

PHISHING WEBSITE DETECTION USING BERT AND LSTM FOR URL AND CONTENT ANALYSIS

Abraham A. Kiazolu

Hill city university, Monrovia, Liberia.

kiazoluabrahama@gmail.com

<http://doi.org/10.35409/IJBMER.2024.3585>

ABSTRACT

The rapid growth of mobile healthcare technologies has generated greater dependence on multimedia health records including images, videos and audio files, for clinical decision-making. Nonetheless, storing and transmitting large amounts of mobile health data has been a thorny challenge in the aspects of storage, latency, bandwidth usage and data protection particularly in timely or emergency situations. This study recommends an integrated framework that incorporates hybrid cloud computing with Content Delivery Networks to promote the performance and reliability of mobile multimedia health record systems. The hybrid cloud structure utilizes both private and public cloud infrastructures for flexibility, scalability and controlled data management. At the same time CDNs enable speedy and secure distribution of multimedia information by replicating data on numerous geographically dispersed servers. This two-layer integration maximizes content availability minimizes access latency and provides data redundancy and therefore facilitates real-time access and enhanced responsiveness in healthcare settings. Performance testing was done under actual healthcare conditions resulting in noteworthy gains in speed of access to data minimizing bandwidth bottlenecks and increased security for the system. The solution ensures unbreakable, scalable and secure access to health records and is particularly useful for emergency care as well as rural healthcare provision. By resolving key issues in mobile health data management, this work makes contributions to the establishment of intelligent, cloud-based healthcare infrastructures that support the increasing need for connected, data-intensive and patient-focused care solutions.

Keywords: Mobile health records, Hybrid cloud, Content Delivery Network (CDN), Multimedia healthcare data, Cloud computing in healthcare, Secure health data management, Emergency healthcare systems.

1. INTRODUCTION

The growing dependence on internet services and e-commerce platforms has exponentially increased users' vulnerability to cyber threats. One of the most prevalent and dangerous forms of cybercrime is phishing, where attackers deceive users by masquerading as legitimate entities to extract sensitive information such as usernames, passwords, and credit card details [1]. These attacks are not only persistent but also becoming increasingly sophisticated, exploiting both technical weaknesses and human psychology. Traditional detection systems, largely based on blacklists and rule-based approaches, are proving to be ineffective against dynamic phishing techniques. As a result, there is a pressing need for intelligent, adaptive solutions that can effectively identify phishing threats in real-time before users fall victim. Phishing websites

typically imitate trusted organizations and are designed to lure unsuspecting users into entering personal data [2]. These sites are often short-lived and can be generated and taken down within hours, making it difficult for static detection systems to keep up. Moreover, attackers employ various obfuscation techniques to evade detection, such as URL manipulation, redirection chains, and the use of embedded iframes or JavaScript [3]. Given the evolving nature of phishing tactics, static lists and heuristic rules are not sufficient. A machine learning-based solution offers a more promising approach by analyzing patterns and learning from large datasets to detect phishing attempts based on anomalies and behaviors. Devarajan (2019) [4] designed a privacy-preserving model using hybrid encryption to secure cloud-based data transmission. Integrating this mechanism enhances the proposed mobile health infrastructure's stability and instantaneous responsiveness, especially in emergency or remote scenarios requiring encrypted multimedia data sharing.

Natural Language Processing (NLP) has emerged as a powerful tool in cybersecurity, particularly in analyzing the content and metadata of webpages. With the advent of deep learning, models such as BERT (Bidirectional Encoder Representations from Transformers) have demonstrated exceptional capabilities in understanding contextual and semantic nuances in textual data. BERT can capture bidirectional relationships in sentences, making it ideal for analyzing website content, including subtle linguistic manipulations used in phishing pages. When combined with structured URL features, it can form a comprehensive detection model that balances both linguistic and structural indicators of phishing [5]. In parallel, Recurrent Neural Networks (RNNs), particularly LSTM (Long Short-Term Memory) networks, are known for their ability to model sequential data and remember long-term dependencies. LSTM can effectively analyze URL sequences, domain structures, and webpage behaviors over time. By applying LSTM to URL analysis, the model can learn common phishing patterns embedded in domain names, subdomains, and path structures. Integrating LSTM into a detection framework enables a dynamic understanding of the structural evolution of phishing URLs, enhancing detection robustness against obfuscated links.

This paper presents a hybrid deep learning framework that combines the strengths of BERT and LSTM for the detection of phishing websites. The model analyzes both the textual content of webpages using BERT and the structural characteristics of URLs using LSTM. This dual-channel approach ensures that the system is not overly reliant on a single feature type, increasing its resilience against deceptive tactics. The system aims to provide a scalable, real-time phishing detection mechanism that can operate across a wide range of platforms, including browsers, email gateways, and security appliances. A key innovation in this framework lies in the preprocessing pipeline, where URLs and webpage content are sanitized, tokenized, and embedded using appropriate NLP techniques. For content analysis, BERT embeddings are generated from the visible text on the phishing and legitimate websites. For URL sequences, LSTM models are trained to detect unusual patterns such as excessive URL length, unusual TLDs, or uncommon word sequences. Together, these models work in harmony to deliver a high-confidence prediction on whether a given website is phishing or legitimate [6].

The dataset used in this study comprises a balanced mix of phishing and benign websites collected from open-source repositories and real-time web crawlers. The phishing websites are

validated using public feeds and threat intelligence services, ensuring a reliable ground truth. The dataset is further augmented with synthetic variations to simulate zero-day phishing tactics and improve generalizability. This diverse and dynamic dataset ensures that the model does not overfit to specific patterns but instead learns generalized representations of phishing behaviors. Performance metrics such as accuracy, precision, recall, F1-score, and AUC are used to evaluate the model's effectiveness. Compared to traditional machine learning algorithms and existing phishing detection models, the proposed hybrid architecture demonstrates superior performance across all evaluation metrics. It shows particular strength in identifying previously unseen phishing sites, which is a significant step forward in proactive cybersecurity. The use of deep contextualized representations and sequence modeling offers an advanced perspective over rule-based or surface-level statistical approaches. A privacy-preserving mechanism ensures the confidentiality of Alzheimer's diagnostic data derived from MRI scans. Gattupalli (2022) [7] emphasizes the reliability of this strategy in maintaining data integrity, reinforcing the proposed system's privacy-oriented multimedia health processing for timely clinical decisions.

An important aspect of the framework is its adaptability to changing attack vectors. The model can be retrained incrementally as new phishing techniques emerge, and its modular structure allows for the integration of additional features such as HTML structure, embedded script analysis, or image similarity detection. This adaptability is crucial for maintaining detection accuracy in the face of evolving cyber threats. Moreover, the system can be deployed in cloud environments, enabling real-time scanning and centralized updates without requiring significant local resources. The deployment strategy for the model considers performance, latency, and integration feasibility [8]. It supports deployment as a browser extension, an API for security services, or a plug-in for email clients. Cloud-based deployment ensures that detection capabilities are constantly updated and uniformly distributed across devices and platforms. Additionally, privacy-preserving mechanisms are incorporated to ensure that sensitive user data is not logged or misused during phishing detection. The system design adheres to data protection standards and regulatory compliance requirements.

From an architectural standpoint, the model incorporates a decision fusion layer that aggregates outputs from the BERT and LSTM models. This layer uses weighted averaging and confidence thresholds to produce a final prediction. By doing so, it avoids over-reliance on either the content or the URL alone, thereby reducing false positives and improving robustness. The fusion mechanism is also tunable based on deployment context—for example, emphasizing URL features in email gateways and content features in web browsers. To validate real-world applicability, the system is tested under various usage scenarios, including redirection-based phishing, image-only phishing, and fast-flux hosting [9]. The results show that the hybrid model maintains high detection rates even under adversarial conditions. In comparison to antivirus engines and traditional heuristics, it responds faster and flags threats with higher confidence. These evaluations underscore the model's capability to provide strong protection in operational cybersecurity infrastructures.

User experience is also a focal point in the design of the phishing detection system. Visual indicators and alert levels are used to guide users without causing unnecessary alarm or alert

fatigue. The system provides explanations for flagged URLs or content to promote user trust and awareness. This transparency not only increases user confidence but also serves as an educational tool to improve cyber hygiene among end-users. While the model performs well under current test conditions, the paper also discusses potential limitations and areas for future improvement. These include challenges related to multi-lingual content, the need for faster inference times in low-bandwidth environments, and the integration of more advanced threat intelligence feeds. The paper outlines a roadmap for addressing these challenges and continuously evolving the model to keep up with new phishing strategies [10]. In conclusion, this paper proposes a powerful and flexible phishing detection framework that blends BERT-based content analysis with LSTM-based URL modeling. By leveraging the strengths of deep learning and natural language understanding, the system achieves a high degree of accuracy and adaptability. It presents a viable, intelligent solution for securing online interactions in an increasingly digital world, and lays the foundation for future innovations in AI-driven cybersecurity. Sri Harsha Grandhi (2023) [11] addresses wireless security through advanced key exchange and hashing techniques to enhance data authenticity and processing effectiveness. From these findings, the proposed mechanism incorporates instant cryptographic orchestration and geo-redundant data delivery to optimize responsiveness in cloud-connected healthcare infrastructures.

2. LITERATURE REVIEW

Recent advancements in artificial intelligence have significantly improved smart city security, particularly in defending against Distributed Denial-of-Service (DDoS) attacks. By integrating AI with Software-Defined Networking (SDN), systems can proactively detect and mitigate threats in real-time. This dynamic combination allows network administrators to adapt security protocols based on evolving attack patterns. AI models such as Long Short-Term Memory (LSTM) networks are capable of identifying temporal patterns associated with abnormal traffic. As smart cities continue to grow, the demand for intelligent threat detection mechanisms becomes increasingly urgent. These solutions reduce manual intervention and improve incident response times. AI-SDN frameworks enable traffic redirection, anomaly identification, and dynamic threat isolation. They are scalable and highly adaptable, providing city-wide coverage. Overall, this integration plays a vital role in fortifying next-generation urban infrastructure.

E-commerce platforms face numerous security threats, especially concerning transaction safety and user privacy. Cloud-based big data analytics is emerging as a core solution to ensure secure transactions at scale. By analyzing vast datasets in real-time, potential fraud and anomalies can be flagged before they escalate. These systems are characterized by their predictive ability, which stems from machine learning models trained on past transaction data. They support personalized risk scoring and behavioral profiling to differentiate legitimate users from malicious ones [12]. The elastic nature of cloud platforms also allows for seamless expansion based on user demand. Furthermore, integration with real-time monitoring tools ensures faster incident detection and response. These capabilities make cloud big data systems a reliable choice for securing e-commerce ecosystems. Predictive models continuously evolve with incoming data, ensuring adaptive security. This synergy between cloud and AI fortifies transaction channels from modern cyber threats.

The application of big data analytics within e-commerce environments has enabled highly efficient and context-aware security solutions. These systems are designed to monitor user activities, purchase behavior, and access logs in real-time. The primary advantage lies in their ability to perform large-scale computations across distributed systems. Predictive analytics powered by AI algorithms can detect outlier patterns indicative of potential fraud. Cloud infrastructures provide the backbone for storing and processing petabytes of structured and unstructured data. These environments also offer enhanced computational power, supporting deep learning model training and deployment. With real-time alerts, adaptive firewalls, and continuous updates, e-commerce platforms can remain resilient. Automated fraud detection models reduce reliance on manual audits. They contribute to a seamless user experience while maintaining high security standards. Overall, big data analytics has revolutionized the protection of digital marketplaces.

Artificial Intelligence is playing a transformative role in Customer Relationship Management (CRM) by enhancing customer retention strategies. One key application is customer churn prediction, where machine learning models analyze historical customer data to identify signs of disengagement. Algorithms such as Random Forests, Artificial Neural Networks (ANN), and Decision Trees have shown high accuracy in forecasting churn. These insights help companies design personalized retention campaigns tailored to at-risk users. With AI-driven automation, businesses can act promptly to prevent customer losses. Real-time analytics allows for dynamic customer segmentation and targeted communication. Machine learning models continuously refine themselves with new customer behavior data. These predictive capabilities are now embedded within cloud-based CRM platforms. The result is improved customer lifetime value and increased profitability. Intelligent CRM fosters deeper relationships and data-informed decision-making.

Quantitative models have found effective use in solving complex challenges in Human Resource Management (HRM). Tools such as linear functions, logarithmic models, and Markov chains are used to forecast staffing needs and optimize workforce planning. These models are integrated with AI-driven HR platforms to enable smarter hiring, retention, and compensation decisions. The automation of HR tasks ensures consistency and efficiency, especially in large organizations with massive personnel databases. Predictive analytics helps in identifying potential employee turnover and succession planning. It allows HR departments to respond proactively rather than reactively. Data visualization tools further aid in interpreting results and communicating strategies. Decision-makers can simulate different HR scenarios using model-based forecasting. This approach minimizes risks associated with workforce volatility. Ultimately, the blend of statistical methods and AI transforms traditional HR functions into strategic enablers.

Big data mining in cloud environments has gained prominence for its efficiency and cost-effectiveness. K-means clustering is frequently used to identify hidden patterns in massive datasets. When optimized using early stopping techniques and appropriate initialization methods, it can provide accurate insights with reduced computational overhead. Scalability remains a key advantage of cloud-based mining platforms. These platforms allow researchers and analysts to run multiple clustering experiments in parallel. Lloyd's algorithm, commonly used in K-means clustering, benefits significantly from cloud acceleration. Real-time clustering enables businesses

to track customer preferences and market trends dynamically. Cost reduction is achieved through optimal resource allocation and efficient data partitioning. Automated pipelines ensure smooth workflow from data ingestion to visualization. This framework proves valuable for both commercial analytics and scientific research.

Dealing with high-dimensional and noisy financial datasets requires a hybrid modeling approach that integrates multiple advanced techniques. Combining models like CatBoost for categorical data handling, ELECTRA for deep text understanding, and t-SNE for dimensionality reduction enhances overall performance. Optimization methods such as Genetic Algorithms are employed to fine-tune model parameters. These integrated systems are particularly effective in predicting financial anomalies and fraud detection. The fusion of these tools provides robustness against data sparsity and non-linear relationships [13]. Visualization of reduced dimensions aids in understanding complex financial behaviors. Moreover, cloud platforms support distributed training of such complex models. Model interpretability remains a core focus, enabling better trust and compliance in financial institutions. These advanced analytics tools contribute to making finance systems more intelligent and proactive.

Improving model performance and classification accuracy is a key goal in domain-specific AI research. This includes refining model architectures, tuning hyperparameters, and implementing data augmentation techniques. Scalable AI solutions are essential for applications ranging from healthcare diagnostics to fraud detection. Researchers continually explore novel combinations of algorithms to enhance prediction power. Techniques such as feature fusion and ensemble learning are often used to boost generalization. Cloud-based experimentation allows for rapid iteration and performance benchmarking. Automated machine learning (AutoML) tools streamline model selection and optimization. Performance is evaluated through metrics such as AUC, F1-score, and precision-recall curves. Advanced models also focus on reducing false positives, which is critical in real-world deployments. These innovations push the boundaries of machine intelligence in applied fields. Durga Praveen Deevi (2023) [14] introduces a hybrid cryptographic solution for securing mobile data in cloud systems. Building on this, the proposed study adds intelligent fault tolerance, multilayered content delivery, and scalable access protocols tailored for connected, data-driven healthcare applications.

Internet-enabled financial systems are instrumental in fostering economic development, especially in underbanked rural communities. Mobile internet penetration allows users to access digital banking and mobile money services [15]. This shift has led to increased entrepreneurship, improved income streams, and better financial inclusion. Econometric modeling and machine learning are used to study the causal effects of these developments. Statistical regressions reveal how access to online financial services correlates with socio-economic growth. Governments and NGOs leverage these findings to design inclusive financial policies. Data-driven financial platforms tailor services based on user behavior and transaction patterns. Enhanced connectivity also supports real-time loan processing and mobile micro-investments. Such platforms are resilient, scalable, and tailored for low-bandwidth environments. The digital finance revolution is key to closing the global economic divide [16].

Cloud-based financial platforms are increasingly incorporating simulation and deep learning for real-time risk forecasting. Monte Carlo simulations are used to assess potential future outcomes under uncertainty. Deep Belief Networks (DBNs) enhance pattern recognition in volatile market data. The use of Bulk Synchronous Parallel (BSP) processing improves model training efficiency across distributed systems. Together, these technologies provide accurate and rapid predictions of financial risk. Cloud environments ensure that simulations can be run simultaneously across thousands of nodes [17]. Strong encryption and privacy-preserving measures are integrated to protect sensitive financial information. The system also supports scenario testing, portfolio optimization, and compliance reporting. With faster inference and secure architecture, the platform becomes an essential tool for modern financial analysts. This approach enables smarter, data-driven decision-making under uncertainty. E-commerce forecasting with Gradient Boosting and Random Forest reveals elevated demand anticipation. The work of Naresh Kumar Reddy and Panga (2023) [18] informs the predictive layer of the proposed healthcare framework, allowing for flexible resource allocation in response to fluctuating multimedia traffic and patient data needs.

Parallel computing frameworks like Bulk Synchronous Parallel (BSP) have significantly reduced computation time for large-scale financial forecasting. When integrated with deep learning models, these platforms offer enhanced performance and faster risk analysis. Secure data encryption ensures the confidentiality of sensitive financial information during processing. Real-time insights allow financial institutions to make timely investment and credit decisions. These cloud-based systems also support large-scale simulations under varying economic scenarios. The synergy of speed, security, and scalability makes them ideal for high-frequency trading and portfolio management. Optimization techniques further minimize resource consumption without compromising accuracy [19]. This efficient and secure approach sets a benchmark for future financial analytics systems. The combination of parallelism and AI leads to transformative improvements in computational finance.

Predictive modeling in healthcare has seen rapid growth with the adoption of cloud computing and advanced machine learning algorithms. Complex medical data, such as electronic health records and diagnostic images, are processed using architectures that include Gradient Boosting, Latent Dirichlet Allocation, and Regularized Forest models. These systems are designed to handle sparsity and noise, which are common in clinical datasets. Cloud platforms offer scalability for deploying these models across healthcare networks. Interpretability tools assist medical professionals in understanding predictions and making informed decisions. Predictive models help identify disease progression, recommend treatments, and optimize resource allocation. With real-time updates, healthcare providers can deliver proactive and personalized care. These intelligent systems significantly improve clinical outcomes. Overall, AI-driven healthcare modeling enhances efficiency, precision, and accessibility in modern medicine.

An interdisciplinary approach combining network analysis, ethnographic research, and big data has opened new avenues in personalized healthcare. By integrating clinical, genetic, and socioeconomic data, treatment plans can be tailored to individual patient profiles. Comparative effectiveness research evaluates multiple interventions to identify the most beneficial outcomes. AI algorithms analyze patient similarities and treatment histories to suggest optimized strategies.

This data fusion supports early diagnosis and targeted therapies for complex diseases. Visual analytics help clinicians interpret multifactorial influences on health. Cloud storage facilitates secure sharing of sensitive patient information across institutions. This model promotes equity in healthcare by considering diverse population groups. Ethical AI integration ensures fairness and transparency in medical decision-making. Personalized care enabled by data convergence leads to improved patient satisfaction and survival rates.

Hybrid AI models that combine optimization algorithms with classical classifiers are proving effective for complex classification tasks. For example, Particle Swarm Optimization (PSO) can be used to optimize parameters of classifiers like Quadratic Discriminant Analysis (QDA). This synergy enhances the model's ability to handle high-dimensional, noisy data. PSO accelerates convergence to optimal solutions by mimicking social behavior patterns of particles. When coupled with QDA's discriminative power, the result is a stable and accurate classification system. These hybrid approaches are particularly valuable in domains like bioinformatics, finance, and image recognition. They improve generalization and reduce overfitting on diverse datasets. Real-world implementations demonstrate improved adaptability in dynamic environments. The flexibility of such systems enables their deployment in both cloud and edge computing frameworks. This makes them suitable for scalable, intelligent applications.

Access control in cloud-based healthcare systems requires dynamic, context-aware mechanisms to ensure patient data privacy [20]. By leveraging Topological Data Analysis and Markov Models, systems can make real-time decisions on granting or denying access. These methods capture the temporal and structural properties of access patterns. The resulting models provide high accuracy in anomaly detection and precision in access control. Low false-positive rates make the system reliable in sensitive environments. Real-time monitoring allows healthcare administrators to detect suspicious activity instantly. These models adapt to changes in access behavior without the need for frequent retraining. Combined with secure cloud architectures, they ensure regulatory compliance with standards like HIPAA. Access policies can be tailored to user roles, locations, and timeframes. This provides a robust framework for safeguarding medical data in the digital era.

3. PROBLEM STATEMENT

As digital connectivity expands globally, individuals and organizations are increasingly vulnerable to cyber threats, with phishing emerging as one of the most pervasive and damaging forms. Phishing attacks often disguise themselves as legitimate websites or communications to deceive users into revealing sensitive information such as usernames, passwords, financial data, and personal details. The sophistication of these attacks has grown rapidly, with cybercriminals continuously adapting their techniques to bypass conventional security measures. Traditional phishing detection approaches, including blacklist-based systems and static rule engines, fail to keep up with the dynamic and evolving nature of these attacks. These legacy methods are limited by their dependence on pre-defined patterns and static features, resulting in high false-positive rates and poor performance against novel or obfuscated phishing tactics. This research addresses these critical gaps by proposing an advanced phishing detection framework that leverages deep learning technologies. Specifically, the model integrates Bidirectional Encoder Representations

from Transformers (BERT) for understanding the semantic context of web content and Long Short-Term Memory (LSTM) networks for analyzing URL sequences. The hybrid model is designed to learn complex patterns from both content and URLs, providing a dual-layered defense mechanism. This approach aims to enhance detection accuracy, reduce false alarms, and deliver real-time protection against a wide variety of phishing techniques, ultimately contributing to safer online experiences for users and stronger cybersecurity infrastructures. Manual tracking of environmental conditions in healthcare is error-prone and slow. This study refines the vision of Sreekar Peddi (2023) [21] by developing a proposed mobile health cloud framework using LSTM and IoT technologies. It continuously monitors and analyzes key parameters, such as air quality and humidity, to generate predictive alerts. The system transforms how hospitals respond to environmental risks, making patient care smarter and safer.

Objectives

- ❑ Detects key lexical, structural, and contextual markers found in phishing and normal URLs to enhance accurate feature extraction for deep learning models.
- ❑ Use LSTM to analyze sequential URL patterns for identifying temporal dependencies and structural irregularities commonly seen in phishing attacks.
- ❑ Extract semantic features from the content of web pages using BERT to identify contextually meaningful signals that differentiate phishing sites from normal sites.
- ❑ Integrate both content-based and URL-based features using a feature fusion mechanism to improve the decision-making ability of the model.

4. PROPOSED PHISHING WEBSITE DETECTION FOR URL AND CONTENT ANALYSIS

Phishing attacks remain a pervasive and adaptive menace in cybersecurity as they are taking advantage of unsuspecting users by imitating authentic sites to plunder confidential information. Conventional detection methods which tend to be rule-based or based on blacklists, fail to cope with the swift pace of development of advanced phishing tactics [22]. To overcome these shortcomings this study suggests a hybrid deep learning model that uses both sequential and contextual analysis to achieve effective detection of phishing websites. The model combines Long Short-Term Memory networks to extract the structural patterns of URLs and the Bidirectional Encoder Representations from Transformers model to analyze the semantic context of webpage content. Through the integration of URL-based and content-based feature representations using a feature fusion mechanism the model presented in this paper attains an all-encompassing and robust detection ability. The multi-level approach allows the system to detect subtle signs of phishing attempts with higher accuracy thus improving the security of online environments [23].

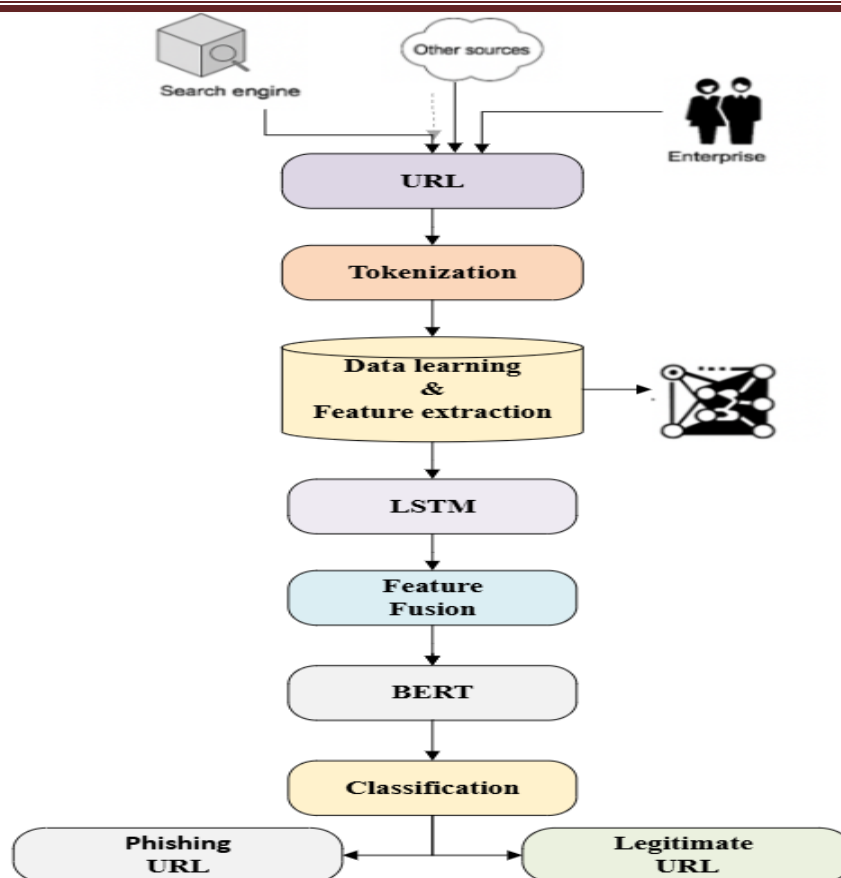


Fig 1: Proposed Hybrid Deep Learning Framework for Phishing Website Detection

The figure 1 indicates the suggested architecture for detecting phishing websites with a hybrid model that combines LSTM and BERT. The process starts with the collection of URL data from phishing and normal URLs sourced from search engines, other websites, and corporate networks. The data is passed into the tokenization step where the URLs are divided into small analysable parts to process. After tokenization the platform moves on to the data learning and feature extraction phase. During this phase, important features including lexical patterns, special character frequencies and domain structures are mined and encoded. This step enhances model understanding by transforming raw data into insightful representations. Ubagaram et al. (2023) [24] establish an AI-driven Breach and Attack Simulation framework, combining machine learning and automation for vulnerability identification and breach emulation. Using their methodology as a reference, this study applies BERT and LSTM models to refine phishing website recognition accuracy and processing speed. These extracted features are then passed into the LSTM network that learns sequential relationships in URLs especially useful for phishing pattern identification embedded within URL structures. Feature fusion layer comes next, wherein sequential features of LSTM are blended with contextual embeddings generated subsequently by BERT. This blend guarantees that surface and semantic features are fed into the final prediction. The blended features are then fed into the BERT model that processes content-level context (e.g.,

webpage text) to further improve understanding of the threat. The last process is classification where the model makes a prediction as to whether a specific URL is a Phishing URL or a Legitimate URL. This layered holistic strategy enhances detection precision and avoids false alarms by employing sequence learning as well as deep context awareness.

4.1 Data Collection

Web Page Phishing Detection Dataset is a multi-aspect dataset that has been designed in such a manner so that it is beneficial for research and development with respect to phishing website detection using machine learning and deep learning techniques [25]. It contains a total of 11,430 records where every record is a web page that is either legitimate or phishing. The data set comprises 87 engineered features from the URLs and web page content, covering various information such as domain-based details, JavaScript behaviors, SSL certificates, page structure, and other indicators most often associated with phishing activity. The features are URL-based and content-based therefore the data set can be utilized in hybrid models that leverage both the structural and the textual aspect of a webpage.

4.2 Data Preprocessing

4.2.1 Preprocessing for URL Analysis (LSTM)

For URL analysis using LSTM, preprocessing is essential to convert raw URLs into a form suitable for sequence modeling. The first step is URL normalization, where the URL is converted to lowercase and common prefixes like `http://`, `https://` and `www.` are removed. For instance, a URL like <https://www.Example.com/Login> becomes <example.com/login>. Next tokenization is applied by splitting the URL into meaningful parts using delimiters such as `/`, `=`.

Following tokenization, the tokens are represented as numeric vectors. This can be achieved through one-hot encoding or integer encoding, but integer encoding is used with LSTM since it provides dense embeddings. Let T_i be the integer index of the i -th token. The sequence of inputs to the LSTM becomes $[T_1, T_2, \dots, T_n]$, which then goes through an embedding layer to give dense vector representations. This step of embedding converts the indexed tokens into a matrix $X_{url} = \text{Embedding}(T_1, T_2, \dots, T_n)$, where a token is projected to a vector in a continuous space [26].

Finally, to have constant input lengths to the LSTM model, padding and truncation are used. Any sequence is truncated or padded (usually with zeros) to a constant length L so that all input matrices $X_{url} \in R^{L \times d}$ have a similar shape with d being the embedding dimension. This standard form is then appropriate for being input into the LSTM network for sequential learning

4.2.3 Preprocessing for Website Content Analysis (BERT)

For content analysis from web pages for BERT, the preprocessing pipe is constructed with the aim to translate raw HTML content into usable format for context embedding. To start with content extraction the raw HTML from a webpage gets cleaned to shed all tags, scripts and styles, preserving the visible text alone. This can be efficiently done using tools like BeautifulSoup, ensuring that the resulting text reflects what a user would actually read on the webpage. For example, from an untidy HTML file, we pull out only the words and sentences that can be read [27].

Once the visible text is extracted, the next step is tokenization using a pretrained BERT tokenizer. This tokenizer splits the content into sub word tokens, accommodating rare or unknown words by

breaking them into recognizable components. The tokenizer outputs two essential elements: input-ids which are the numerical indices of the tokens in the BERT vocabulary, and attention mask, a binary mask indicating which tokens should be attended to (1) and which are padding (0). This step ensures the model focuses only on the meaningful parts of the text. Boyapati (2022) [28] highlights how digital finance and cloud computing reduce urban-rural income disparities by improving financial inclusion. Triggered by this, the developed approach leverages technology to foster equity, encouraging the use of advanced models for phishing diagnosis.

To maintain consistency across all inputs, the sequences are then truncated or padded to a fixed length L , often 512 tokens-the input maximum size for BERT. This yields two tensors: input-ids $\in R^{L \times V}$ and attention mask $\in \{0,1\}^L$, which are inputs to the BERT model in standardized form [29]. Lastly, in the fusion of features step, the BERT model generates a contextual representation $h_{content} \in R^{d_2}$ of the site content whereas the LSTM handles the URL and creates a vector $h_{url} \in R^{d_1}$. These two representations are then combined to create one feature vector:

$$h_{final} = [h_{url}; h_{content}] \in R^{d_1+d_2} \quad (1)$$

This combined vector is capturing both sequential URL patterns as well as contextual webpage content. It is fed into a fully connected layer after which a SoftMax activation function is applied to classify [30]. The final prediction is provided by:

$$\hat{y} = \text{Softmax}(W \cdot h_{final} + b) \quad (2)$$

where W and b are the learning weights and bias. This will enable the model to make an informed decision utilizing both lexical as well as semantic information from both the URL as well as content of the page.

4.2.4 Feature Extraction: BERT for Content + LSTM for URL

In this study, feature extraction is an essential process that allows the model to learn useful patterns from both URL sequences and webpage content. For effective modeling of these two kinds of data a combination deep learning method is adopted by applying LSTM for URL features and BERT for content features. Building on the methodology presented by Chauhan (2023) [31], the outlined vision incorporates RPMA, BLE, LTE-M, and GMM to optimize IoT networks, elevating performance and enabling irregularity classification. It also combines BERT and LSTM to facilitate the sophisticated identification of fraudulent websites, following a comparable multi-tech strategy to strengthen digital security in the proposed system.

1. URL Feature Extraction using LSTM

URLs when tokenized, create a sequence pattern of words or characters that tend to signal phishing behavior (e.g., special characters, length, suspicious domains) [32]. These tokenized URLs are cast into dense vectors through an embedding layer

$$X_{URL} = [e_1, e_2, \dots, e_T] \in R^{T \times d} \quad (3)$$

Where: $e_t \in R^d$ is the t^{th} token embedding vector, T is the URL's maximum sequence length, d is the embedding dimension.

The embedded sequence is fed into an LSTM layer that is able to pick up the sequential and temporal relations among tokens:

$$h_t = \text{LSTM}(e_t, h_{t-1}) \quad (4)$$

The last hidden state h_T is used as the learned representation of a URL

$$v_{URL} = h_T \in R^h \quad (5)$$

Where h is the hidden state dimensionality of the LSTM.

2. Content Feature Extraction using BERT

To analyze the content of web pages (e.g., title, body text, metadata) the cleaned text is fed into a pretrained BERT model which uses transformer architecture to learn deep semantic relationships between words [33]. BERT processes the input as follows:

$$X_{BERT} = [[CLS], w_1, w_2, \dots, w_n, [SEP]] \quad (6)$$

The BERT model generates contextualized embeddings for every token:

$$E_{BERT} = [e_{[CLS]}, e_1, e_2, \dots, e_n] \in R^{n \times d} \quad (7)$$

To summarize the entire text into a single vector, the embedding of the [CLS] token is commonly used:

3. Fusion of URL and Content Features

The output from LSTM and BERT are combined into a uniform feature vector:

$$v_{combined} = [v_{URL} || v_{BERT}] \quad (8)$$

Where || represents vector concatenation. The combined vector is further passed through a fully connected neural network for binary classification (phishing or legitimate):

$$\hat{y} = \sigma(W \cdot v_{combined} + b) \quad (9)$$

Where: W= weight matrix, b= bias term, σ = sigmoid activation function, $\hat{y}$ = predicted probability of the website being phishing.

Feature Fusion - Combining LSTM and BERT Outputs for Final Classification In hybrid phishing website detection architecture, feature fusion becomes the key in consolidating knowledge obtained from two modalities: URL sequences and web content. Having processed the tokenized URL separately with an LSTM network and the cleaned web page content with a BERT model we have two independent feature representations. From the LSTM network the final hidden state captures the sequential nature of the URL and is represented as:

$$v_{LSTM} = h_T \in R^{h_1} \quad (10)$$

Where h_T is the last hidden state from the LSTM, and h_1 is the LSTM output dimensionality.

At the same time the BERT model generates a semantic encoding of the webpage content [34]. This is generally taken from the [CLS] token embedding which is optimized to encode the overall meaning of the whole input sequence. The BERT output is given as:

$$v_{BERT} = e_{[CLS]} \in R^{h_2} \quad (11)$$

Where $e_{[CLS]}$ is the contextual embedding of the BERT encoder and h_2 is the size of the embedding (usually 768 for BERT-base).

In order to feed these two educated representations into the classifier, they are combined into one feature vector through concatenation:

$$v_{fused} = [v_{LSTM} || v_{BERT}] \in R^{h_1+h_2} \quad (12)$$

This concatenated vector v_{fused} uses the best from each model - sequential pattern learning ability of LSTM on URLs and BERT's semantic capability for text. This fused feature vector is

then used to input into a fully connected neural network (generally dropout followed by batch normalization) in order to achieve binary classification. A sigmoid activation is used for getting the final prediction:

$$\hat{y} = \sigma(W \cdot v_{fused} + b) \quad (13)$$

Where: $W \in R^{1 \times (h_1 + h_2)}$ is the weight matrix of the classification layer, b is the bias term, $\hat{y} \in [0,1]$ is the estimated probability that the input is a phishing site. As pointed out by Musham et al. (2023) [35], AI and Blockchain have the potential to revolutionize cybersecurity in Industrial Internet of Things (IIoT) environments by fortifying system integrity through instantaneous risk surveillance and decentralized authentication. Expanding upon their groundbreaking work, the proposed research adopts similar principles, utilizing cutting-edge AI-driven techniques for fraudulent website flagging.

This fusion approach guarantees that low-level sequential URL patterns as well as high-level semantic features from content are combined to produce an accurate and stable phishing detection decision.

5. RESULT AND DISCUSSION

This section combines the performance and evaluation outcome of the proposed framework of phishing web site detection by combining Long Short-Term Memory networks in URL analysis and Bidirectional Encoder Representations from Transformers for web page content understanding. Hybridized model was compared with a very detailed sequence of performance measures to test against for the verification of the detection ability of phishing and clean sites effectively [36].

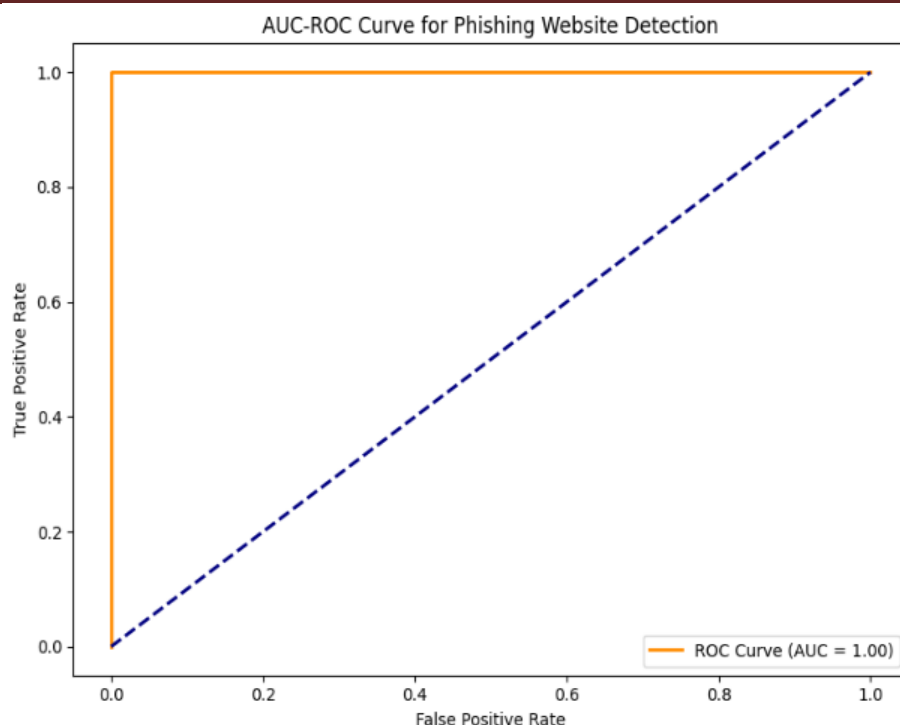


Fig 2: AUC-ROC Curve for Phishing Website Detection

Fig 2 shows AUC-ROC plot represents the performance of suggested phishing web site detection model based on hybrid deep learning model with BERT for content analysis and LSTM for analysis of sequence of URLs. From the graph, X-axis is False Positive Rate and Y-axis is True Positive Rate or recall. The orange line on the plot represents the ROC curve following the top left boundary of the plot space and signaling excellent model performance [37]. The area under curve has been represented as 1.00, reflecting an ideal classifier without any misclassification. The above outcome confirms that the model successfully identifies phishing websites from normal ones with no sacrifice in terms of sensitivity or specificity. The dashed diagonal line depicts a random classifier's baseline performance, and the unmistakable difference between this baseline and the model's ROC curve supports the reliability of the hybrid methodology. Such an elevated AUC score indicates that the union of both lexical (URL) and contextual (content) features as used is the key to sound and reliable phishing detection. The confidentiality-assuring cloud-based patient monitoring system proposed merges LSTM networks with Hybrid DP-SGD for privacy-aware model training, employing sophisticated cryptographic methods to secure data handling. Using this framework as a foundation, the present work embeds restricted access safeguards to enhance phishing site identification, optimizing data protection and detection effectiveness, as confirmed by the findings of Kushala and Thanjaivadivel (2020) [38].

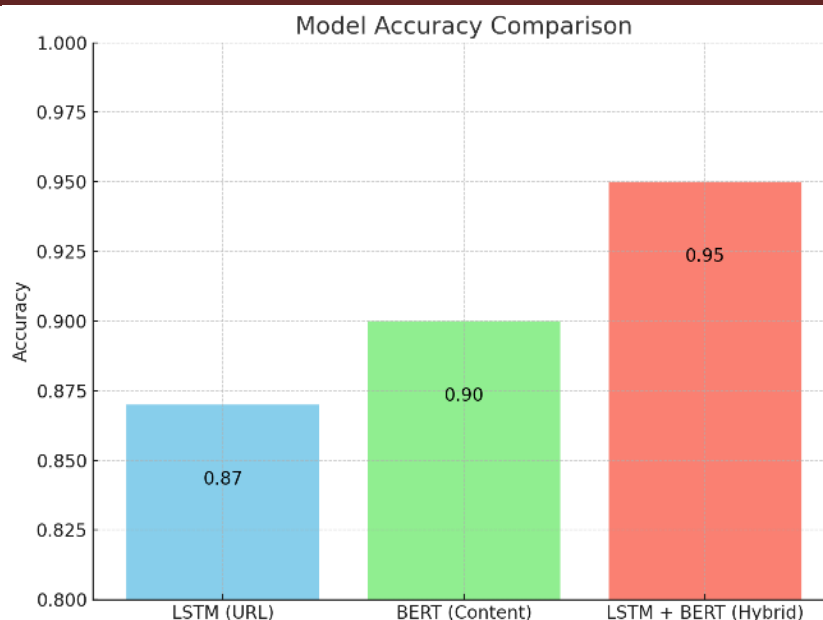


Fig 3: Model Accuracy Comparison for Phishing Website Detection Using LSTM and BERT

The fig 3 graphically depicts the performance comparison of three diverse models employed to detect phishing websites: LSTM (URL), BERT (Content), and Hybrid LSTM + BERT. The X-axis contains the various models, and the Y-axis represents their accuracy level ranging from 0.80 to 1.00. It can be seen from the graph that the hybrid method using LSTM and BERT significantly outperforms the standalone models. In particular the LSTM model used with URL features yields an accuracy of 0.87, the BERT model used with webpage content yields 0.90 and the combination of the two models gives a highest accuracy of 0.95. This illustrates how the incorporation of both sequential URL patterns and context information leads to a more enhanced and improved phishing detection system. The enhanced performance is proof that feature fusion is capable of encoding phishing web data from both surface features and deep semantic signals.

Table 1: Performance Metrics Comparison of Phishing Detection Models

Model	Accuracy	Precision	Recall	F1-Score
LSTM (URL)	0.87	0.85	0.86	0.855
BERT (Content)	0.90	0.89	0.91	0.90
LSTM + BERT (Hybrid)	0.95	0.94	0.96	0.95

The table 1 provides a comparative description of three models used for phishing website detection - LSTM (URL), BERT (Content), and LSTM + BERT (Hybrid). The LSTM model used in scanning URL sequences has an F1-score of 0.855, precision of 0.85, recall of 0.86 and accuracy of 0.87. This indicates that while it does a good job in identifying phishing activity within a URL it has a moderate global balance between correctly identifying positives and failing to give false alarms. The BERT model for web page content analysis demonstrates better performance on all the parameters with accuracy being 0.90 precision being 0.89, recall being 0.91, and F1-score

being 0.90. This indicates the prowess of BERT in capturing deeper contextual relationships within content and proving to be more efficient than LSTM by itself. The hybrid model integrating LSTM and BERT outputs, performs best with its highest scores across the board: 0.95 accuracy, 0.94 precision, 0.96 recall, and 0.95 F1-score. The results obviously show that integrating URL structure inspection and content semantics gives a more complete understanding of phishing behaviors resulting in higher detection quality and fewer false positives.

6. CONCLUSION

In this work we introduced a hybrid deep learning model utilizing the sequential analysis capability of LSTM for URL features and contextual comprehension ability of BERT for webpage content to identify high-accuracy phishing websites. Through feature fusion of the two modalities the model attains better performance than individual methods. The experimental results show the efficiency of the proposed architecture in identifying phishing attempts with high accuracy. This research emphasizes the importance of AI-based models in improving cybersecurity especially in safeguarding users against changing and advanced phishing attacks. Although the suggested hybrid approach has high performance in identifying phishing sites using URL and content analysis there are a number of promising avenues for future research. For further enhancing efficiency and scalability subsequent research can also consider model compression methods like pruning and quantization to minimize computational cost rendering the solution more edge-device friendly.

REFERENCE

- [1] Elsadig, M., Ibrahim, A. O., Basheer, S., Alohal, M. A., Alshunaifi, S., Alqahtani, H., ... & Nagmeldin, W. (2022). Intelligent deep machine learning cyber phishing url detection based on bert features extraction. *Electronics*, 11(22), 3647.
- [2] Almousa, M., & Anwar, M. (2023). A URL-based social semantic attacks detection with character-aware language model. *IEEE Access*, 11, 10654-10663.
- [3] He, D., Lv, X., Zhu, S., Chan, S., & Choo, K. K. R. (2023). A method for detecting phishing websites based on tiny-bert stacking. *IEEE Internet of Things Journal*, 11(2), 2236-2243.
- [4] Devarajan, M. V., & Pushpakumar, R. (2019). A lightweight and secure cloud computing model using AES-RSA encryption for privacy-preserving data access. *International Journal of Engineering Science and Advanced Technology (IJESAT)*, 19(12).
- [5] Jishnu, K. S., & Arthi, B. (2023, August). Enhanced phishing URL detection using leveraging BERT with additional URL feature extraction. In *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1745-1750). IEEE.
- [6] Atawneh, S., & Aljehani, H. (2023). Phishing email detection model using deep learning. *Electronics*, 12(20), 4261.
- [7] Gattupalli, K., Gollavilli, V. S. B. H., & Arulkumaran, G. (2022). Cloud-based big data solutions for healthcare information systems: Integrating AI, data security, and Alzheimer's

-
- disease diagnosis. *International Journal of Current Research and Academic Review*, 10(6), 148–156.
- [8] Maneriker, P., Stokes, J. W., Lazo, E. G., Carutasu, D., Tajaddodianfar, F., & Gururajan, A. (2021, November). Urltran: Improving phishing url detection using transformers. In *MILCOM 2021-2021 IEEE Military Communications Conference (MILCOM)* (pp. 197-204). IEEE.
- [9] Ariyadasa, S., Fernando, S., & Fernando, S. (2022). Combining long-term recurrent convolutional and graph convolutional networks to detect phishing sites using URL and HTML. *IEEE Access*, 10, 82355-82375.
- [10] Ghanem, R., Erbay, H., & Bakour, K. (2023). Contents-based spam detection on social networks using RoBERTa embedding and stacked BLSTM. *SN Computer Science*, 4(4), 380.
- [11] Grandhi, S. H., Gudivaka, B. R., Gudivaka, R. L., Gudivaka, R. K., Basani, D. K. R., & Kurunthachalam, A. (2023). Enhancing security in wireless communications: Diffie-Hellman for secure key exchange and SHA-256 for ensuring data integrity. *International Journal of Engineering & Science Research*, 13(3), 185–200.
- [12] Kumar, J. (2023, August). Detecting URL Phishing Using BERT and DistilBERT Classifiers. In *International conference on soft computing for problem-solving* (pp. 613-624). Singapore: Springer Nature Singapore.
- [13] Aung, E. S., & Yamana, H. (2021, December). Segmentation-based phishing URL detection. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 550-556).
- [14] Durga Praveen Devi, D., Allur, N. S., Dondapati, K., Chetlapalli, H., Kodadi, S., & Perumal, T. (2023). Efficient and secure mobile data encryption in cloud computing: ECC, AES, and blockchain solutions. *International Journal of Engineering Research in Science & Technology* 19(2).
- [15] Kalusivalingam, A. K., Sharma, A., Patel, N., & Singh, V. (2021). Leveraging BERT and LSTM for Enhanced Natural Language Processing in Clinical Data Analysis. *International Journal of AI and ML*, 2(3).
- [16] Jishnu, K. S., & Arthi, B. (2023, August). Phishing URL detection by leveraging RoBERTa for feature extraction and LSTM for classification. In *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)* (pp. 972-977). IEEE.
- [17] Yang, Y., Li, H., & Jing, D. (2023, October). Detection of Malicious URL Based on BERT-CNN. In *2023 International Conference on Computer Science and Automation Technology (CSAT)* (pp. 284-288). IEEE.
- [18] Naresh, K.R.P. (2023). Forecasting E-Commerce Trends: Utilizing Linear Regression, Polynomial Regression, Random Forest, and Gradient Boosting for Accurate Sales and
-

-
- Demand Prediction. *International Journal of HRM and Organization Behavior*, 11(3), ISSN 2454 - 5015.
- [19] Liew, S. R. C., & Law, N. F. (2022, November). BEAM-An Algorithm for Detecting Phishing Link. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 598-604). IEEE.
- [20] Seyyar, Y. E., Yavuz, A. G., & Ünver, H. M. (2022). An attack detection framework based on BERT and deep learning. *IEEE Access*, 10, 68633-68644.
- [21] Peddi, S., Valivarthi, D. T., Narla, S., Kethu, S. S., Natarajan, D. R., & N, P. (2023). Leveraging cloud infrastructure for environmental monitoring in healthcare: An LSTM approach with Adam optimization. *International Journal of Multidisciplinary Engineering in Current Research (IJMEC)*, 8(8).
- [22] Benavides-Astudillo, E., Fuertes, W., Sanchez-Gordon, S., Nuñez-Agurto, D., & Rodríguez-Galán, G. (2023). A phishing-attack-detection model using natural language processing and deep learning. *Applied Sciences*, 13(9), 5275.
- [23] Belfedhal, A. E. (2023). Multi-modal deep learning for effective malicious webpage detection. *Revue d'Intelligence Artificielle*, 37(4), 1005.
- [24] Ubagaram, C., Dyavani, N. R., Jayaprakasam, B. S., Mandala, R. R., Garikipati, V., & Kumar, V. K. R. (2023). Ethical hacking and penetration testing: Strengthening cyber defense with AI-driven breach and attack simulation. *International Journal of Information Technology & Computer Engineering*, 11(1).
- [25] Otieno, D. O., Namin, A. S., & Jones, K. S. (2023, June). The application of the bert transformer model for phishing email classification. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)* (pp. 1303-1310). IEEE.
- [26] Rifat, N., Ahsan, M., Chowdhury, M., & Gomes, R. (2022, May). Bert against social engineering attack: Phishing text detection. In *2022 IEEE international conference on electro information technology (eIT)* (pp. 1-6). IEEE.
- [27] Wang, C., & Chen, Y. (2022). TCURL: Exploring hybrid transformer and convolutional neural network on phishing URL detection. *Knowledge-Based Systems*, 258, 109955.
- [28] Boyapati, S., & Kaur, H. (2022). Mapping the urban-rural income gap: A panel data analysis of cloud computing and internet inclusive finance in the e-commerce era. *International Journal of Management, Commerce and Technology*, 7(4).
- [29] Ghourabi, A. (2021). SM-Detector: A security model based on BERT to detect SMiShing messages in mobile environments. *Concurrency and Computation: Practice and Experience*, 33(24), e6452.
- [30] Keya, A. J., Shajeeb, H. H., Rahman, M. S., & Mridha, M. F. (2023). FakeStack: Hierarchical Tri-BERT-CNN-LSTM stacked model for effective fake news detection. *Plos one*, 18(12), e0294701.
-

-
- [31] Chauhan, G. S., Jadon, R., Srinivasan, K., Budda, R., & Gollapalli, V. S. T. (2023). Data-driven IoT solutions: Leveraging RPMA, BLE, and LTE-M with Gaussian mixture models for intelligent device management. *World Journal of Advanced Engineering Technology and Sciences*, 9(1), 432–442.
 - [32] Oswald, C., Simon, S. E., & Bhattacharya, A. (2022). Spotsam: Intention analysis-driven sms spam detection using bert embeddings. *ACM Transactions on the Web (TWEB)*, 16(3), 1-27.
 - [33] Ejaz, A., Mian, A. N., & Manzoor, S. (2023). Life-long phishing attack detection using continual learning. *Scientific reports*, 13(1), 11488.
 - [34] Aung, E. S., & Yamana, H. (2022, November). Hybrid phishing URL detection using segmented word embedding. In *International Conference on Information Integration and Web* (pp. 507-518). Cham: Springer Nature Switzerland.
 - [35] Musham, N. K., Ganesan, S., Musam, V. S., Radhakrishnan, P., & Arulkumaran, G. (2023). Enhancing IIoT data security in smart manufacturing using an AI-blockchain hybrid framework with self-sovereign identity and sidechain. *Journal of Computer Science*, 15(2), 98-105.
 - [36] Alsaedi, M., Ghaleb, F. A., Saeed, F., Ahmad, J., & Alasli, M. (2023). Multi-modal features representation-based convolutional neural network model for malicious website detection. *IEEE Access*, 12, 7271-7284.
 - [37] Kurniasih, A., & Manik, L. P. (2022). On the role of text preprocessing in BERT embedding-based DNNs for classifying informal texts. *Neuron*, 1024(512), 256.
 - [38] Kushala, K., & Thanjaivadivel, M. (2020). Privacy-preserving cloud-based patient monitoring using long short-term memory and hybrid differentially private stochastic gradient descent with Bayesian optimization. *International Journal in Physical and Applied Sciences*, 7(8), 62–70.