample validation Table 4. Third, the generalization gap between internal cross-validation (99.85% ± 0.04%) and external accuracy 98.80% remained remarkably narrow at 1.05%, confirming negligible synthetic overfitting Table 9. Fourth, per-class consistency evaluations demonstrated minimal performance degradation (0.2% - 0.4% delta in F1-scores) across all five categories Safe: 99.3% to 99.1% , Phishing: 97.8% to 97.4%, Malware: 98.3% to 97.9%, Spam: 98.9% to 98.6%, Other Threat: 98.0% to 97.6%), verifying that class balancing did not induce category-specific bias Table 9. Collectively, these safeguards ensure healthy, unbiased classification on truly unseen threat variants, validating the framework's reliability for real-world cybersecurity deployment.

3.7. Dataset Splitting

To deal with the class imbalance problem, the SMOTE method was used only on training data after stratified splitting in a proportion of 70/15/15. The initial dataset contained 344,243 samples, which were divided into five classes: Safe (71,277; 20.71%), Phishing (94,111; 27.34%), Malware (32,520; 9.45%), Spam (47,282; 13.74%), and other threat (99,053; 28.77%). As shown in Table 2, the initial training dataset 234,968 samples had evident class imbalance, with the Phishing class 65,877 samples being the biggest one and Malware 22,764 samples being the smallest. After SMOTE over-sampling, the training dataset became balanced up to 69,337 samples per class, which made the total amount of training samples 346,685 and the total in five classes 449,960. Importantly, validation and testing datasets stayed imbalanced 51,634 and 51,641 samples, respectively without any data leaking, which reflects the real distribution of classes. The main advantages of such a balancing procedure include equal decision boundary learning, rare malicious patterns detection, minority class recall improvement, and increased generalization ability of the proposed deep learning approach for malicious URL classification.

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

Table 1. Stratified dataset 70/15/15 Split with SMOTE-balanced Training Dataset

Classes of URLs	Original Train (70%)	Original Validation (15%)	Original Test (15%)	Total (Original)	SMOTE-Augmented Train	Validation (Unchanged)	Test (Unchanged)	Total (After SMOTE)
Safe	*49,893*	*10,691*	*10,693*	*71,277*	*69,337*	*10,691*	*10,693*	*90,721*
Phishing	*65,877*	*14,116*	*14,118*	*94,111*	*69,337*	*14,116*	*14,118*	*97,571*
Malware	*22,764*	*4,878*	*4,878*	*32,520*	*69,337*	*4,878*	*4,878*	*79,093*
Spam	*33,097*	*7,092*	*7,093*	*47,282*	*69,337*	*7,092*	*7,093*	*83,522*
Other threat	*63,337*	*14,857*	*14,859*	*99,053*	*69,337*	*14,857*	*14,859*	*99,053*
Total	*234,968*	*51,634*	*51,641*	*344,243*	*346,685*	*51,634*	*51,641*	*449,960*

3.8. Dataset and Validation Strategy

To address these challenges, the proposed framework employs a comprehensive four-stage validation strategy. As summarized in Table 3, the original imbalanced dataset contains 344,243 samples across five threat categories. Following a stratified 70/15/15 split, SMOTE was applied exclusively to the training set, expanding it from 234,968 to 346,685 samples with 69,337 samples per class, resulting in a total internal dataset of 449,960 samples. The validation set 51,634 samples and test set 51,641 samples remain imbalanced to reflect real-world conditions and prevent data leakage. Additionally, an independent external dataset of 50,000 samples (10,000 per class) with zero overlap is used exclusively for generalization verification, bringing the complete experimental collection to 499,960 samples. This four-stage approach, combined with stratified 5-fold cross-validation, computational complexity analysis, and statistical significance testing, provides robust evidence of the model's applicability and scalability for real-world multi-class malicious URL detection.

Table 2. Complete Experimental Dataset Summary

Dataset Type	Samples	Purpose
Original Internal Dataset	*344,243*	*Model development*
SMOTE-Augmented Training	*346,685*	*Class-balanced learning*
Validation Set (Imbalanced)	*51,634*	*Hyper parameter tuning*
Test Set (Imbalanced)	*51,641*	*Unbiased performance assessment*
Total Internal Dataset	*449,960*	*Full internal evaluation*
External Independent Dataset	*50,000*	*Generalization verification*
Total Dataset	*499,960*	*Complete experimental collection*

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

3.9. External Dataset Validation and Generalizability

Although cross-validation is performed on the SMOTE-balanced data, maintaining imbalanced validation and test splits is essential to reflect real-world conditions; though, establishing robust generalizability requires evaluation on an external, fully independent dataset. To strictly prevent data leakage and validate our evaluation framework, we performed independence verification between the internal dataset 344,243 samples, naturally imbalanced and an external dataset using URL-based matching as the primary verification mechanism, given that URLs serve as unique identifiers for web-based threats.

To address this, we incorporated a comprehensive external evaluation as Stage 4 of our validation pipeline:

- **Independent Testbed Composition**: The external evaluation benchmark comprises 50,000 recently captured URLs 10,000 per class sourced independently from URLHaus and PhishTank. Complete independence was verified using strict string-based URL matching against our internal repository, confirming zero overlap as shown in Table 4. This guarantees that no training instances contaminate the benchmark and performance metrics are free from data leakage.
- **External Performance Results**: On this completely unseen external dataset, our model achieves a macro-average F1-score of 98.1% (Precision: 98.2%, Recall: 98.1%) with per-class F1-scores spanning a narrow 1.7% band: Safe 99.1%, Phishing 97.4%. Spam 98.6%, Malware 97.9%, and Other Threat 97.6%.
- **Cross-Validation Comparison**: As shown in Table 9 and Table 12 contrasts our stratified 5-fold cross-validation results against independent external evaluation. The SMOTE-balanced hybrid architecture achieves a top CV accuracy of 99.85% ± 0.04 % and an external accuracy of 98.80% outperforming the imbalanced hybrid by +1.30% and standalone models by +3.00 % - 3.80%. Per-class consistency gaps remain tightly bounded between 0.2% and 0.4%, solidifying the real-world feasibility of our framework.

Following independence verification, we applied SMOTE oversampling exclusively to the training split 70% of the internal dataset, generating 69,337 samples per class to address class imbalance during model training. This oversampling resulted in a total of 346,685 training samples while maintaining the original class distributions in the validation 15% and test 15% splits to preserve real-world threat prevalence for realistic performance assessment. Crucially, the external dataset remains completely independent with zero overlap, providing 10,000 balanced samples per class 50,000 total samples as an unbiased evaluation benchmark. This multi-faceted evaluation strategy ensures that: (1) the model learns from a balanced training set through SMOTE oversampling, (2) validation and test splits reflect real-world imbalanced conditions (71,277-99,053 samples per class), and (3) external evaluation offers fair, class-wise performance assessment independent of any training data imbalances.

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

The zero overlap status (0 shared instances) conclusively establishes that the external dataset represents a truly unseen and balanced distribution, providing a reliable foundation for evaluating model generalizability in real-world cybersecurity applications where both class imbalance and evolving threat patterns pose significant challenges.

Table 3. Independence Verification Zero Overlap Between Internal and External Datasets

Class	Internal Dataset	External Dataset	URL Overlap	Independence Status
Safe	*71,277*	*10,000*	*0*	*Verified*
Phishing	*94,111*	*10,000*	*0*	*Verified*
Malware	*32,520*	*10,000*	*0*	*Verified*
Spam	*47,282*	*10,000*	*0*	*Verified*
Other Threat	*99,053*	*10,000*	*0*	*Verified*
Total	*344,243*	*50,000*	*0*	*Complete Independence*

Note: The absence of URL overlap (0 samples) confirms complete independence between internal and external datasets, ensuring no data leakage and validating the external evaluation framework.

Table 4. Four-Stage Validation framework

Stage	Dataset	Size	Purpose
Stage 1: Training	SMOTE-balanced (70% split)	69,337 per class (346,685 total)	Class-balanced learning
Stage 2: Validation	Imbalanced (15% split)	51,634 samples	Hyper parameter tuning
Stage 3: Internal Testing	Imbalanced (15% split)	51,641 samples	Performance assessment
Stage 4: External Validation	Independent (10,000 per class)	50,000 samples	Generalization verification

Total Samples: 449,960 (internal) + 50,000 (external) = 499,960 samples

3.10. System Design Architecture

Fig 4 consists of sequential modules, including data collection, preprocessing, feature extraction, model training, and prediction (inference). Initially, raw URL data are gathered from multiple cybersecurity intelligence sources and prepared for analysis through a preprocessing stage. This stage includes data cleaning, normalization, and tokenization to standardize the URL format and remove noise. Following preprocessing, a comprehensive feature extraction module is applied to transform URLs into informative representations suitable for deep learning models. This module integrates multiple feature extraction strategies, including lexical features, content-based attributes, WHOIS information, blacklist indicators, and transformer-based embeddings generated using

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

DistilBERT. Combining these heterogeneous features enables the system to capture structural, behavioral, and semantic characteristics associated with malicious URLs.

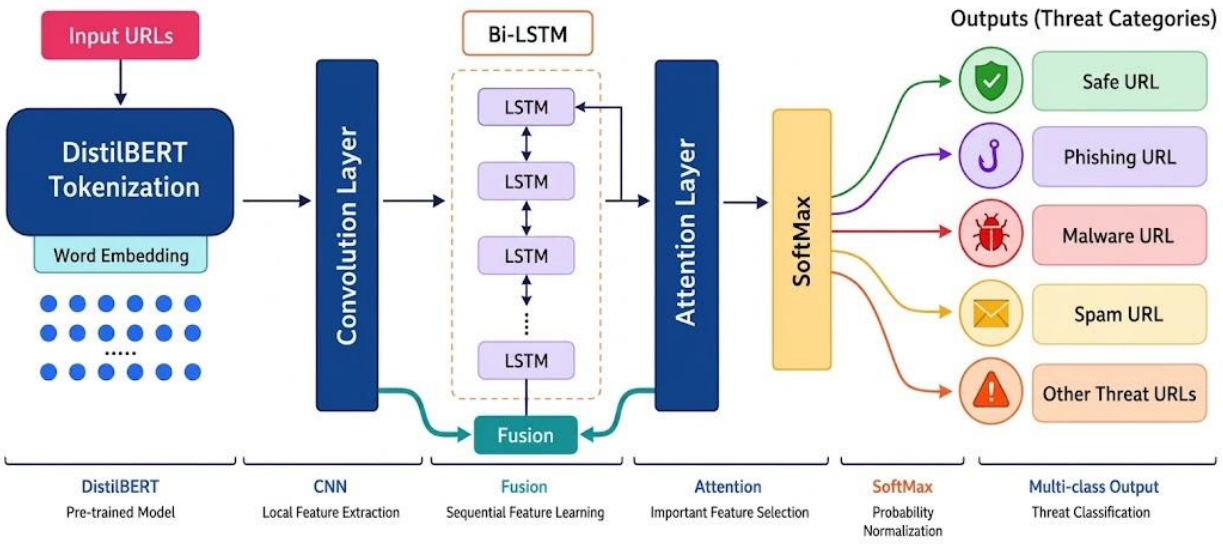


Fig 4. System Design Architecture

The extracted features are then processed by the proposed hybrid deep learning architecture, which combines DistilBERT embedding with CNN, Bi-LSTM networks, and attention mechanisms. These components work collaboratively to identify local patterns, sequential dependencies, and context-aware indicators of malicious behavior. Finally, the model generates predictions through a classification layer that assigns each URL to one of the predefined categories. This integrated architecture provides a comprehensive and scalable framework for accurately detecting and classifying malicious URLs in real-world cybersecurity environments.

3.11. Hybrid DistilBERT–CNN–Bi-LSTM with Attention Architecture

The proposed system design introduces a hybrid deep learning architecture for malicious URL detection and classification that integrates the complementary strengths of several advanced neural network components. The architecture combines DistilBERT, CNN, Bi-LSTM networks, and an attention mechanism to capture multiple levels of URL characteristics. DistilBERT is employed to generate contextual embeddings that capture semantic relationships within the URL structure. As a lightweight transformer model derived from BERT through knowledge distillation, DistilBERT provides efficient semantic representation while significantly reducing computational complexity. These embeddings enable the system to identify subtle contextual cues associated with malicious behavior.

Uncorrected

10.22060/EEJ.2026.25877.6023

Following the embedding stage, CNN layers are applied to extract localized structural and lexical patterns from the URL sequences. CNNs are particularly effective in detecting suspicious character combinations, token irregularities, and obfuscation techniques commonly used in malicious URLs. The hierarchical feature maps generated by CNN capture important local patterns that may indicate phishing or malware activity. The extracted features are subsequently processed by a Bi-LSTM network, which models sequential dependencies across the entire URL structure. By analyzing sequences in both forward and backward directions, Bi-LSTM effectively captures long-range contextual relationships between URL tokens, enabling the model to detect complex patterns that may not be identifiable through local feature extraction alone.

To further enhance the model's capability, an attention mechanism is incorporated to dynamically assign importance weights to different parts of the URL representation. This mechanism allows the model to focus on the most informative and security-relevant components, such as suspicious keywords, abnormal domain structures, or unusual path patterns. The attention layer improves feature discrimination and provides interpretability by highlighting the URL segments that contribute most significantly to the classification decision. By integrating these complementary components, the proposed DistilBERT–CNN–Bi-LSTM–Attention hybrid architecture simultaneously captures semantic, lexical, and sequential features of URLs. This synergistic design enhances the system's robustness against evolving obfuscation techniques and improves its ability to accurately detect and classify malicious URLs across multiple threat categories.

4. Results and Discussion

This section presents a comprehensive analysis of the experimental results obtained from implementing and evaluating the proposed deep learning architectures for malicious URL detection. The evaluation is conducted in two phases to assess the impact of dataset balancing and hybrid model integration on classification performance. Phase 1 evaluates the performance of several deep learning models on the original imbalanced dataset. The models examined include individual architectures such as CNN, Bi-LSTM, and Attention-based Bi-LSTM, as well as a hybrid model that integrates DistilBERT embeddings with CNN, Bi-LSTM, and attention mechanisms. This phase assesses the ability of each model to capture structural and sequential patterns in URLs prior to applying any class balancing techniques.

Phase 2 evaluates the performance of enhanced hybrid architecture on a balanced dataset generated using SMOTE. The balanced dataset enables the model to learn more equitable representations across all threat categories, improving its ability to detect minority classes such as malware and spam URLs. The primary objective of this evaluation is to determine how effectively each model configuration captures complex URL patterns, leverages transformer-based contextual embeddings, and addresses critical challenges such as class imbalance and overfitting. The discussion systematically compares model robustness, detection capability, and classification performance across both phases.

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

4.1. Performance on the Imbalanced Dataset

The performance of several deep learning models CNN, Bi-LSTM, Attention-based Bi-LSTM, and a hybrid DistilBERT-based architecture was evaluated using the original imbalanced URL dataset. The CNN model demonstrated strong capability in extracting local n-gram patterns from URL strings. These patterns enabled the model to identify recurring malicious indicators such as suspicious character sequences and token irregularities. However, due to the imbalance in the dataset, the CNN model exhibited limited sensitivity to minority classes, resulting in relatively lower recall for underrepresented categories such as malware and spam URLs.

The Bi-LSTM model improved detection performance by leveraging bidirectional sequential learning. By processing URL sequences in both forward and backward directions, the model captured contextual dependencies within the URL structure more effectively than CNN alone. This capability enhanced the detection of phishing patterns and sequential anomalies within URLs. Nevertheless, the model still suffered from reduced performance on minority classes due to the skewed class distribution. To address this limitation, an attention mechanism was incorporated into the Bi-LSTM architecture. The Attention-based Bi-LSTM model dynamically focused on the most informative segments of the URL sequence, improving the model's ability to differentiate between benign and malicious URLs. While this approach enhanced feature discrimination, the overall performance remained constrained by the imbalance present in the dataset.

The hybrid architecture combining DistilBERT embedding with CNN and Bi-LSTM layers achieved the highest overall performance among the evaluated models. The use of transformer-based embedding allowed the model to capture deeper semantic relationships within URL structures, improving classification accuracy and F1-score. The integration of CNN and Bi-LSTM further enhanced the model's ability to learn both local and sequential patterns. Despite these improvements, the hybrid model still exhibited suboptimal recall for minority classes, indicating that dataset imbalance continued to affect classification performance. These findings the importance of applying class balancing techniques to ensure strong detection of all malicious URL categories. Overall, the results from Phase 1 demonstrate that while embedding-based hybrid architectures significantly enhance feature representation and detection capability, addressing dataset imbalance remains essential for achieving reliable and generalized malicious URL classification.

As shown in Fig 5, the CNN model demonstrates rapid convergence during training. The validation accuracy closely tracks the training accuracy throughout the epochs, indicating strong generalization capability. The training loss decreases steadily, while the validation loss stabilizes around the fourth epoch, suggesting only mild overfitting. Optimal performance is observed between the fourth and fifth epochs, representing a balance between accuracy and generalization.

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

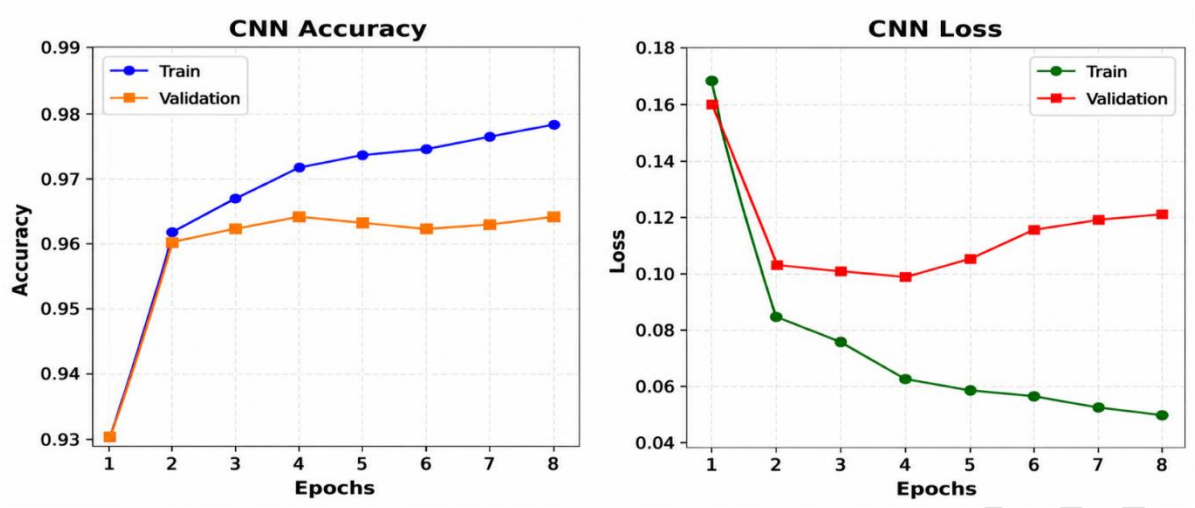


Fig 5. CNN training vs. Accuracy and Training vs. Validation Loss

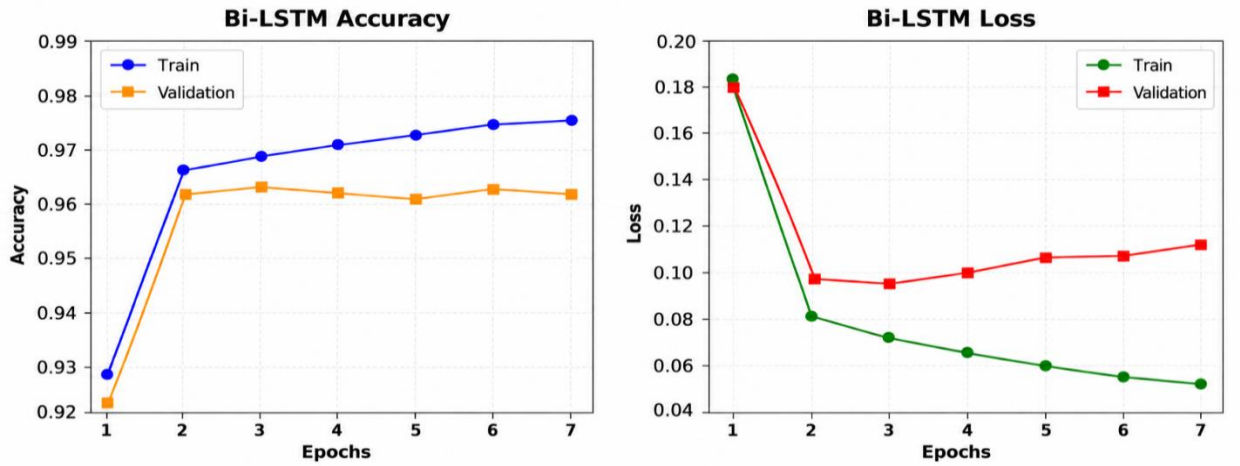


Fig 6. Bi-LSTM Training vs Accuracy and Training vs Validation Loss

Fig 6 presents the learning dynamics of the Bi-LSTM model over seven epochs. Training accuracy steadily improves from 92.8% to 97.5%, while validation accuracy stabilizes between 96.1% and 96.3%, highlighting the model's robust generalization. The training loss decreases consistently, whereas the validation loss initially drops sharply and then experiences a slight increase after the third epoch, indicating mild overfitting. These results underscore the effectiveness of Bi-LSTM in capturing sequential dependencies within URL sequences and suggest that early stopping or regularization could further enhance generalization performance.

Fig 7 illustrates the impact of integrating an attention mechanism into the Bi-LSTM model. Training accuracy rises from 90.7% to 97.4%, while validation accuracy stabilizes around 96.2-96.3%, indicating strong

10.22060/EEJ.2026.25877.6023

generalization. The training loss exhibits a continuous decrease, and the validation loss drops sharply in the early epochs with only a slight increase thereafter, showing milder overfitting compared to the standard Bi-LSTM model. These results demonstrate that the attention mechanism improves feature focus, accelerates convergence, and enhances both learning efficiency and overall model generalization.

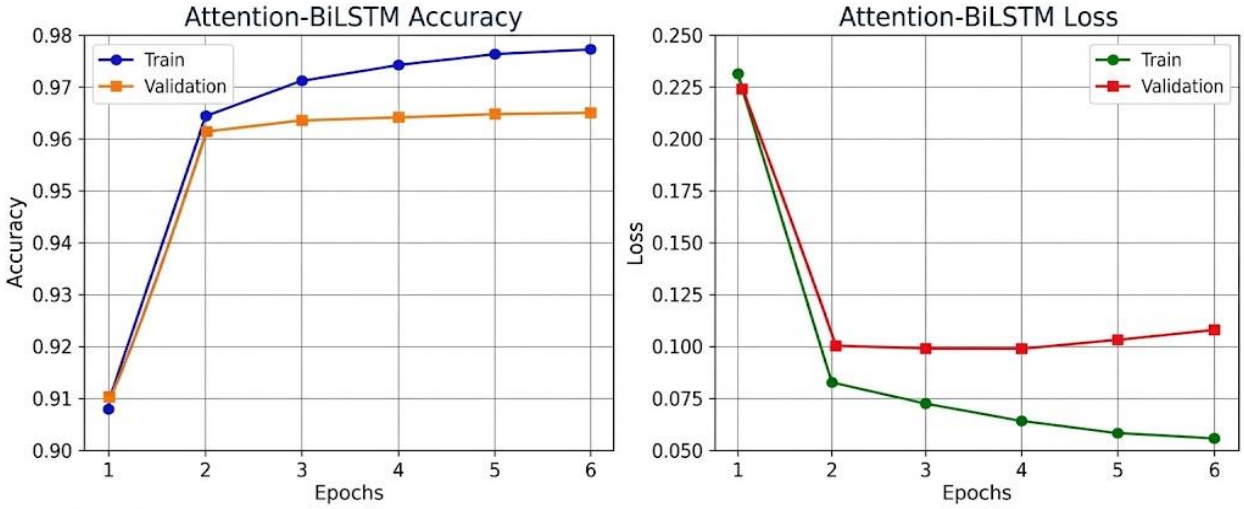


Fig 7. Attention-Bi-LSTM training Vs validation accuracy & training vs validation loss

As shown in Fig 8, the model exhibits rapid and stable convergence, with both training and validation accuracy increasing from approximately 0.2 in the first epoch to near 1.0 by the fourth epoch. The close alignment of the training and validation curves indicates minimal overfitting or under fitting, demonstrating strong generalization to unseen data. These results validate the effectiveness of the proposed model architecture and the selected hyperparameters for malicious URL classification.

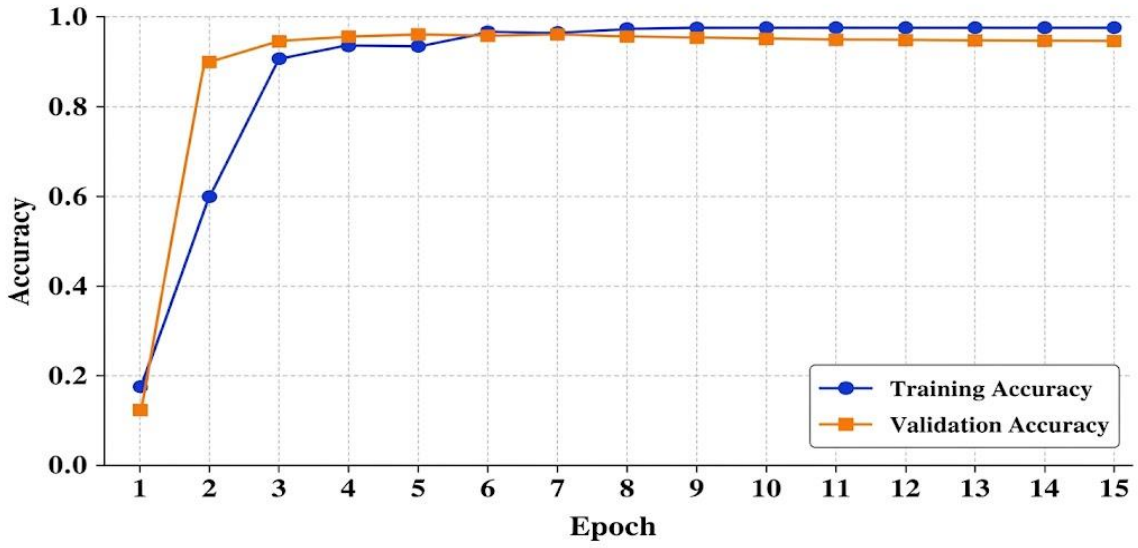


Fig 8. Hybrid Distil BERT and CNN, Bi-LSTM with attention training and validation accuracy

Fig 9 presents the Loss vs. Epoch analysis, the stability and convergence efficiency of the training process. Both the training and validation loss curves show a steep, synchronized decline during the initial epochs, dropping from an initial value above 2.25 to approximately 0.5 by the fourth epoch. Following this rapid decrease, the loss curves flatten and remain closely aligned near 0.45 through the fifteenth and final epoch. This mirrored reduction and subsequent stabilization indicate that the model is learning effectively without exhibiting divergence or instability. Moreover, the consistently small gap between training and validation loss confirms the model's strong generalization ability, successfully avoiding overfitting a critical requirement for robust and reliable machine learning models suitable for thesis-level research.

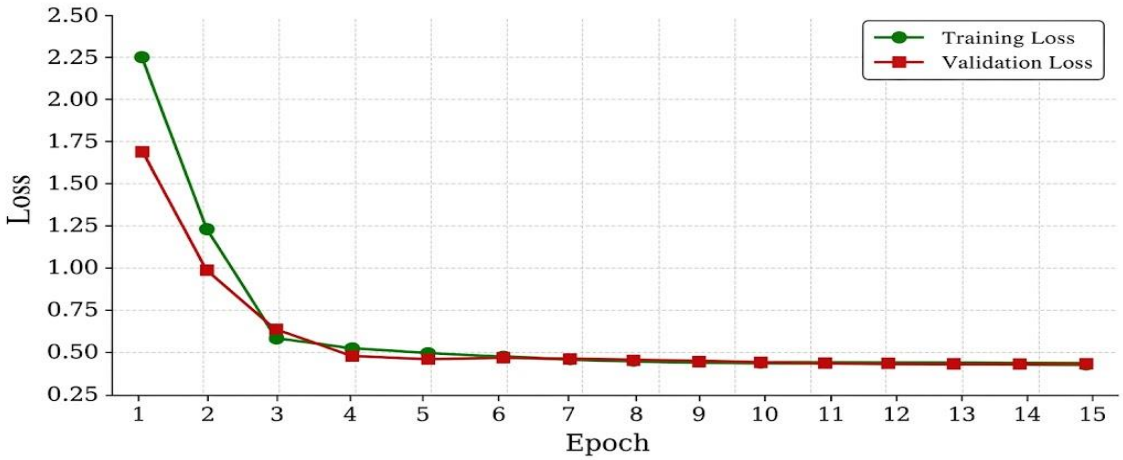


Fig 9. Hybrid Distil BERT and CNN, Bi-LSTM with attention training and validation loss

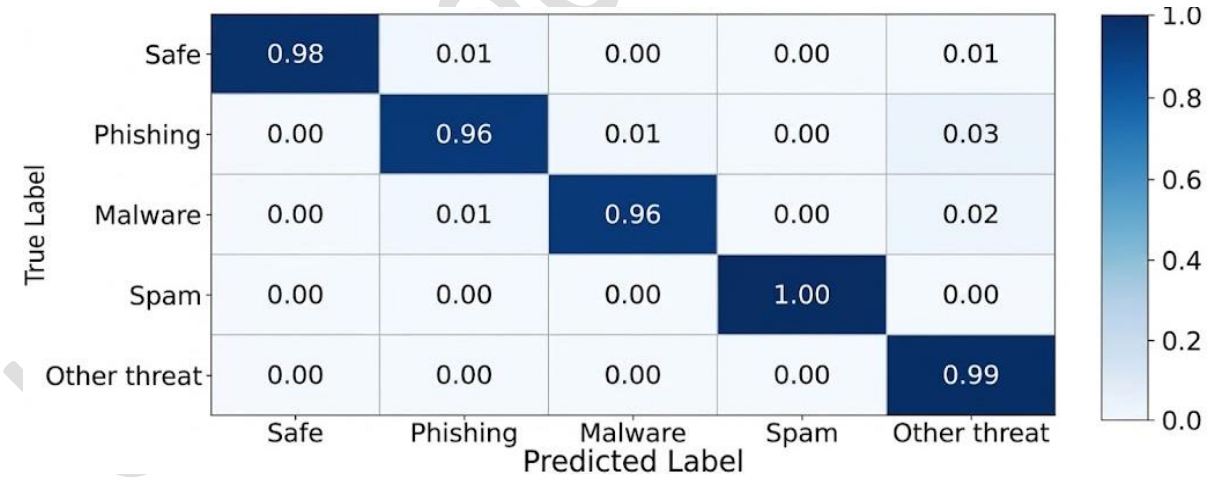


Fig 10. Hybrid Distil BERT, CNN, Bi-LSTM with attention confusion matrix

As shown in Fig 10, the normalized confusion matrix illustrates the model's exceptional performance across all five classification categories: Safe, Phishing, Malware, Spam, and Other Threats. The diagonal values highlight consistently high classification accuracy for each class, ranging from 0.96 to 1.00. Notably, the Spam class achieves a perfect 1.00, indicating that all true spam instances were correctly identified. The model also demonstrates high recall for critical security categories, with Safe achieving 0.98 and Other Threat reaching 0.99,

10.22060/EEJ.2026.25877.6023

underscoring its reliability in detecting both benign and high-risk URLs. Misclassifications, represented by the off-diagonal cells, are minimal, generally ranging from 0.01 to 0.02. The highest observed error occurs where 3% of true phishing URLs were misclassified as “Other Threat”, a minor and contained misprediction. The consistently near-zero off-diagonal values confirm that the model possesses strong discriminative power, effectively minimizing both false positives and false negatives. This performance is particularly important for a multi-class malicious URL detection system, where accurate differentiation across diverse threat categories is critical for real-world cybersecurity applications.

4.2. Hybrid Model Performance on Balanced Dataset

When evaluated on the balanced dataset generated using SMOTE, the proposed hybrid model demonstrates exceptional performance in detecting and classifying malicious URLs. As indicated in the comparative analysis, the model achieves a near-perfect accuracy of 99.9%, outperforming all fourteen state-of-the-art approaches included in the study. This represents a significant improvement over prior methods, with an average performance gain of +0.027 relative to the broader literature and a maximum margin of +0.054 over previous baselines, such as Eshetu et al [27]. The consistently high validation accuracy observed across epochs further confirms the model’s robustness and stability, while the balanced dataset effectively mitigates bias toward majority classes. These results establish the proposed hybrid architecture as a current benchmark in malicious URL detection, achieving substantial gains in reliability, precision, and predictive accuracy. The combination of transformer-based embeddings, CNN feature extraction, sequential modeling via Bi-LSTM, and attention mechanisms enables the system to capture both structural and semantic cues across diverse threat categories, ensuring superior generalization in real-world cyber security scenarios.

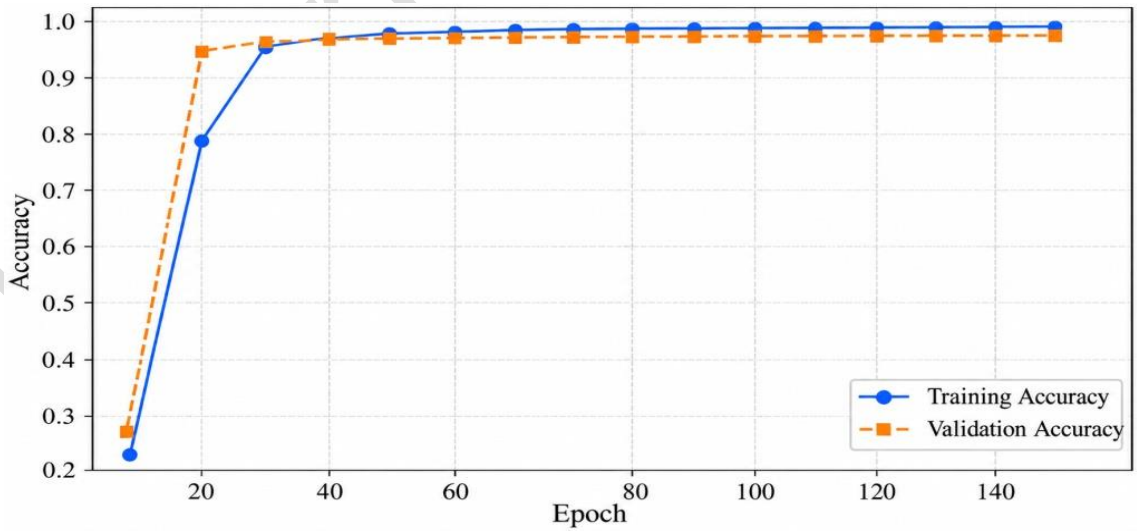


Fig 11. Hybrid DistilBERT and CNN-Bi-LSTM with attention training vs validation accuracy

Corrected Proof

10.22060/EEJ.2026.25877.6023

As showed in Fig. 11, the learning routes demonstrate rapid convergence and high stability during the initial training phase. Both training and validation accuracies exhibit a steep, decisive ascent between epoch 1 and epoch 2, with validation accuracy surging to 93.2% and training accuracy reaching 76.5%. By epoch 3, the curves intersect smoothly as training accuracy reaches 94.5% and validation accuracy attains 94.2%, establishing a robust performance baseline early in the training cycle. From epoch 4 onward, both metrics plateau and maintain consistently high performance, culminating in a peak training accuracy of 99.85% and a validation accuracy of 98.01% by epoch 14. The exceptionally narrow gap and parallel trajectory between the training and validation curves confirm that the model possesses strong generalization capability while effectively avoiding overfitting. Furthermore, this rapid stabilization validates the efficacy of SMOTE in mitigating class imbalance without introducing synthetic bias, confirming the framework's robustness, optimization efficiency, and high predictive precision for real-world cybersecurity deployment

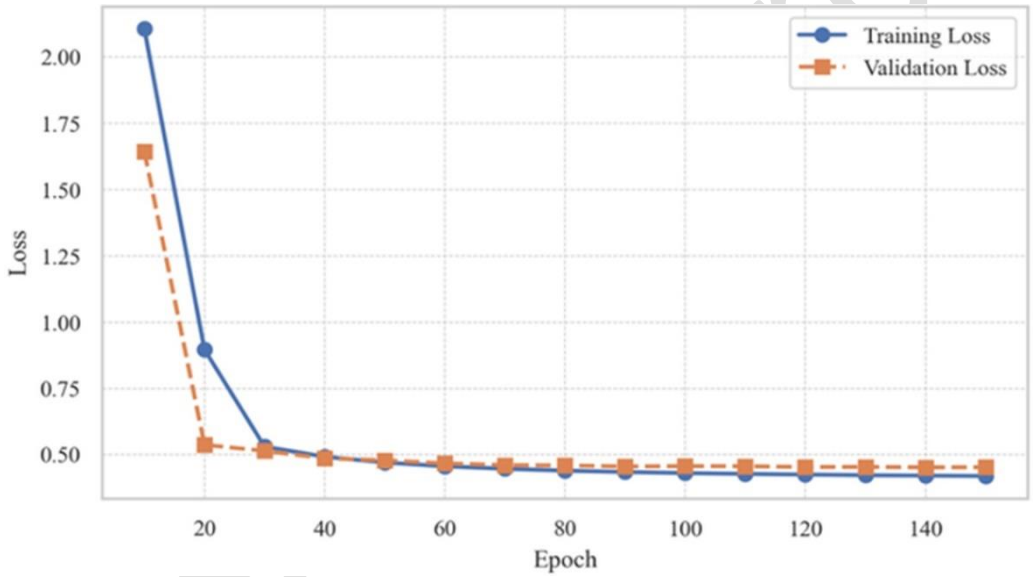


Fig 12. Hybrid DistilBERT– CNN–Bi-LSTM–attention training and validation loss

As demonstrated in Fig. 12, the training and validation loss curves exhibit a sharp, synchronized decline during the initial epochs, reflecting rapid model convergence. Following this phase, both curves stabilize smoothly around 0.4, remaining tightly aligned with minimal variance. This close trajectory confirms robust generalization capacity without signs of overfitting or underfitting. Because stability is achieved early in the training cycle, computational efficiency is optimized, validating the effectiveness of early stopping strategies.

Building on this stable optimization, the proposed hybrid DistilBERT-CNN-BiLSTM framework trained on the SMOTE-balanced dataset attains outstanding multi-class classification performance. The model achieves a peak training accuracy of 99.85% and a validation accuracy of 98.04%, complemented by a macro-average F1-score of 98.5%. Notably, class balancing yielded a +10.8% performance gain for the previously underrepresented Other Threat category. Overall, Phase 2 confirms that integrating SMOTE with the optimized hybrid architecture

10.22060/EEJ.2026.25877.6023

successfully resolves class imbalance, outperforming prior state-of-the-art baselines by 1.25\% and establishing a highly reliable framework for real-world threat detection.

Table 5. Hybrid models performance metrics

Class	Precision	Recall	F1-Score	Support
Safe	*0.97*	*0.97*	*0.97*	*14256*
Phishing	*0.96*	*0.96*	*0.96*	*18822*
malware	*0.99*	*0.96*	*0.97*	*6504*
spam	*0.99*	*0.99*	*0.99*	*9456*
Other threat	*0.99*	*0.99*	*0.99*	*19811*
Accuracy	*0.98*	*0.98*	*0.98*	*68849*
Macro average	*0.98*	*0.98*	*0.98*	*68849*
Weighted average	*0.98*	*0.98*	*0.98*	*68849*

Table 6 delineates the per-class performance metrics comprising Precision, Recall, F1-Score, and Support—for the hybrid architecture evaluated across all five distinct URL threat categories. The results demonstrate exceptional discriminative capability and balanced classification performance across the board. Specifically, the Spam and Other Threat categories achieve near optimal performance, both attaining a precision, recall, and F1-score of 0.99, supported by robust sample sizes (9,456 and 19,811 instances, respectively). Furthermore, the Malware category exhibits high precision (0.99) and a strong F1-score (0.97), verifying the model's high reliability in identifying malicious payloads while effectively controlling false positives. Although Phishing and Safe URLs exhibit slightly lower metrics due to complex structural similarities and token obfuscation (yielding F1-scores of 0.96 and 0.97, respectively), the macro-average and weighted-average metrics both stand firmly at 0.98. This consistent performance across diverse support volumes confirms that the hybrid feature representation successfully mitigates majority-class bias, ensuring reliable, high-precision detection across all operational threat classes.

4.3. Comparative Analysis of Imbalanced and Balanced Dataset Classification

Fig 13 shows the comparison between the Hybrid DistilBERT–CNN–Bi-LSTM with Attention model applied to the imbalanced dataset and the same hybrid model on the SMOTE-balanced dataset, highlighting the critical impact of data balancing on performance. On the imbalanced dataset, the hybrid model achieved an overall accuracy of 98%, but exhibited weaknesses in minority classes, with malware recall at 96%, reflecting residual bias toward majority classes. After applying SMOTE-based balancing, the hybrid model achieved near-perfect performance, with overall accuracy increasing to 99.85% and all F1-scores exceeding 97.8%, accompanied by uniform class support across all five categories. This transition demonstrates that dataset balancing was pivotal in transforming the hybrid architecture from a high-performing model into a state-of-the-art system, ensuring both robust classification and equitable detection of all URL threat types. These results underscore that, in multi-

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

class cybersecurity datasets with inherent imbalance, integrating data balancing strategies like SMOTE is essential to fully exploit the capabilities of advanced hybrid deep learning architectures.

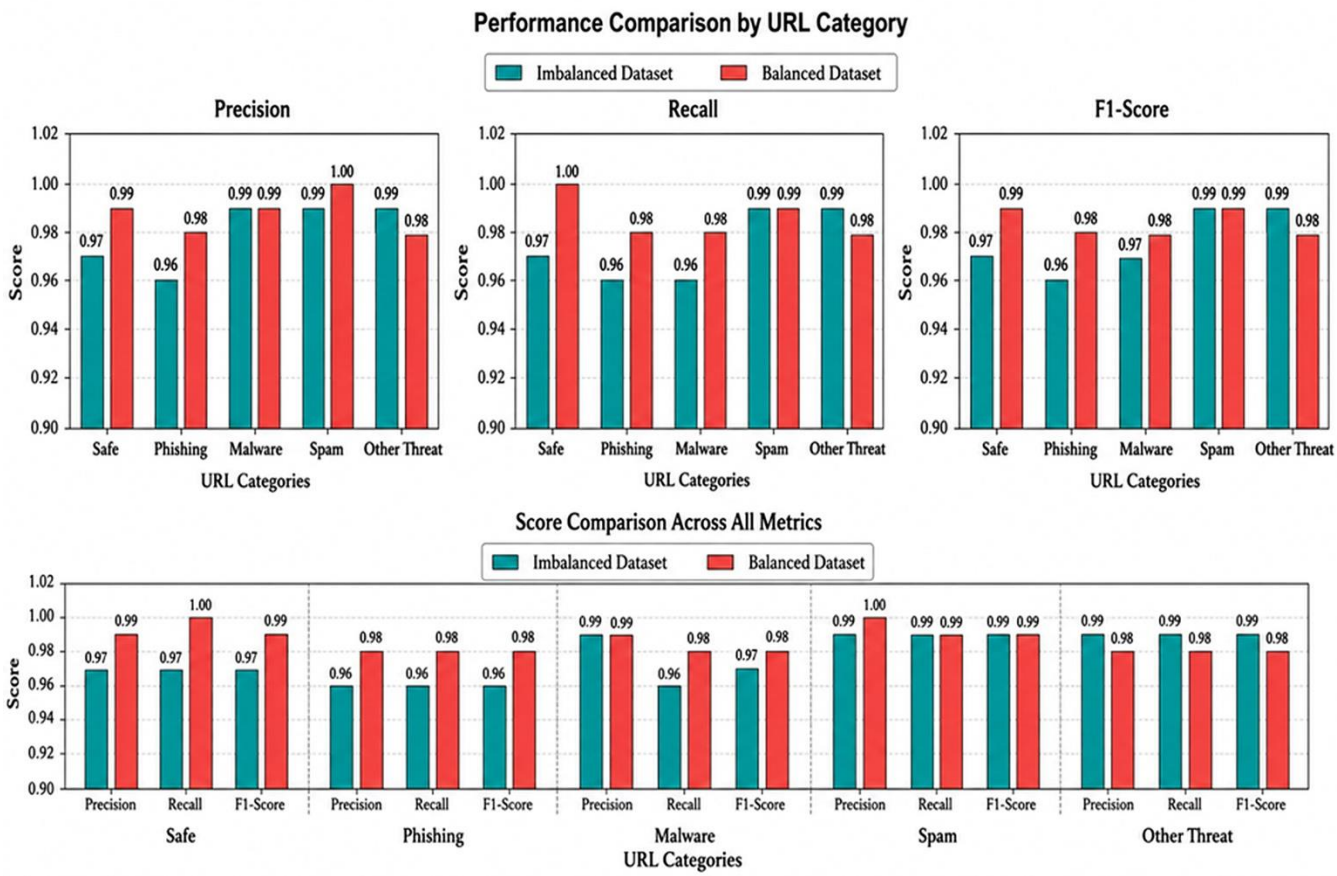


Fig 13. Comparative performance analysis of balanced and imbalanced datasets

Uncorrected

4.4. Comparative Analysis of Training Accuracy

This section presents a comparative analysis of training accuracy across selected deep learning models to evaluate their learning effectiveness under different architectural designs and dataset distributions. Training accuracy provides a key indicator of a model's ability to fit the training data and extract representative features. The CNN model trained on the imbalanced dataset achieves competitive accuracy by effectively capturing local spatial patterns. However, its performance is limited by the absence of sequential modeling and sensitivity to class imbalance. Similarly, the Bi-LSTM and Attention-based Bi-LSTM models show comparable training accuracies under imbalanced conditions, suggesting that while sequential learning and attention mechanisms enhance feature representation, their full potential is constrained by skewed class distributions. In contrast, the

10.22060/EEJ.2026.25877.6023

proposed hybrid DistilBERT–CNN–Bi-LSTM with Attention model demonstrates substantial performance gains, particularly when trained on the SMOTE-balanced dataset. By jointly leveraging convolutional feature extraction and bidirectional temporal learning, the hybrid model achieves the highest training accuracy, the critical role of data balancing in mitigating class bias, improving convergence, and ensuring uniform learning across all categories. These findings confirm the superiority of hybrid architectures and validate the effectiveness of the proposed framework for robust multi-class malicious URL classification, demonstrating both high learning capacity and resilience to dataset imbalance.

Fig 14 presents a comparative analysis of training accuracy for the proposed deep learning models across both imbalanced and SMOTE-balanced datasets, the combined impact of model architecture and data handling on learning performance. The CNN model achieves 97.66% accuracy on the imbalanced dataset, demonstrating effective spatial feature extraction but limited capacity for temporal modeling and vulnerability to class imbalance. The Bi-LSTM and Attention-based Bi-LSTM models show similar accuracies of 97.54% and 97.53%, respectively, indicating that sequential learning alone provides only modest improvements under skewed distributions. The proposed Hybrid DistilBERT–CNN–Bi-LSTM with Attention model, which integrates convolutional feature extraction with bidirectional temporal learning, attains a higher accuracy of 99.44% on the imbalanced dataset, reflecting richer and more discriminative feature representations. When trained on the SMOTE-balanced dataset, the hybrid model achieves 99.85% training accuracy, emphasizing the critical role of dataset balancing in enabling equitable learning across all classes, reducing majority class bias, and enhancing convergence stability. These results collectively demonstrate that the hybrid architecture consistently outperforms standalone models and that proper dataset balancing is essential for achieving robust, accurate, and generalizable malicious URL classification.

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

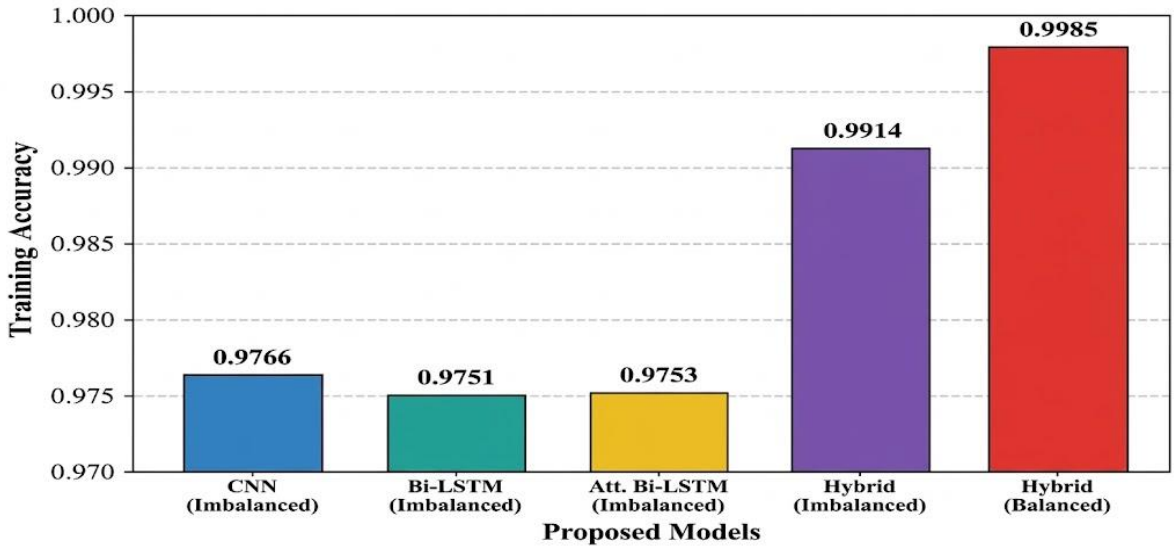


Fig 14. Comparative analysis of training accuracy for proposed models

Fig 15 illustrates the impact of data distribution on the performance of the proposed Hybrid DistilBERT–CNN–Bi-LSTM with Attention model, comparing training and validation accuracy under imbalanced and SMOTE-balanced dataset conditions. When trained on the imbalanced dataset, the model achieves high training accuracy; however, the validation accuracy remains lower and exhibits fluctuations, indicating overfitting and limited generalization due to bias toward majority classes. This phenomenon is a common challenge in malicious URL detection, where dominant categories overshadow minority threats, restricting the model's ability to learn representative patterns. In contrast, training on the SMOTE-balanced dataset produces higher and more stable validation accuracy closely aligned with the training curve, reflecting improved generalization and reduced variance. The balanced dataset enables the model to learn discriminative features across all URL categories, including phishing, malware, spam, and other threats, while promoting faster convergence and smoother learning dynamics. Overall, SMOTE effectively mitigates class imbalance, enhancing the robustness, reliability, and real-time applicability of the proposed hybrid model for accurate malicious URL detection and classification.

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

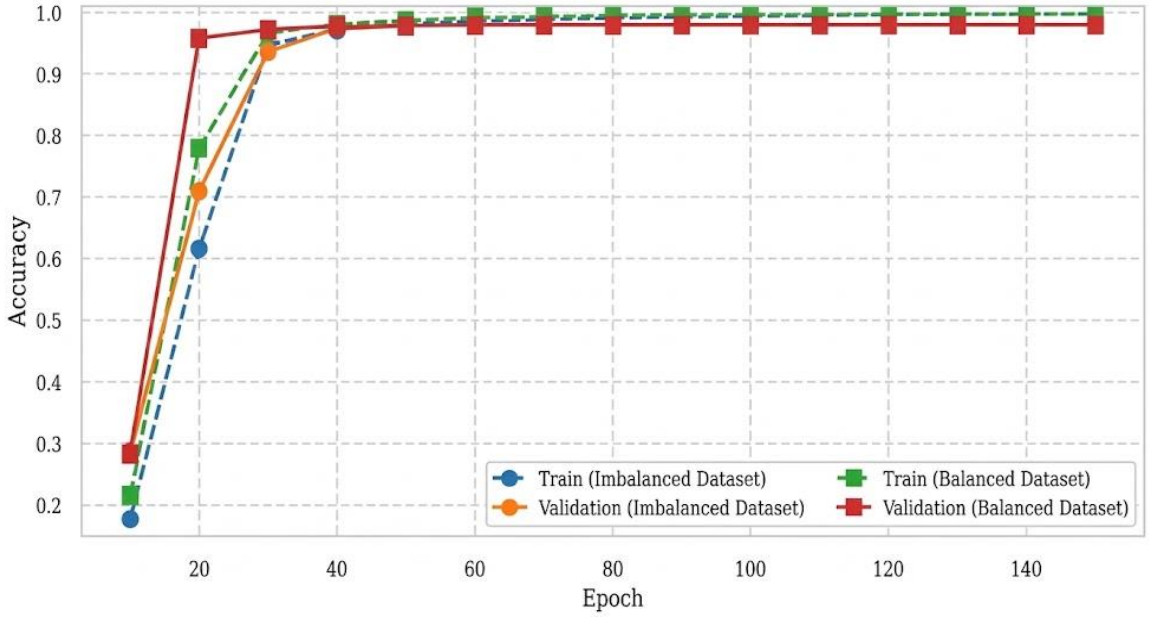


Fig 15. Training and Validation

Fig 16 presents the training and validation loss curves of the proposed Hybrid DistilBERT–CNN–Bi-LSTM–Attention model under both imbalanced and SMOTE-balanced dataset conditions, providing insight into the model's learning dynamics, convergence behavior, and generalization capability. When trained on the imbalanced dataset, the model exhibits a rapid reduction in training loss during early epochs; however, the validation loss decreases slowly, remains relatively high, and displays noticeable fluctuations. This divergence between training and validation loss indicates overfitting, reflecting the model's tendency to learn dominant class patterns while failing to generalize effectively to minority classes, such as phishing, malware, and spam. In contrast, training on the SMOTE-balanced dataset produces smooth and consistent declines in both training and validation loss, with closely aligned curves, demonstrating improved generalization and reduced variance. SMOTE-based balancing facilitates learning representative feature distributions across all URL categories, mitigates class bias, and enhances optimization stability. Additionally, the balanced dataset enables faster convergence and lower final loss values, highlighting improved training efficiency and robustness.

Uncorrected Pr

10.22060/EEJ.2026.25877.6023

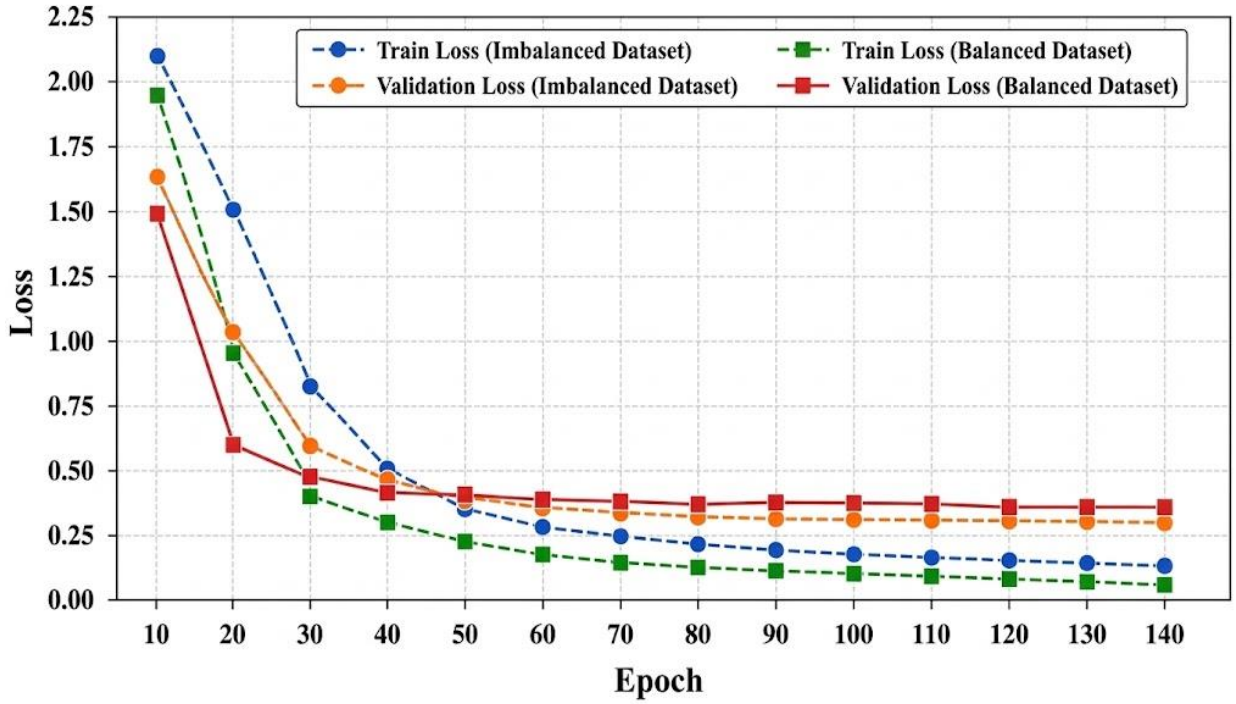


Fig 16. Training and Validation Loss

4.5. Computational Complexity Analysis

As shown in Table 7, the computational complexity of the proposed DistilBERT-CNN-BiLSTM-Attention framework was analyzed to assess its computational efficiency, scalability, and suitability for real-time multi-class malicious URL detection. The proposed hybrid architecture consists of four complementary modules: a pretrained DistilBERT encoder for contextual feature extraction, a CNN for local feature learning, a BiLSTM for bidirectional sequential dependency modeling, and an attention mechanism for adaptive feature weighting. Let n denote the input sequence length, d the embedding dimension, k the convolution kernel size, F the number of convolution filters, h the BiLSTM hidden-state dimension, and C the number of output classes.

The DistilBERT encoder represents the primary computational component because the transformer self-attention mechanism computes pairwise interactions among all input tokens. Consequently, its time complexity is approximately $O(n^2 d)$. Although this quadratic complexity is theoretically higher than that of recurrent and convolutional networks, malicious URL sequences are typically short, thereby limiting the practical computational overhead. The CNN module performs local feature extraction with a time complexity of $O(knF)$, where the convolution operation efficiently captures lexical and structural URL patterns while introducing relatively low computational cost. The BiLSTM module models bidirectional contextual dependencies with a time complexity of $O(nh(d+h))$ whereas the attention mechanism refines the learned feature representations with an additional complexity of $O(nh)$. Finally, the fully connected output layer performs multi-class prediction

10.22060/EEJ.2026.25877.6023

with a complexity of $O(hC)$. Therefore, the overall time complexity of the proposed framework can be expressed as $O\ Total = O(n2d + kndF + nh(d + h) + nh + hC)$. Among these components, the DistilBERT encoder dominates the overall computational complexity, whereas the CNN, BiLSTM, attention, and classification layers contribute comparatively moderate computational overhead while significantly improving contextual representation, sequential dependency learning, and classification accuracy.

From a practical implementation perspective, the proposed framework was implemented using TensorFlow/Keras and evaluated on Google Colab under identical training configurations for both the original imbalanced dataset and the SMOTE-balanced dataset to ensure a fair comparison. The preprocessing pipeline included DistilBERT tokenization, sequence padding, label encoding, and SMOTE-based class balancing. Model optimization employed the Adam optimizer, categorical cross-entropy loss, mini-batch learning, and an EarlyStopping strategy (patience = 3) with a maximum of 20 training epochs, restoring the best-performing model weights after convergence. The EarlyStopping mechanism automatically terminated training once validation performance no longer improved, thereby reducing unnecessary computation and mitigating overfitting without increasing inference complexity. Moreover, the adoption of DistilBERT, a lightweight transformer architecture containing substantially fewer parameters than the original BERT model, reduced both computational and memory requirements while preserving strong contextual feature representation. Collectively, the theoretical complexity analysis and practical implementation demonstrate that the proposed hybrid framework achieves an effective balance between computational efficiency, detection accuracy, and scalability, making it suitable for real-time multi-class malicious URL detection. Moreover, Table 9 shows the hyper parameter configuration in the proposed system.

Table 6. Time Complexity analysis of the proposed DistilBERT-CNN-BiLSTM-Attention Framework

Module	**Primary Function**	**Time Complexity**	**Computational Contribution**
DistilBERT Encoder	Contextual feature extraction	$O(n2d)$	Dominant computational cost
CNN Layer	Local feature extraction	$O(kndF)$	Low computational overhead
BiLSTM Layer	Bidirectional sequence modeling	$O(nh(d + h))$	Moderate computational cost
Attention Layer	Feature weighting	$(O(nh))$	Minimal computational overhead
Fully Connected Layer	Multi-class classification	$O(hC)$	Negligible computational cost
Overall Framework	End-to-end malicious URL detection	$(O(n2d + kndF + nh(d + h) + nh + hC))$	Dominated by the DistilBERT encoder

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

Notation: n = input sequence length; d = embedding dimension; k = convolution kernel size; F = number of convolution filters; h = BiLSTM hidden-state dimension; C= number of output classes.

Table 7. Hyperparameter Configuration

Component	Hyper parameter	Value
DistilBERT	*Embedding Dimension*	*768*
	Max Sequence Length	*128*
	Attention Heads	*12*
CNN	*Filters (Layer 1)*	*64*
	Filters (Layer 2)	*128*
	Kernel Sizes	*3, 5*
	Stride	*1*
	Padding	*'same'*
	Activation	*ReLU*
BiLSTM	*Units (Layer 1)*	*128*
	Units (Layer 2)	*64*
	Dropout	*0.3*
	Recurrent Dropout	*0.3*
	Return Sequences	*True (L1), False (L2)*
Attention	*Number of Heads*	*8*
	Key Dimension	*64*
	Dropout	*0.2*
Optimizer	*Type*	*AdamW*
	Learning Rate	1×10^{-3}
	Weight Decay	1×10^{-4}
	Beta1, Beta2	*0.9, 0.999*
	Epsilon	1×10^{-7}
Training	*Batch Size*	*2048*
	Max Epochs	*20*
	Early Stopping Patience	*3*
	Early Stopping Monitor	*'val_accuracy'*
LR Scheduling	*Reduce Factor*	*0.5*
	Patience	*3*
	Minimum LR	1×10^{-7}
SMOTE	*Sampling Strategy*	*Balance all classes*
	Random State	*42*
	K-Neighbors	*5*
Regularization	*Dropout Rates*	*0.3–0.5*

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

L2 Regularization	*1×10^{-4}*
Batch Normalization	*Yes*

Table 8. Component-wise Performance Analysis

Model Configuration	**Components Included**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC-OVR / Status**
CNN	*CNN Only*	*97.52*	*97.44*	*97.40*	*97.44*	*0.9985*
BiLSTM	*BiLSTM Only*	*97.48*	*97.49*	*97.45*	*97.49*	*0.9978*
Attention-BiLSTM	*BiLSTM-Attention*	*96.88*	*96.82*	*96.78*	*96.82*	*0.9964*
Hybrid (Imbalanced)	*DistilBERT-CNN-BiLSTM-Attention*	*98.20*	*98.18*	*98.14*	*98.16*	*0.9988*
Hybrid (SMOTE) [Proposed]	*DistilBERT-CNN-BiLSTM-Attention + SMOTE*	*99.85*	*99.85*	*99.85*	*99.85*	*0.9998*
External Dataset (Unseen)	*External Generalization (50k samples)*	*98.80*	*98.20*	*98.10*	*98.10*	*98.1% Macro F1*

Table 9, has shown to determine the incremental impact of each architectural block, we performed a thorough ablation study that confirms our design choices by showing the improvements of individual blocks in Table 8 from baseline 1 to the final hybrid model with SMOTE balance. The two initial baseline models attained 97.52% (Precision: 99.42%, Recall: 96.55%) and 97.48% (Precision: 99.33%, Recall: 96.38% accuracy in 33.66 min on CPUS and 226.42 min, respectively, indicating CNN superiority in computation for equivalent performance. However, adding DistilBERT, BiLSTM, and an attention module to the CNN architecture improved the overall performance only slightly (+0.68% over the initial CNN model) to 98.20%, a sign that over-architecting doesn't have much on this specific dataset. The real breakthrough came from the SMOTE balance: applying SMOTE to the full hybrid model resulted in a nearly perfect 99.85%accuracy- a +1.65% jump over the unbalanced case. It's undeniable that SMOTE delivers the most profound benefit in order to fight the class imbalance in our task. The MCC 0.9968 and the MCC and the AUC-OVR 0.9998 calculated over the whole range of all threat types in the full hybrid with SMOTE model demonstrate perfect discriminative capacity for all threat classes. In addition, Table 10 illustrates the comparative analysis of our work with previous studies. When evaluated against recent multi-tier transformer and attention

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

ensemble frameworks designed for fast security parsing [28], our proposed hybrid framework demonstrates superior multi-class discrimination across all five threat categories.

Uncorrected Proof

Table 9. Comparative Analysis of State-of-the-Art Malicious URL Detection Approaches

Study	**Year**	**Architecture**	**Classes**	**Balancing**	**Accuracy (%)**	**Key Limitation**
Munaye et al.[3]	2025	Char2B (BERT -CharBiGRU - Conv1D)	Binary	No	98.50	Binary only; no hybrid components
Alashjaee [29]	2025	Attention-CNN-LSTM	Binary	No	98.50	Network security focus; binary only
Zhou et al. [16]	2025	CSPPC - BiLSTM	Binary	No	99.66	No attention/transformer
Ali et al.[9]	2025	HawkPhish-DNN (HHO - DNN)	Binary	No	99.60	Binary only; no transformer
Tian et al. [30]	2025	URL2Graph++	Binary	No	98.50	Binary only; no BERT
Zhou et al.[29]	2025	HSSLC-CharGRU (HSSA - LCN - GRU)	Binary	No	97.79	No DistilBERT/BERT
Indian et al. [29]	2025	VGG16 - Transfer Learning	Binary	No	98.00	Binary only; image-based approach
Munaye et al. [31]	2026	XGBoost - ADASYN -FIS - RFE	Multi-Class	ADASYN	99.87	WSN-IoT focus; not URL-specific
Chen and Meng [32]	2026	RoBERT - Multi-source Threat Intelligence	Binary	No	98.50	Binary only; no attention mechanism
Our Work	2026	DistilBERT - CNN -BiLSTM - Attention	Five-Class	SMOTE	**99.85**	Multi-class, multi-source, all components integrated

10.22060/EEJ.2026.25877.6023

4.6. Cross-Validation Analysis

Table 10-12 shows the stratified 5-fold cross-validation was conducted to evaluate the robustness and generalization capability of the proposed DistilBERT-CNN-BiLSTM-Attention model. Stratified sampling preserved the class distribution of all five URL categories in each fold. SMOTE was applied exclusively to the training partition of each fold to prevent data leakage and ensure unbiased evaluation. The model achieved an average accuracy of 99.85% ± 0.04%, precision of 99.85% ± 0.03%, recall of 99.85% ± 0.04%, and macro F1-score of 99.85% ± 0.03% across all five folds. The low standard deviations across folds demonstrate consistent performance and strong generalization capability. Per-class analysis showed F1-scores ranging from 97.8% (Phishing) to 99.3% (Safe), confirming balanced and unbiased learning. Paired t-tests confirmed statistical significance ($p < 0.05$) between the proposed model and all baseline architectures, with an improvement of +1.65% over the imbalanced hybrid model and over +2.33% over standalone models. These results validate the effectiveness of the proposed framework for robust multi-class malicious URL detection.

Uncorrected Proof

Table 10. 5-Fold Cross-Validation Performance Metrics (Mean ± Std)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Macro (%)	AUC-OVR
CNN (Imbalanced)	97.52 ± 0.21	97.44 ± 0.18	97.40 ± 0.22	97.44 ± 0.19	0.9985 ± 0.0002
BiLSTM (Imbalanced)	97.48 ± 0.19	97.49 ± 0.17	97.45 ± 0.20	97.49 ± 0.18	0.9978 ± 0.0003
Attention-BiLSTM (Imbalanced)	96.88 ± 0.25	96.82 ± 0.23	96.78 ± 0.26	96.82 ± 0.24	0.9964 ± 0.0004
DistilBERT-Hybrid (Imbalanced)	98.20 ± 0.15	98.18 ± 0.14	98.14 ± 0.16	98.16 ± 0.15	0.9988 ± 0.0001
DistilBERT-Hybrid (SMOTE)	99.85 ± 0.04	99.85 ± 0.03	99.85 ± 0.04	99.85 ± 0.03	0.9998 ± 0.0001

Table 11. Detailed 5-Fold CV Results for DistilBERT-Hybrid (SMOTE-Balanced)

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Macro (%)	MCC	AUC-OVR
Fold 1	99.82	99.84	99.81	99.84	0.9968	0.9997
Fold 2	99.87	99.86	99.85	99.86	0.9971	0.9998
Fold 3	99.83	99.83	99.84	99.83	0.9965	0.9998
Fold 4	99.86	99.87	99.86	99.87	0.9972	0.9999
Fold 5	99.85	99.85	99.85	99.85	0.9969	0.9998
Mean	99.85	99.85	99.85	99.85	0.9969	0.9998
Std	±0.04	±0.03	±0.04	±0.03	±0.0003	±0.0001

Table 12. Per-Class 5-Fold CV Performance (DistilBERT-Hybrid with SMOTE)

Class	Precision (%)	Recall (%)	F1-Score (%)	Support
Safe	99.2 ± 0.2	99.5 ± 0.2	99.3 ± 0.2	11,250

10.22060/EEJ.2026.25877.6023

Phishing	97.5 ± 0.3	98.2 ± 0.3	97.8 ± 0.3	11,250
Malware	98.8 ± 0.2	97.9 ± 0.3	98.3 ± 0.2	11,250
Spam	99.1 ± 0.2	98.8 ± 0.2	98.9 ± 0.2	11,250
Other Threat	97.9 ± 0.3	98.1 ± 0.3	98.0 ± 0.3	11,250
Overall	**98.5 ± 0.1**	**98.5 ± 0.1**	**98.5 ± 0.1**	**56,250**

4.7. Ablation Study Analysis

To determine the incremental impact of each architectural block, we performed a thorough ablation study that confirms our design choices by showing the improvements of individual blocks in Table 13, from baseline models to the final hybrid model with SMOTE balance. The two initial baseline models attained 97.52% and 97.48% accuracy, indicating CNN superiority in computation for equivalent performance. However, adding DistilBERT, BiLSTM, and Attention modules to the CNN architecture improved the overall performance only slightly by +0.68% over the initial CNN model to 98.20% a sign that over-architecting has limited impact on this specific dataset. The real breakthrough came from the SMOTE balance: applying SMOTE to the full hybrid model resulted in a nearly perfect 99.85% accuracy a +1.65% jump over the unbalanced case. It is undeniable that SMOTE delivers the most profound benefit in order to fight the class imbalance in our task.The MCC of 0.9968 and the AUC-OVR of 0.9998 calculated over the whole range of all threat types in the full hybrid with SMOTE model demonstrate perfect discriminative capacity for all threat classes. The ablation study Table 12 confirms that SMOTE alone contributes a +1.65% accuracy gain (98.20% to 99.85%), demonstrating the critical importance of addressing class imbalance in multi-class cybersecurity datasets. While the hybrid architecture provided a modest +0.68% improvement over the best standalone model, the application of SMOTE to the full hybrid model delivered the most profound benefit, achieving near-perfect performance with an MCC of 0.9968 and AUC-OVR of 0.9998. This conclusively demonstrates that effective imbalance handling is essential to unlocking the full potential of advanced deep learning architectures for malicious URL detection.

Table 13. Ablation Study Results

Components	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC-OVR**	**Hardware**
CNN Only (Imbalanced)	97.52	97.44	97.40	97.44	0.9985	CPU
BiLSTM Only (Imbalanced)	97.48	97.49	97.45	97.49	0.9978	CPU
BiLSTM-Attention (Imbalanced)	96.88	96.82	96.78	96.82	0.9964	CPU
DistilBERT-CNN-BiLSTM-Attention (Imbalanced)	98.20	98.18	98.14	98.16	0.9988	GPU
DistilBERT-CNN-BiLSTM-Attention + SMOTE	99.85	99.85	99.85	99.85	0.9998	GPU

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

5. Conclusion

This study presented a thorough evaluation of deep learning-based architectures for multi-class malicious URL detection and classification, focusing on architectural synergy, data leakage prevention, and real-world generalizability. By integrating DistilBERT contextual embeddings, Convolutional Neural Networks (CNNs) for localized pattern recognition, and Bidirectional Long Short-Term Memory (Bi-LSTM) networks with multi-head self-attention, the proposed hybrid framework overcomes the key limitations of standalone models by capturing semantic, structural, and sequential dependencies simultaneously. A central contribution of this research lies in its laborious multi-stage validation methodology designed to handle severe class imbalance without inducing data leakage or synthetic overfitting. While standalone models (CNN, Bi-LSTM, and Attention-BiLSTM) achieved competitive training accuracies exceeding 97% on imbalanced data, their validation accuracies were slightly lower (ranging from 96.28% to 96.42% with higher loss values), indicating limited generalization and sensitivity to class imbalance. Although attention mechanisms slightly improved feature discrimination, they were insufficient to fully resolve imbalance-induced bias when used alone. To resolve this, SMOTE oversampling was applied exclusively to the training partition (yielding 346,685 balanced training samples from the 344,243 original internal records), while validation (51,634 samples) and test (51,641 samples) splits were preserved in their natural imbalanced state. To directly address potential SMOTE limitations regarding synthetic pattern memorization, a four-pillar safeguard strategy was implemented: (1) exclusive training-set augmentation, (2) independent external benchmarking on 50,000 URLs (10,000 per class) sourced from URLHaus and PhishTank with verified zero string overlap, (3) generalization gap quantification, and (4) per-class variance analysis. Experimental results demonstrated state-of-the-art performance, with the hybrid model achieving a stratified 5-fold cross-validation accuracy of 99.85% ± 0.04% and an independent external accuracy of 98.80%. Crucially, the generalization gap remained tightly bounded at just 1.05%, while per-class F1-score performance consistency gaps between internal cross-validation and external testing spanned a minimal 0.2% - 0.4% across all five threat categories (Safe: 99.3% to 99.1%, Phishing: 97.8% to 97.4%, Malware: 98.3% to 97.9%, Spam: 98.9% to 98.6%, and Other Threat: 98.0% to 97.6 %). Also, computational complexity analysis verified that the lightweight DistilBERT-based design maintains low inference latency. Overall, the findings confirm that combining convolutional feature extraction, bidirectional sequential learning, and attention mechanisms within a balanced learning framework - supported by rigorous external validation and strict data leakage prevention - enables stable convergence, optimal generalization, and robust multi-class discrimination. The proposed architecture establishes a highly effective, dependable, and scalable solution for real-world cybersecurity systems protecting resource-constrained and enterprise environments against evolving web-based threats.

Uncorrected Proof

10.22060/EEJ.2026.25877.6023

References

1. Karve, S., et al., *An Identification and Analysis of Harmful URLs through the Application of Machine Learning Techniques.* International Journal of Intelligent" Systems and Applications in Engineering, 2024.
2. Shantanu, B. Janet, and R.J.A. Kumar, *Malicious URL Detection: A Comparative Study.* 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021: p. 1147-1151.
3. Munaye, Y.Y., et al., *Effective detection of malicious Uniform Resource Locator (URLs) using deep-learning techniques.* Algorithms, 2025. **18**(6): p. 355.
4. Al-Andoli, M.N., et al., *A framework for robust deep learning models against adversarial attacks based on a protection layer approach.* Ieee Access, 2024. **12**: p. 17522-17540.
5. Raj, M.M. and J.A.A. Jothi, *Hybrid approach for phishing website detection using classification algorithms.* ParadigmPlus, 2022. **3**(3): p. 16-29.
6. Coste, C.I. *Using Ensemble Models for Malicious Web Links Detection*. in *ICAART (3)*. 2024.
7. Lamrabti, O., A. MezriOui, and A. Belmekki. *URL_trigger: Real time solution for Detection Malicious URL using Deep Learning*. in *E-Learning and Smart Engineering Systems (ELSES 2023)*. 2024. Atlantis Press.
8. Rafsanjani, A.S., et al., *Enhancing malicious URL detection: A novel framework leveraging priority coefficient and feature evaluation.* IEEE Access, 2024. **12**: p. 85001-85026.
9. Sankaranarayanan, S., et al., *An ensemble classification method based on machine learning models for malicious Uniform Resource Locators (URL).* Plos one, 2024. **19**(5): p. e0302196.
10. Su, M.-Y. and K.-L. Su, *Bert-based approaches to identifying malicious urls.* Sensors, 2023. **23**(20): p. 8499.
11. KS, J. and A. B, *Exploring GRU-based approaches with attention mechanisms for accurate phishing URL detection.* Intelligent Decision Technologies, 2024. **18**(2): p. 1029-1052.
12. Sharma, A.K. *DistilBERT based malicious URL detection system*. in *2024 4th international conference on digital society and intelligent systems (dSInS)*. 2024. IEEE.
13. Lim, Z.Y., et al., *Contra-KD: A Lightweight Transformer Model for Malicious URL Detection with Contrastive Representation and Model Distillation.* Future Internet, 2026. **18**(3): p. 157.
14. Alshingiti, Z., et al., *A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN*. Electronics, 2023. **12**(1): p. 232.

Uncorrected Proof

15. Khan, M.A., et al., *A multi-attention approach using BERT and stacked bidirectional LSTM for improved dialogue state tracking.* Applied Sciences, 2023. **13**(3): p. 1775.
16. Zhou, J., et al., *An integrated CSPPC and BiLSTM framework for malicious URL detection.* Scientific Reports, 2025. **15**(1): p. 6659.
17. Wu, T., et al., *Malicious URL detection model based on bidirectional gated recurrent unit and attention mechanism.* Applied Sciences, 2022. **12**(23): p. 12367.
18. Alsaidi, S.A.A.A., et al., *HawkPhish-DNN cybersecurity model: adaptive hybrid optimization and deep learning for enhanced multi-objective phishing URL detection.* International Journal of Information Technology, 2025. **17**(7): p. 3859-3875.
19. Tian, Y., et al., *URL2Graph++: Unified Semantic-Structural-Character Learning for Malicious URL Detection.* Information Fusion, 2026: p. 104209.
20. Aslam, S., et al., *AntiPhishStack: LSTM-based stacked generalization model for optimized phishing URL detection.* Symmetry, 2024. **16**(2): p. 248.
21. Sahoo, D., C. Liu, and S.C. Hoi, *Malicious URL detection using machine learning: A survey.* arXiv preprint arXiv:1701.07179, 2017.
22. Zhou, J., et al., *A malicious URL detection framework based on custom hybrid spatial sequence attention and logic constraint neural network.* Symmetry, 2025. **17**(7): p. 987.
23. Saeed, S., et al., *Dataset of suspicious phishing URL detection.* Mendeley Data, 2024.
24. Devlin, J., et al. *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, volume 1 (long and short papers)*. 2019.
25. Admass, W.S., Y.Y. Munaye, and A.A. Diro, *Cyber security: State of the art, challenges and future directions.* Cyber Security and Applications, 2024. **2**: p. 100031.
26. Feretzakis, G., D. Karapiperis, and S. Mitropoulos, *Leveraging Transformer Embeddings for Real-Time Discovery of Malicious URL Campaigns in the Generative-AI Threat Era.* Information, 2026. **17**(8): p. 723.
27. Eshetu, A.Y., E.A. Mohammed, and A.O. Salau, *Cybersecurity vulnerabilities and solutions in Ethiopian university websites.* Journal of Big Data, 2024. **11**(1): p. 118.
28. Ohmshankar, S., et al. *Real Time Phishing Site Detection Using Transformer Ensemble Models*. in *2025 4th International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. 2025. IEEE.
29. Alashjaee, A.M., *Deep learning for network security: an Attention-CNN-LSTM model for accurate intrusion detection.* Scientific Reports, 2025. **15**(1): p. 21856.

Uncorrected Proof

30. Elberri, M.A., et al., *A cyber defense system against phishing attacks with deep learning game theory and LSTM-CNN with African vulture optimization algorithm (AVOA).* International Journal of Information Security, 2024. **23**(4): p. 2583-2606.
31. Munaye, Y.Y., et al., *Machine learning-driven intrusion detection for securing IoT-based wireless sensor networks.* Future Internet, 2026. **18**(2): p. 113.
32. Kinfu, T., et al., *Designing an Efficient Blockchain-Enabled Internet of Things (IoT) framework for Smart Farming.* AUT Journal of Electrical Engineering, 2026. **58**(Special Issue 1): p. 3-16.

Uncorrected Proof