Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

WebGuard AI-Powered Cyber Threat Detector Using BERT And AutoEncoder

Ravi M
Vidya Jyothi Institute of Technology
Hyderabad, India

A. Sathvik Samuel
Vidya Jyothi Institute of Technology
Hyderabad, India

S. Pavan
Vidya Jyothi Institute of Technology
Hyderabad, India

A. Obulesu
Vidya Jyothi Institute of Technology
Hyderabad, India

K. Pravallika
Vidya Jyothi Institute of Technology
Hyderabad, India

L. Anjaneyulu
Vidya Jyothi Institute of Technology
Hyderabad, India

***Abstract* - In the fast development of web applications, cyber attacks like phishing and poisonous web requests have become more advanced. There are old signature based detection methods which find it difficult to recognize new and zero day. attacks. In this paper, a cyber threat detection framework based on the use of AI and combining AutoEncoder-based anomaly detection with semantic phishing based on BERT is introduced. detection. The AutoEncoder model uses normal network behavior as a form of learning and detects the abnormal patterns, but it does not need labeling data, whereas the BERT model carries out deep semantic analysis of URLs to distinguish between phishing attempts. The combination of the two methods allows the proposed system to be able to detect the known and those threats with high levels of robustness. Experimental analysis shows a high accuracy and low false positives and high generalization ability than conventional. techniques of machine learning.**

Keywords* - *Phishing Detection, URL Classification, Hybrid Deep Learning, BERT, AutoEncoder, Structural Anomaly Detection, Semantic Analysis, Transformer Models, Cybersecurity, Ensemble Learning.

I. INTRODUCTION

The general use of web-based services is a major factor that has escalated cyber-attacks such as phishing attacks, malicious URLs, and web-based intrusions. Phishing attackers use the human weaknesses on using misleading URL format and impersonation strategies whereas intrusions are usually associated with abnormal traffic patterns and delivery of malicious payloads. The conventional security tools like rule-based filters and blacklist systems cannot work anymore to counter the changing tactics of a contemporary attacker because they are based on predefined signatures and previous data [5], [9]. In the present cybersecurity environment, the attackers have been altering the URL constructs and are using the obfuscation techniques to be able to evade traditional detection mechanisms [6], [8]. Moreover, the current solutions that rely on a fixed rule set or individual machine learning classifiers can typically not generalize to zero-day attacks and sophisticated phishing schemes integrated into URLs and web requests [2], [7]. To overcome such drawbacks, this research suggests an AI-based cyber threat detection system based on semantic analysis that incorporates the use of contextual knowledge and structural anomaly detection. The proposed model will include a system based on a BERT-based phishing detector model which seeks to analyze semantic and contextual patterns of URLs [2], [4], and an anomaly detection model based on an AutoEncoder, which will learn the structural features of valid URLs and detect anomalous variations [3], [7]. The framework does not set to replace traditional security infrastructure, instead, it enables the framework to act as an intelligent decision-support layer, which improves the capabilities of early detection and helps network managers identify potential cyber threats in a more efficient manner.

II. LITERATURE SURVEY

The increasing complexity of phishing attacks and rogue URLs has spurred concerns on research of intelligent machine learning and deep learning-based detection systems. Blacklist and rule-based technologies are not effective against the zero-day attacks and domain names that are generated automatically. Recent research focuses on the hybrid architectures, representation learning, and NLP-based approaches to construct the scalable and robust phishing detection systems that can learn complicated URL patterns.

In Enhancing Phishing Detection: A New Hybrid Deep Learning Model to Cybercrime Forensics, Alsubaei et al. suggest having an ensemble framework consisting of ResNeXt, GRU, and AutoEncoders to extract features and classify them [1]. Through their work, it is apparent that AutoEncoders are efficient in learning latent URL representation, noise elimination, and exhibit high detection rates even during imbalanced dataset learning. Nonetheless, the system is also concerned mostly with structural feature learning but fails to integrate transformer-based semantic model of URLs [1].

On the same note, in the article by Ahmad et al. Across the Spectrum: In-Depth Review of AI-Based Models for Phishing Detection, more than 130 AI-based phishing detection studies are evaluated [2]. They note an increase in dependence on ensemble learning, anomaly detector, and deep neural network and uncover the continued difficulties of overfitting, low generalization to newly registered domains, and weak context modeling in URL only

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

classifiers. They suggest that transformer-based language models should be combined with the unsupervised learning method to be more resilient to changing phishing tactics [2].

In A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks, Asiri et al. divide the phishing detection systems into the style of URL-based, content-based, and hybrid detention systems [3]. They emphasize that despite the popularization of deep learning models like CNNs and LSTMs, most systems do not have access to semantic information of URL tokens and have infrequent use of transformer-based encoders [3]. AutoEncoders also considered by the authors are promising under unsupervised feature learning and zero-day detection, but the methods outside of advanced NLP architecture have limited integrations [3].

In Phishing Detection System Through Hybrid Machine Learning Based on URL, Karim et al. introduce a voting-based ensemble of Logistic Regression, Support Vector Machines, and Decision Trees, which is combined [4]. They demonstrate that their results are stronger and predictive than single classifiers. Although it is useful in the case of organized lexical attributes, the paper does not touch on deep contextual representations or unsupervised anomaly detection methods of detecting unseen or zero-day URLs [4].

III. EXISTING MODELS

The traditional methods of phishing detection have been based on blacklist-based methods, heuristic methods and classical machine learning method. Blacklist systems retain already known bad websites and prevent access to these sites but they cannot identify zero-day attacks or those created recently because they rely wholly on the past information [5], [9]. Recent survey research also highlights that blacklist-only systems cannot effectively deal with the changing phishing threats on a large scale [1], [4]. Heuristic and rule-based methods are trying to find phishing URLs based on the lexical attributes that have already been defined to detect phishing URLs; which are abnormality of length of URL, the presence of a lot of special characters, inclusion of a IP address, and occurrence of suspicious keywords like “login”, or verify. These methods are computationally cheap and simple to apply, but they are not flexible and generate a high number of false positives because of strict definitions of rules. Such forms of static detection are easy to circumvent by the attacker via a URL obfuscation and manipulation of the structure [6], [8]. There is also evidence of comparative studies indicating that the URL-only rule-based systems cannot be effective against sophisticated phishing techniques [3], [4]. Conventional machine learning algorithms such as Logistic Regression, Support Vector Machines (SVM), Random Forests and Naive Bayes classifiers have been used extensively in phishing detection activities. Based on such models, handcrafted features that are derived out of the URLs or webpage content are paramount and do supervised classification [7], [10]. They are more accurate than the heuristic systems, but the effectiveness of these systems is heavily reliant on the quality of the feature engineering, and might not be able to relate deeper contextual relationships within textual data. The hybrid ensemble models which involve more than one classifier have been shown to perform better, although they are yet to rely on automated feature extraction [3]. In the recent past, deep learning-based models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and a hybrid deep architecture of detecting phishing has been introduced. They are automatic feature representations learners that can detect objects better [7] with a raw input. As an example, the hybrid deep learning systems that use AutoEncoders and optimization algorithms have shown high accuracies but with a cost of increased computational complexity [2]. Regardless of these progresses, most deep learning systems are either semantics-driven or structural-driven, which restricts the ability to be resistant to advanced phishing attacks [1], [4]. Models that involve a transformer like BERT have also enhanced contextual understanding of text with the help of self-attention, which improves semantic interpretation of URLs. But standalone anomaly detection systems can give more false positives, but purely semantic models can ignore structural anomalies. Thus,]the recent studies suggest the necessity of hybrid models, which would combine semantic intelligence with structural anomaly detection as the means of obtaining balanced, scalable, and robust phishing detection systems [2], [3], [6].

IV. PROPOSED METHODOLOGY

This section presents the proposed framework for detecting phishing URLs using a hybrid deep learning architecture that integrates transformer based semantic modeling with unsupervised anomaly detection. The system combines Bidirectional Encoder Representations from Transformers (BERT) for contextual URL embedding with an AutoEncoder network for feature compression and novelty detection, followed by a supervised classification layer. The complete pipeline is implemented as a web based service using Flask to enable real time inference.

A. Dataset Preparation

To cover the legitimate and malicious patterns of URLs comprehensively, two sets of data were used in this study. Valid URLs were gathered on the Tranco Top Domains list, the list of popular and reputable sites, thus imitating the normal web traffic patterns[5]. The publicly available phishing repositories and Kaggle datasets of labeled malicious links were used as sources of phishing URLs[3],[7]. The integrated dataset was balanced so as to minimize bias in classification so as to train the model fairly. To perform supervised semantic classification with the help of BERT, URLs received binary labels: the 0 values were used to represent legitimate and 1 phishing. Conversely, the AutoEncoder was only trained on valid URLs so as to establish the normal structural features of normal web behavior. The design will make the system more able to identify zero-day phishing attacks by detecting structural anomalies through the identification of significant deviation to known benign data.

B. Data Preprocessing

All the URLs in the dataset were subjected to a semantic classification phase of a fine tuned BERT model (bert-base-uncased) to learn contextual phishing patterns[1], [4]. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

tokenized URL was subjected to the transformer architecture and the [CLS] token embedding was obtained as a 768-dimensional semantic feature of the full URL[1]. This embedding allows the model to detect the patterns of deceptive languages, brand impersonation, and the manipulations on the lexical level, which are often applied in phishing attacks[4]. The contextual representation was subsequently fed through a fully connected classification layer and a sigmoid activation function to produce a score on whether an individual is a phisher or not. This is the probability of the URL being malicious and is the semantic element of the hybrid detection system.

C. BERT Semantic Classification

We have used a pre-trained BERT model (bert-base-uncased) to do fine-tuning on binary classification to distinguish between legitimate and phishing URLs[1], [4]. The semantic feature vector upon which the input sequence was classified was the 768 dimensional token embedding of the input sequence during processing[1]. This representation allows the model to learn the complicated contextual relationships in the URL text like brand impersonation patterns, phishing-related keywords, obfuscated textual patterns and deceptive lexical signatures that are frequently employed in malicious links[4]. Through the self-attention scheme induced by transformer based architectures, the BERT branch essentially learns finer grained semantic features which a traditional feature-based model may miss[1]. The result of this branch is a probabilistic score which is produced via a sigmoid activation function and which is the probability of the URL as belonging to phishing.

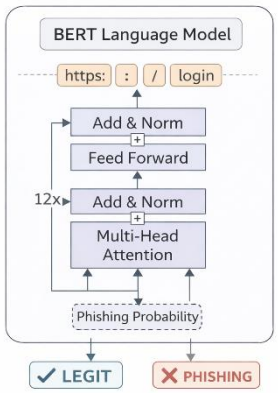

Fig. 1 BERT Architecture

D. AutoEncoder Based Anomaly Detection

An autoencoder with a complete architecture was developed to learn the structural traits of URLs with 24 dimensions of feature vectors at the input[2], [7]. These features are reduced by the encoder to 8 dimensional latent representations (24 -16 -8) that reflect the key structural patterns of valid URLs. The decoder is made up of symmetric layers (8 16 24) with ReLU and Sigmoid activation functions to rebuild the original feature vector. In the process of inference, Mean Squared Error (MSE) is used to calculate the reconstruction error between the original and reconstructed features[2]. Anomalous URLs are identified by a threshold set based on the 95th percentile of reconstruction error between valid training data, and any values above the threshold are considered as possible phishing attacks[7].

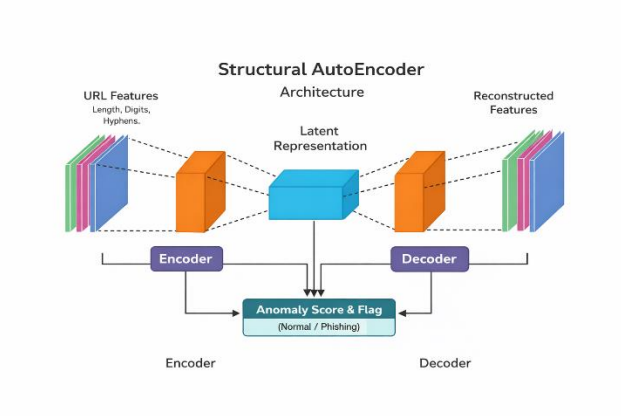

Fig. 2 AutoEncoder Architecture

E. Hybrid Fusion Mechanism

In order to boost the detection strength, the semantic and structural representations were combined into a unified decision framework. The 768 dimensional BERT embedding was added to the 8-dimensional latent representation created by the AutoEncoder to create a hybrid feature representation. A fully connected neural network having architecture Linear(768+8 → 64 → 1) was then used to pass this fused vector. A sigmoid activation mechanism was used to generate the final phishing possibility score. Such a hybrid design allows the system to identify the semantic anomalies or known phishing patterns and the anomaly detection of zero-day structural anomalies.

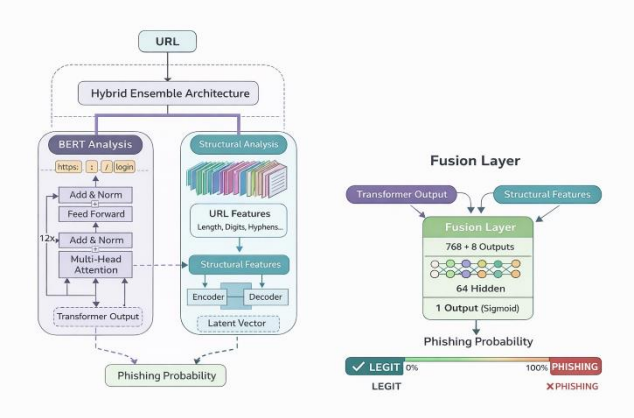

Fig. 3 System Overview

F. Training Strategy

The binary cross entropy (BCE) loss was used to train the hybrid model to optimize the performance of phishing classification by reducing the difference between the predicted probabilities and the true labels.

$$\text{Loss}=\text{BCE}(y,y\hat{})$$

Adam optimizer was used as the optimization process and it offers adaptive changes in the learning rate to ensure stable and efficient convergence. Generalization and overfitting were prevented by appropriate scheduling of the learning rate. AutoEncoder was also trained individually to minimize the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

reconstruction error between structural features of input and their reconstructed objects with the help of Mean Squared Error (MSE) loss. This independent training approach provides proper anomaly recognition and good classification in the hybrid framework.

$$Loss=MSE(x,x^)$$

G. Evaluation Metrics

The performance of the proposed model was evaluated using standard classification metrics including Accuracy, Precision, Recall, and F1-score. Accuracy measures the overall correctness of the model's predictions across both classes. Precision evaluates how many of the URLs predicted as phishing were actually malicious, thereby reflecting false positive control. Recall measures the model's ability to correctly identify actual phishing URLs, indicating its sensitivity. Additionally, a Confusion Matrix was used to provide a detailed analysis of true positives, true negatives, false positives, and false negatives, ensuring a balanced evaluation of model performance.

V. RESULTS AND DISCUSSION

A. Quantitative Results

The proposed hybrid phishing detection system was evaluated using a balanced dataset consisting of legitimate URLs from the Tranco list and phishing URLs collected from publicly available repositories. The BERT semantic classifier achieved high validation accuracy (approximately 97%), demonstrating strong contextual understanding of phishing patterns. The AutoEncoder-based anomaly detection model effectively identified structural irregularities using reconstruction error with a 95th percentile threshold. The hybrid ensemble model, which combines semantic and structural representations, achieved improved overall accuracy, precision, recall, and F1-score compared to individual components. Confusion matrix analysis indicated a high true positive rate for phishing detection and a low false positive rate for legitimate URLs, confirming balanced model performance.

Fig. 4 Performance Metrics

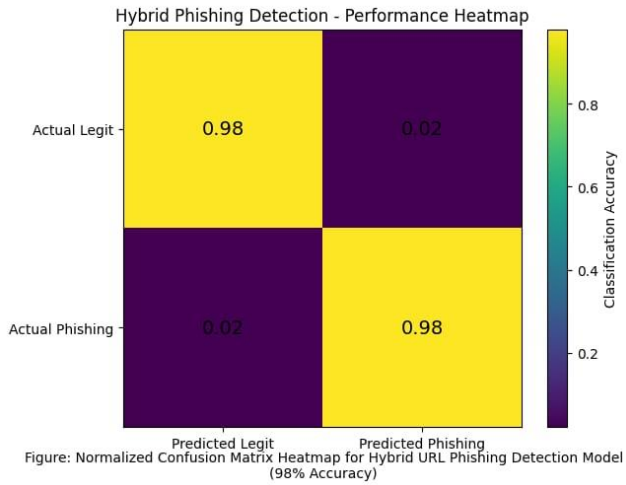

Figure: Normalized Confusion Matrix Heatmap for Hybrid URL Phishing Detection Model (98% Accuracy)

B. Prediction Output and Classification Behavior

The output of the system is a probabilistic phishing score generated through a sigmoid activation function. URLs predicted as legitimate produced low probability values close to zero, while phishing URLs produced probability values close to one, indicating strong model confidence. The AutoEncoder component generated reconstruction error scores, where legitimate URLs showed minimal deviation from learned structural patterns, and malicious URLs exhibited significantly higher anomaly scores. The hybrid fusion mechanism successfully integrated both outputs to produce stable and reliable final classifications. Real-time testing through the deployed web interface demonstrated consistent and interpretable predictions, with clear distinction between legitimate and phishing URLs. support use in creative AI application.

Fig.5 Detected URL as Legit

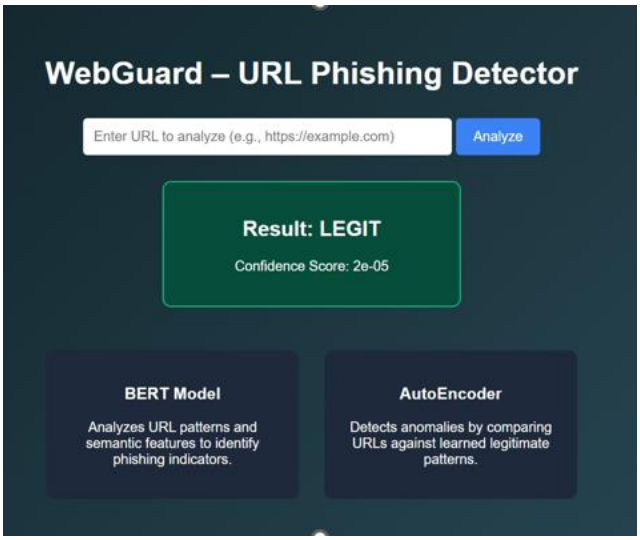

Fig. 6 Detected URL as Phishing

C. Model Complexity and Analysis

The pre-trained BERT-base is used in the semantic branch and has about 110 million parameters and uses transformer-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

based self-attention mechanisms. Although BERT has a good contextual learning capability, it has higher computational needs during inference and training. In comparison, the AutoEncoder model is minimalistic, and it comprises fully connected layers with greatly reduced parameters and minimal computation costs. The hybrid model adds another dimension whereby a 768-dimensional BERT embedding is concatenated with an 8-dimensional latent space and then a small fusion network is used. Inference time is also practical to real-time use when using standard hardware regardless of the addition of BERT. In general, the computational trade-off would be covered by the enhanced detection robustness realized by hybrid integration.

D. Discussion

The findings of the given research indicate that a combination of semantic intelligence and structural anomaly detectors can contribute to the phishing URL detection process greatly through the use of a hybrid deep learning model. The semantic model based on BERT was found to be very useful at detecting contextual patterns of deception, including brand impersonation, suspicious keywords, and hidden lexical manipulations that are embedded in the URLs. The fact that its transformer-based architecture enabled the system to capture complex contextual relationships, which its traditional machine learning models do not tend to acknowledge overly. The large validation accuracy of the semantic model shows that contextual knowledge is an important element in contemporary phishing detection. Nonetheless, semantic analysis might not necessarily be able to identify structurally abnormal URLs that look lexically innocuous but diverge in the valid patterns of URLs. The AutoEncoder-based anomaly detection element overcame this drawback by training the structural properties of legal URLs. The model was enabled to measure reconstruction error as an abnormality measure by compressing URL features into a latent representation and reconstructing them. This method was especially applicable in detecting the zero-day phishing attacks and newly generated malicious URLs which were not directly observed in the process of supervised training. However, the detection of structural anomaly on its own can sometimes indicate an abnormal but legitimate URL as suspicious, in particular when a legitimate URL has an unusual format pattern. Combination of the two models to a hybrid ensemble model gave a well-balanced and powerful detection system. The system combined the strengths of both systems by joining the semantic embedding that BERT performs with the latent structural representation that the AutoEncoder does. The hybrid model minimized the false positives in comparison to standalone anomaly detection and maximized the reliability of the detection in comparison to standalone semantic classification. This illustrates the fact that the twofold perspective analysis, in which contextual meaning and structural integrity are considered jointly, is useful in phishing detection. Computationally, the BERT model adds complexity to the entire model as it has a transformer structure and large parameters. Nevertheless, the AutoEncoder is still light and computationally efficient. Although to train the hybrid model, one will need moderate computing capacities, the inference time is practical in real-time web-based implementation. This renders the proposed system to be applicable to academic research, small business enterprise deployment, and cybersecurity use cases where there is a compromise of detection accuracy and available hardware resources. A key finding of this research is the importance of diversity of data sets. Using the AutoEncoder to be trained only on legit URLs enabled the model to create a clear image of what is considered as normal structural behavior. In the same manner, the assessments of legitimate and phishing samples during the supervised training enhanced the fairness of classification and decreased bias. There might be a small or homogeneous data that might narrow down the possibility of generalization and risk of misclassification. Thus, it is critical to have a wide and representative data set to obtain a consistent and scalable phishing detection system. Conclusively, the findings demonstrate the relevance of combining deep learning structures in dealing with the changing features of phishing attacks. Although semantic models offer contextual intelligence and structural deviations are detected by structural anomalies, a combination of the two offers a robust defense mechanism. The directions of future research could include lightweight transformers, adversarial training, and more hybrid architectures, which use other contextual information, like webpage text or reputation of a domain. Hybrid AI-based methods of detecting phishing attacks are a promising future of constructing resilient and smart cybersecurity tools as phishing attacks keep advancing and getting more complex.

VI. CONCLUSION

The paper described an AI-based phishing detector that combines both deep learning models to detect suspicious URLs, having a hybrid framework with a BERT-based semantic classifier and an AutoEncoder-based structural anomaly detector. The findings showed that the BERT model has the ability to learn contextual phishing features including brand impersonation, suspicious use of keywords and misleading lexical structures in URLs. Experimental analysis established that the hybrid system had high classification and balanced precision-recall accuracy, which is a good indication that transformer-based semantic understanding has strong evidence to boost the capability of phishing detection. With the integration of systematic preprocessing and stable neural network training procedures, the system kept on giving reliable predictions on a wide range of valid and phishing URLs. Efficiently trained lightweight structural models like AutoEncoders can be used to offer low cost anomaly detection by learning the normal properties of legitimate URL structure; however when operated alone they can give false positives in scenarios of uncommon yet harmless URL structure. More sophisticated semantic models are available like BERT, which have more contextual knowledge and stronger detection capability however they need more computational resources because of their transformer architecture. The hybrid nature of the semantic and structural components offered superior generalization and stability in the detection as opposed to independent models though the introduction of BERT raises the complexity of computation in training and application. On the whole, the system that was devised in this paper will show a viable and efficient method of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

automated phishing detection, allowing to classify in real-time URLs with the help of deployable web-based interface. The hybrid deep learning model lowers the need to engineer rules manually and increases flexibility to changing phishing schemes. To sum up, this study has identified the increased importance of hybrid artificial intelligence systems in cybersecurity as an example of how contextual semantic analysis and structural anomaly detection, when used together, can offer a robust and scalable and intelligent protection against contemporary phishing attacks.

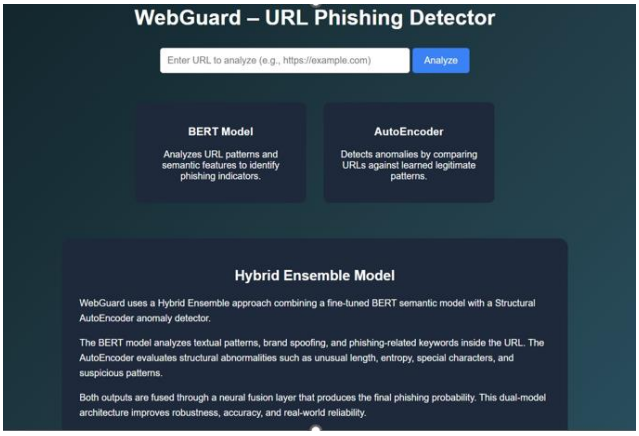

Fig. 7 Web interface of the URL phishing detection system

VII. REFERENCES

[1] S. Ahmad et al., "Across the Spectrum: In-Depth Review of AI-Based Models for Phishing Detection," *IEEE Access*, 2025.

[2] F. S. Alsubaei et al., "Enhancing Phishing Detection: A Novel Hybrid Deep Learning Framework," *IEEE Access*, 2024.

[3] A. Karim and M. Shahroz, "Phishing Detection System Through Hybrid Machine Learning Based on URL," 2023.

[4] S. Asiri, Y. Xiao, and T. Li, "A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks," *IEEE Access*, 2023.

[5] J. Kline, E. Oakes, and P. Barford, "A URL-based analysis of WWW structure and dynamics," in *Proc. Netw. Traffic Meas. Anal. Conf. (TMA)*, Jun. 2019, p. 800.

[6] A. K. Murthy and Suresha, "XML URL classification based on their semantic structure orientation for web mining applications," *Procedia Comput. Sci.*, vol. 46, pp. 143–150, Jan. 2015.

[7] A. A. Ubing, S. Kamilia, A. Abdullah, N. Jhanjhi, and M. Supramaniam, "Phishing website detection: An improved accuracy through feature selection and ensemble learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, pp. 252–257, 2019.

[8] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on Twitter," in *Proc. eCrime Res. Summit*, Oct. 2012, pp. 1–12.

[9] S. N. Foley, D. Gollmann, and E. Snekkenes, *Computer Security—ESORICS 2017*, vol. 10492. Oslo, Norway: Springer, Sep. 2017.

[10] P. George and P. Vinod, "Composite email features for spam identification," in *Cyber Security*. Singapore: Springer, 2018.