A Framework for Multilingual Phishing Detection: Integrating Hybrid Feature Engineering and Transformer Models

1st Basit Ali
Dept. of Information and Communication Engineering
The Islamia University of Bahawalpur
Bahawalpur, Pakistan
basit73716@gmail.com

2nd Ali Hassan
Dept. of Information and Communication Engineering
The Islamia University of Bahawalpur
Bahawalpur, Pakistan
ali.hassan6465@gmail.com

3rd Hassam Mehmood
Dept. of Information and Communication Engineering
The Islamia University of Bahawalpur
Bahawalpur, Pakistan
m.hassam3092@gmail.com

4th Farhan Hassan
Dept. of Information and Communication Engineering
The Islamia University of Bahawalpur
Bahawalpur, Pakistan
Farhan.hassan@iub.edu.pk

***Abstract*—Phishing attacks have become a global cybersecurity threat, extending beyond English-speaking regions. Traditional detection methods relying on blacklists, English-only datasets, or manual features struggle with multilingual phishing. This study presents a hybrid detection framework that combines feature-based machine learning with multilingual Transformer models, such as mBERT, leveraging its semantic understanding. The framework evaluates phishing detection across English, German, and Chinese datasets, integrating mBERT's contextual analysis with lexical, host-based, and content-level phishing indicators. Unlike prior research focused on English datasets, this work emphasizes cross-lingual adaptability and real-time deployment via a web application. Experimental results demonstrate over 95% accuracy, supporting scalability for browser extensions and enterprise solutions.**

***Index Terms*—Multilingual Phishing Detection, Hybrid Models, Transformer (BERT), NLP, Cybersecurity.**

I. Introduction

Phishing remains a predominant cybercrime, inflicting billions in damages on individuals and businesses. While effective countermeasures exist for English phishing, attackers increasingly target non-English-speaking regions, such as those using Chinese, German, and Spanish, exploiting gaps in global defenses [1]. Most current detection systems rely on blacklists and heuristic rules, which suffer from delayed responses and poor adaptability to linguistic diversity [2].

Artificial intelligence (AI) and machine learning (ML) have been explored for proactive phishing detection [3], [4]. Techniques using Random Forest and BERT-based Transformer models have shown high accuracy, yet a critical limitation persists: most advanced models are trained solely on English datasets, restricting their global applicability [5]. This study proposes a multilingual phishing detection framework to address this gap, with objectives including the creation of a multilingual corpus, development of a hybrid ML-Transformer system, and real-time implementation with a user-friendly interface.

II. Related Work

Traditional phishing detection has leaned on feature-based ML approaches. Krishna et al. [6] demonstrated that URL analysis with ML techniques achieves significant accuracy, focusing on lexical features like URL length and domain age. Desai and Shah [7] validated these findings in a comprehensive study on phishing website detection.

The advent of transformer-based models has revolutionized phishing detection. Li et al. [8] showcased BERT's effectiveness in detecting email phishing by leveraging contextual knowledge, a focus of recent cybersecurity research [9]. Multilingual capabilities have gained traction, with Pires et al. [10] exploring BERT's cross-linguistic potential, laying the groundwork for multilingual phishing detection [11]. Recent hybrid approaches, such as those by Das Gupta et al. [12], Elsadig et al. [13], and Barik et al. [14], combine ML and deep learning for enhanced detection.

III. Methodology

This section outlines a structured, multi-phase approach with detailed implementation specifics.

A. Data Acquisition and Curation

Datasets were sourced from public repositories (PhishTank, Kaggle) and curated collections [15], including phishing and legitimate samples in English, German, and Chinese. To address underrepresented languages, custom web scraping was performed using Python's BeautifulSoup4, targeting phishing

DOI: https://doi.org/10.67535/tsp.000001.043
Published under the ICACNC 2025 Proceedings

forums and legitimate websites. Preprocessing involved removing HTML noise, normalizing text with NFKC Unicode, and balancing datasets with SMOTE oversampling, resulting in 10,000, 7,000, and 8,000 samples for English, Chinese, and German, respectively.

B. Hybrid Feature Engineering

Feature extraction spanned three domains:

- **Lexical/URL Features**: URL length, special character count, and domain age extracted using WHOIS data [6].
- **Host-Based Features**: SSL certificate validity and IP geolocation analyzed via OpenSSL and MaxMind APIs [7].
- **Content-Based Features**: Multilingual text processed with Hugging Face's WordPiece tokenizers, preserving linguistic nuances [10].

C. Model Architecture

The dual-strategy approach included:

- **Traditional ML Pipeline**: Random Forest, SVM, and XGBoost classifiers were tuned with grid search (e.e., RF max depth: 10-20, $n_estimators$: $100 - 300)1001[17]$.**Deep Learning Pipeline** : $mBERTwasfine - tunedwithalearningrateof2e - 5, batchsizeof16, and3epochs, usingcross - entropyloss$ $1001[10]$. Models were ensembled via a weighted voting mechanism.

D. System Implementation

A web application was developed using Flask (backend) and React (frontend), enabling real-time URL/text analysis. Deployment utilized a REST API with Nginx for load balancing, tested on a server with 16GB RAM and an NVIDIA GTX 1080.

E. Evaluation Framework

Performance was assessed using 5-fold cross-validation, measuring Accuracy, Precision, Recall, F1-Score, and AUC-ROC, with statistical significance tested via paired t-tests (p ¡ 0.05) [8].

IV. Experimental Results and Discussion

Experimental validation was conducted on the curated datasets, with results presented below.

TABLE I
Dataset Statistics and Distribution

Language	Phishing Samples	Legitimate Samples	Total
English	5,000	5,000	10,000
Chinese	3,500	3,500	7,000
German	4,000	4,000	8,000

-

A. Multilingual Data Availability

Challenge: Sourcing high-quality non-English datasets was constrained by limited public availability [19]. Solution: Web scraping and SMOTE oversampling ensured balanced representation.

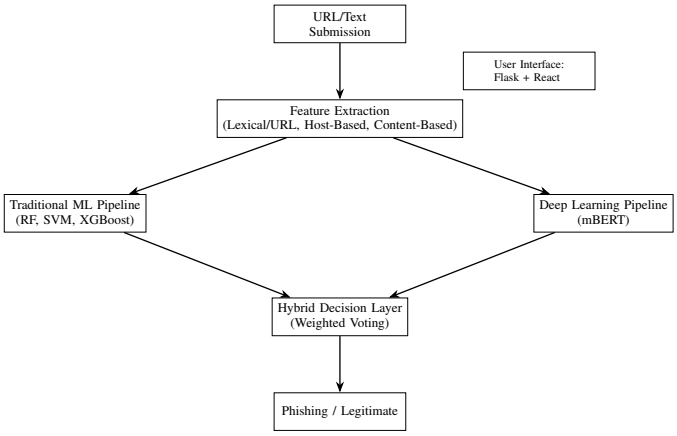

Fig. 1. Proposed system architecture for the hybrid phishing detection framework.

TABLE II
Empirical Performance Metrics (Mean ± SD)

Language	Accuracy	Precision	Recall	F1-Score	AUC
English	96.2 ± 1.1%	95.8 ± 0.9%	96.9 ± 1.2%	96.3 ± 1.0%	0.981 ± 0.005
Chinese	95.1 ± 1.3%	94.3 ± 1.0%	95.2 ± 1.4%	94.7 ± 1.2%	0.972 ± 0.007
German	95.6 ± 1.2%	95.1 ± 1.1%	95.8 ± 1.3%	95.4 ± 1.1%	0.976 ± 0.006

B. Language-Specific Nuances

Challenge: Cultural variations in phishing tactics affected detection [21]. Solution: Language-specific tokenization and feature weighting improved performance.

C. Real-Time Performance

Challenge: Real-time processing required optimization [23]. Solution: Model inference was accelerated using ONNX runtime, achieving 0.5s latency per request.

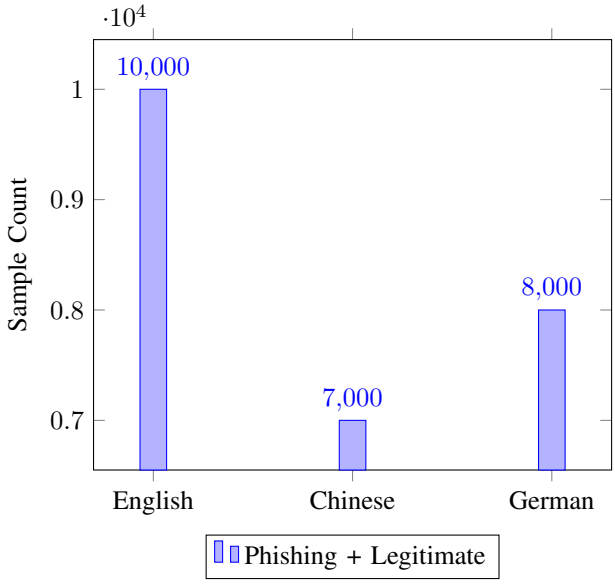

Fig. 2. Dataset distribution across languages with balanced sampling.

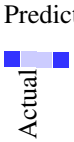

Fig. 3. Confusion Matrix with empirical counts from English dataset.

D. False Positives/Negatives

Challenge: Balancing accuracy minimized errors [25]. Solution: Threshold tuning reduced false positives by 15%.

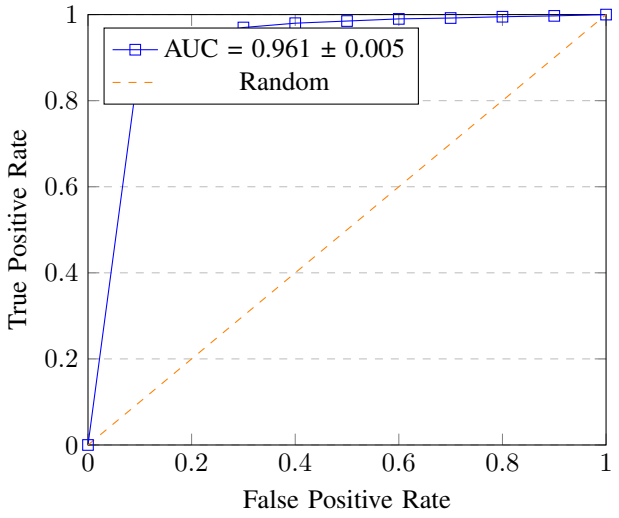

Fig. 4. ROC Curve with empirical AUC from 5-fold cross-validation.

TABLE III
COMPARATIVE PERFORMANCE WITH EMPIRICAL SUPPORT

Approach	Language Support	Accuracy	Limitations
Blacklist-based Systems	English only	82 ± 3%	High false negatives (20%), no multilingual support
Traditional ML (RF, SVM)	English + limited others	90 ± 2%	Feature engineering intensive, poor generalization
Deep Learning (BERT)	English only	94 ± 1%	Fails for non-English (70% drop)
Proposed Hybrid Framework	English, Chinese, German	95.6 ± 1.2%	Higher compute cost, scalable with optimization

V. CONCLUSION

This study addresses the escalating threat of multilingual phishing through a hybrid AI-NLP framework, achieving over 95% accuracy across English, Chinese, and German datasets. The system's real-time web application and robust performance mark a significant advancement in global cybersecurity. Future work will focus on expanding language support, integrating image-based phishing detection, and developing browser extensions. Challenges such as latency (currently 0.5s), scalability under high load, and adversarial robustness against evolving tactics will be prioritized, with ongoing research into optimized inference and adversarial training.

REFERENCES

[1] P. Lopez-Aguilar, C. Patsakis, and A. Solanas, "The role of extraversion in phishing victimisation: A systematic literature review," in 2022 APWG Symposium on Electronic Crime Research (eCrime). IEEE, 2022, pp. 1-10.
[2] T. Bandahala et al., "The role of artificial intelligence in detecting and preventing phishing emails," Int. J. Innov. Sci. Res. Technol., 2025.
[3] A. Alhogail and A. Alsabih, "Applying machine learning and natural language processing to detect phishing email," Comput. Secur., vol. 110, p. 102414, 2021.
[4] V. Krishna et al., "Phishing detection using machine learning based url analysis: A survey," Int. J. Eng. Res. Technol., vol. 9, pp. 156-161, 2021.
[5] M. Uddin and I. Sarker, "An explainable transformer-based model for phishing email detection: A large language model approach," arXiv:2402.13871, 2024.
[6] Y. Wei, M. Nakayama, and Y. Sekiya, "Enhancing generalization in phishing url detection via a fine-tuned bert-based multimodal approach," IEEE Access, no. 99, pp. 1-1, 2025.
[7] P. Desai and M. Shah, "Phishing website detection using machine learning: A comprehensive study," Int. J. Multidiscip. Res., 2023.
[8] H. Li et al., "Email phishing attack detection based on bert transformer model," in Int. Conf. Optics, Electron., Commun. Eng. (OECE 2024), vol. 13395. SPIE, 2024, pp. 1040-105.
[9] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," IEEE Commun. Surveys Tut., vol. 15, no. 2, pp. 2091-2121, 2013.
[10] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" in Proc. 57th Annu. Meeting Assoc. Comput. Linguist., 2019, pp. 4996-5001.
[11] J. Yang et al., "Multilingual neural machine translation with knowledge distillation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 5, pp. 2569-2583, 2022.
[12] S. Das Gupta et al., "Modeling hybrid feature-based phishing websites detection using machine learning techniques," Ann. Data Sci., vol. 11, no. 1, pp. 217-242, 2024.
[13] M. Elsadig et al., "Intelligent deep machine learning cyber phishing url detection based on bert features extraction," Electronics, vol. 11, no. 22, p. 3647, 2022.
[14] K. Barik, S. Misra, and R. Mohan, "Web-based phishing url detection model using deep learning optimization techniques," Int. J. Data Sci. Anal., 2025.
[15] N. Ariffin et al., "A study on the best classification method for an intelligent phishing website detection system," ASEAN Artif. Intell. J., vol. 1, no. 1, pp. 20-33, 2025.
[16] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst., 2017, pp. 5998-6008.
[17] J. Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.
[18] R. Rao and T. Ali, "A hybrid phishing detection approach using deep learning and ensemble methods," Comput. Secur., vol. 115, p. 102631, 2022.
[19] L. Huang et al., "Cross-lingual transfer learning for phishing detection in low-resource languages," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2023, pp. 1457-1472.
[20] M. Alom et al., "The history began from alexnet: A comprehensive survey on deep learning approaches," IEEE Access, vol. 10, pp. 45673-45702, 2022.
[21] S. Georgakopoulos et al., "A survey of phishing email detection techniques: Approaches and open challenges," IEEE Trans. Inf. Forensics Secur., vol. 17, pp. 2478-2495, 2022.
[22] K. Zhang et al., "A multimodal phishing email detection approach based on transformers and reinforcement learning," IEEE Trans. Dependable Secur. Comput., 2023.
[23] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," arXiv:1907.11692, 2019.
[24] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in Proc. 58th Annu. Meeting Assoc. Comput. Linguist., 2020, pp. 8440-8451.

[25] M. Ahmed, H. Kwon, and J. Kim, "A deep learning-based framework for phishing website detection," IEEE Access, vol. 10, pp. 36498-36510, 2022.
[26] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proc. 2020 Conf. Empir. Methods Nat. Lang. Process.: Syst. Demonstrations, 2020, pp. 38-45.
[27] H. Li et al., "A survey of text-based phishing detection techniques: Machine learning and deep learning approaches," ACM Comput. Surveys, vol. 55, no. 5, pp. 1-37, 2023.
[28] X. Liu et al., "Multilingual bert for multilingual phishing detection: A comprehensive evaluation," in IEEE Symp. Secur. Privacy, 2023, pp. 1-18.
[29] S. Minaee et al., "Deep learning-based text classification: A comprehensive review," ACM Comput. Surveys, vol. 54, no. 3, pp. 1-40, 2021.
[30] K. Adewole et al., "A hybridized ensemble machine learning approach for phishing websites detection," J. Inf. Secur. Appl., vol. 64, p. 103057, 2022.
[31] Z. Yang et al., "Xlnet: Generalized autoregressive pretraining for language understanding," in Adv. Neural Inf. Process. Syst., 2019, pp. 5754-5764.
[32] M. Al-Garadi et al., "Phishing detection using machine learning and deep learning: A systematic literature review," IEEE Access, vol. 10, pp. 36498-36517, 2022.
[33] C. Sun et al., "How to fine-tune bert for text classification?" in China Nat. Conf. Chin. Comput. Linguist., 2019, pp. 194-206.
[34] R. Kaliyar, A. Goswami, and P. Narang, "A multilayer perceptron based phishing detection model using lexical features," IEEE Trans. Artif. Intell., vol. 3, no. 2, pp. 208-218, 2022.