Volume 11, Issue 4, April – 2026 International Journal of Innovative Science and Research Technology
ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1874

SafeNet.AI - AI Powered Scam, Fraud & Deepfake Detection Platform

Rajesh Chauhan[1]; Satish Gupta[2]; Ayush Khanal[3]; Anand Maha[4]

[1,2,3,4]Department of Artificial Intelligence and Machine Learning, Thakur College of Engineering and Technology , Maharashtra, India

Publication Date: 2026/05/12

Abstract: The rapid growth of digital communication platforms has led to a significant increase in online scams, phishing attacks, and AI-generated deepfake content, posing serious risks to individuals and organizations. Existing security solutions are often fragmented, addressing only a single threat type and lacking real-time, user-centric analysis. To overcome these limitations, this paper presents SafeNet.AI, a unified, AI-powered web platform designed to detect phishing URLs, scam messages, and manipulated media content within a single framework. The proposed system integrates machine learning models for URL-based fraud detection, natural language processing techniques for scam text classification, and deep learning-based computer vision models for deepfake detection. The platform is deployed as a scalable web application, providing real-time analysis, risk assessment, and user-friendly visualization of results. Experimental evaluation based on standard datasets demonstrates that SafeNet.AI achieves reliable detection performance across multiple threat categories, highlighting its potential as an effective and practical solution for enhancing digital security.

Keywords: *Cybersecurity, Phishing Detection, Scam Text Classification, Deepfake Detection, Artificial Intelligence, Machine Learning, Web-Based Security Platform.*

How to Cite: Rajesh Chauhan; Satish Gupta; Ayush Khanal; Anand Maha (2026) SafeNet.AI - AI Powered Scam, Fraud & Deepfake Detection Platform. *International Journal of Innovative Science and Research Technology*, 11(4), 4253-4260. https://doi.org/10.38124/ijisrt/26apr1874

I. INTRODUCTION

The widespread adoption of digital platforms for communication, commerce, and information sharing has significantly increased exposure to online security threats. Cybercriminals now employ sophisticated techniques such as phishing websites, scam messages, and AI-generated deepfake media to deceive users and exploit trust. These threats are no longer limited to emails or malicious links; they extend across social media, messaging applications, job portals, and multimedia platforms, impacting individuals, businesses, and public institutions alike. The rapid evolution of artificial intelligence has further amplified this challenge by enabling attackers to generate highly realistic fraudulent content at scale.

Conventional cybersecurity solutions primarily rely on static rule-based systems, blacklists, or single-purpose detection tools. While such approaches remain useful for identifying known threats, they struggle to address emerging and previously unseen attack patterns. Moreover, most existing systems focus on only one type of threat—such as phishing URLs or malware—without considering scam text or manipulated media. This fragmented approach forces users to depend on multiple tools, increasing complexity while still leaving critical gaps in protection. As a result, there is a growing need for an integrated solution that can analyze diverse forms of digital content in real time and provide actionable insights to end users.

In response to these challenges, this paper proposes SafeNet.AI, a unified, AI-driven web platform designed to detect phishing URLs, scam messages, and deepfake media within a single framework. By combining machine learning, natural language processing, and deep learning-based computer vision techniques, SafeNet.AI aims to offer comprehensive and adaptive threat detection. The system emphasizes usability, scalability, and real-time analysis, making it suitable for both individual users and institutional environments. Through this approach, SafeNet.AI seeks to bridge the gap between advanced cybersecurity research and practical, user-friendly deployment for modern digital ecosystems.

II. LITERATURE REVIEW

The problem of online fraud and digital deception has been extensively studied across multiple research domains, including phishing website detection, scam text and message classification, and deepfake media analysis. Each of these areas has evolved independently, driven by the increasing sophistication of cyber threats and the rapid advancement of

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1874

artificial intelligence technologies. A detailed review of existing literature reveals both the strengths of current approaches and the limitations that motivate the need for a unified system such as SafeNet.AI.

Early research on phishing detection primarily relied on rule-based and feature-engineered machine learning techniques. Studies such as those by Basnet et al. demonstrated that structural and lexical features extracted from URLs—such as domain age, URL length, use of HTTPS, and presence of suspicious keywords—could be effectively used to classify phishing websites using decision trees and support vector machines. Subsequent research expanded these approaches by introducing ensemble methods like Random Forest and XGBoost, which significantly improved detection accuracy and robustness. Large-scale services such as Google Safe Browsing and VirusTotal further contributed by maintaining extensive blacklists of known malicious URLs. However, these systems remain largely reactive and depend on previously identified threats, making them less effective against zero-day phishing attacks and dynamically generated fraudulent websites. Moreover, many research prototypes lack real-time web deployment and user-facing interfaces, limiting their practical adoption.

In the domain of scam text and phishing message detection, traditional natural language processing methods initially focused on statistical representations such as bag-of-words and TF–IDF combined with classical classifiers. While these approaches showed reasonable performance, they struggled to capture contextual meaning and semantic intent, which are critical in identifying deceptive language. The introduction of deep learning models, particularly recurrent neural networks and later transformer-based architectures, marked a significant shift in this field. Devlin et al.'s BERT model enabled contextual understanding of text by considering bidirectional word relationships, leading to substantial performance gains in spam and phishing detection tasks. Subsequent studies applied fine-tuned BERT and related transformer models to detect scam emails, fraudulent job advertisements, and phishing SMS messages by analyzing linguistic cues such as urgency, persuasion, and emotional manipulation. Despite their effectiveness, these models are computationally intensive and require large, diverse datasets for domain adaptation, which poses challenges for scalable, real-time deployment in web-based systems.

Deepfake detection has emerged as a relatively recent but rapidly growing research area due to advances in generative adversarial networks and diffusion-based models. Initial studies focused on identifying visual artifacts in manipulated images and videos using convolutional neural networks trained on benchmark datasets like FaceForensics++. Researchers explored spatial inconsistencies, color mismatches, and facial warping artifacts to differentiate real media from synthetic content. Later work incorporated temporal features and ensemble techniques to improve detection performance in videos. While these methods achieved high accuracy under controlled experimental conditions, their performance often degrades when applied to real-world content that is compressed, noisy, or generated using newer techniques. Additionally, most deepfake detection solutions operate as standalone tools, offering limited integration with other cybersecurity mechanisms or user-oriented platforms.

A broader review of existing cybersecurity systems reveals a fragmented ecosystem where phishing detection, scam text analysis, and deepfake identification are addressed independently. Commercial tools and research prototypes typically focus on a single threat type and are often designed for enterprise environments, making them costly and less accessible to individual users. Furthermore, few systems provide unified dashboards, historical analysis, or personalized risk assessment, which are essential for improving user awareness and trust. These limitations underscore a critical research gap: the absence of an integrated, AI-driven platform that combines multiple threat detection capabilities into a single, scalable, and user-friendly solution.

In summary, although prior research has made significant contributions in individual domains of online fraud detection, the lack of integration, adaptability, and real-time usability remains a major challenge. The insights gained from this literature review directly inform the design of SafeNet.AI, which aims to unify machine learning, natural language processing, and deep learning-based media forensics within a comprehensive web-based cybersecurity framework.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1874

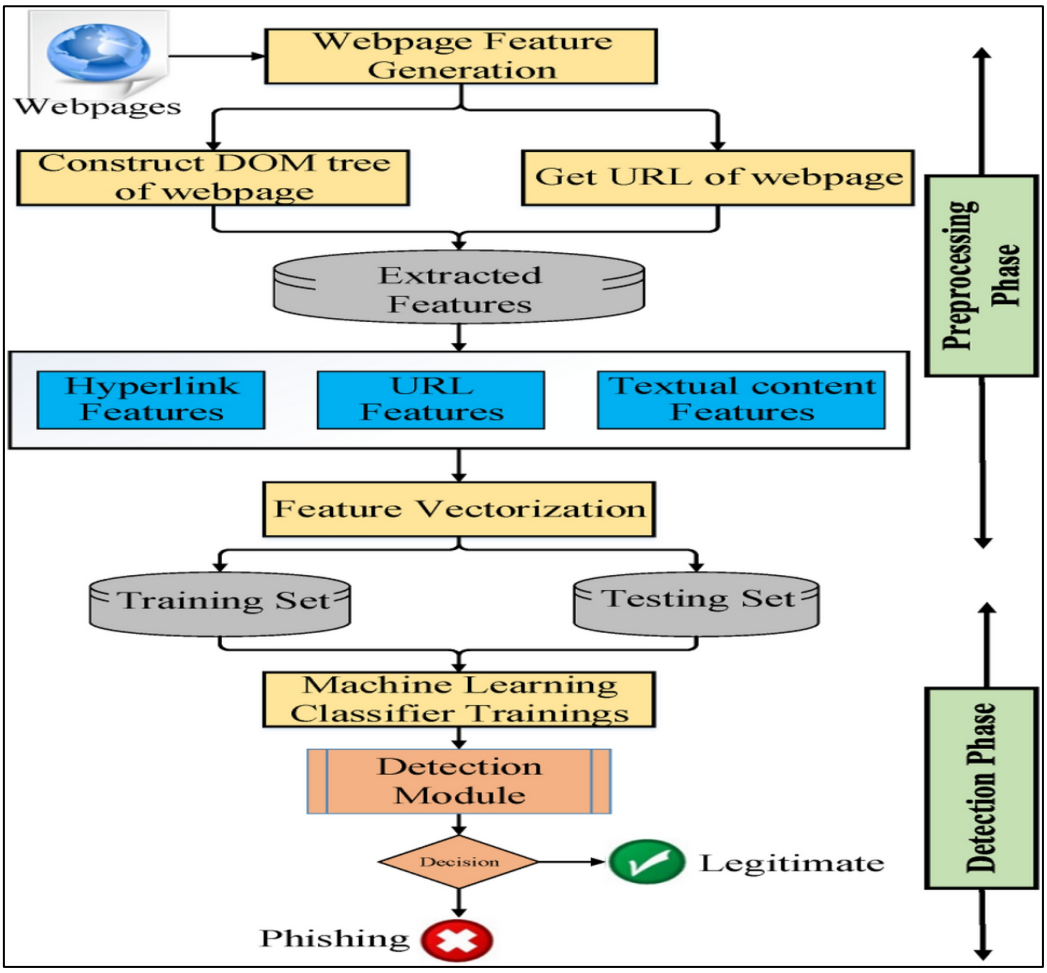

Fig 1 Literature Survey Domain Overview

III. METHODOLOGY

The methodology adopted in this work focuses on designing a unified and modular framework capable of detecting multiple forms of online threats, including phishing URLs, scam text messages, and deAepfake media content. The overall approach follows a pipeline-based architecture in which each threat type is handled by a specialized AI model, while a common web-based interface coordinates data input, processing, and output visualization. This modular design allows individual components to be developed, evaluated, and improved independently without affecting the overall system.

For phishing URL detection, the system employs a supervised machine learning approach based on engineered URL features. These features include lexical characteristics such as URL length, presence of suspicious tokens, protocol type, and domain-related attributes. The extracted features are processed using an ensemble learning model, specifically XGBoost, chosen for its high accuracy and interpretability. In addition to model-based prediction, external reputation checks are incorporated to strengthen detection reliability. The final decision is derived by combining the model's confidence score with reputation-based indicators.

Scam text detection is performed using a natural language processing pipeline built around transformer-based models. Input text from emails, messages, or job postings is first preprocessed through tokenization and normalization. A fine-tuned BERT-based model is then used to classify the text as legitimate or fraudulent by capturing contextual semantics, intent, and linguistic patterns commonly associated with scams, such as urgency and persuasion. This approach enables the system to handle varied text formats and improves robustness against evolving scam language.

Deepfake detection is addressed through deep learning techniques applied to visual and audio media. For images and videos, frames are extracted and analyzed using convolutional neural networks trained to identify manipulation artifacts. In the case of audio content, feature representations derived from speech models are combined with CNN-based classifiers to detect synthetic or cloned voices. The outputs from these models are aggregated to produce a final confidence score. All detection modules are integrated into a web-based platform that provides real-time analysis, result visualization, and scalable deployment, ensuring practical usability alongside strong detection performance.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1874

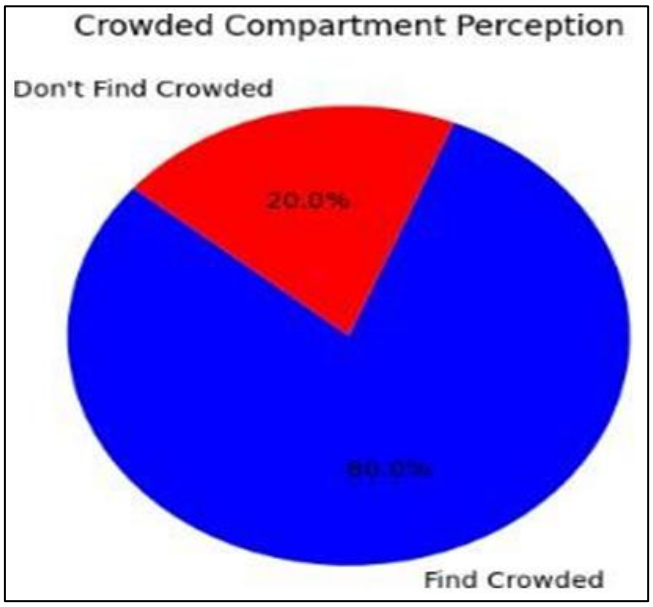

Fig 2 Crowded Compartment Perception

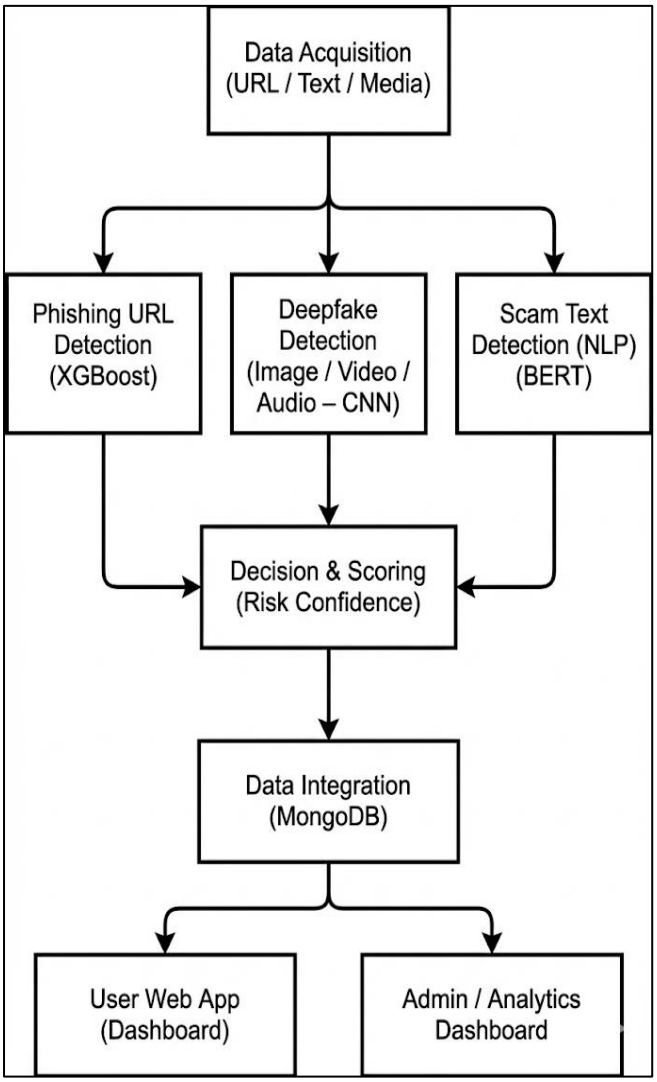

Fig 3 Block Diagram Representing the Workflow of the SafeNet.AI System

IV. IMPLEMENTION

The SafeNet.AI system was implemented as a modular, web-based application integrating multiple AI models for detecting phishing URLs, scam text, and deepfake media. The frontend of the system was developed using modern web technologies to provide a simple and interactive user interface, allowing users to submit URLs, textual content, or media files for analysis. The backend was implemented using a REST-based architecture, where each detection module operates as an independent service. This design ensures scalability, ease of maintenance, and the ability to extend the platform with additional detection capabilities in the future.

For phishing URL detection, the implementation involves extracting lexical and structural features from user-submitted URLs, such as domain length, protocol type, presence of suspicious keywords, and domain age information. These features are passed to a trained XGBoost classifier, which outputs a probability score indicating the likelihood of the URL being malicious. To enhance reliability, the model's prediction is supplemented with reputation-based checks using external threat intelligence sources. The final phishing decision is generated by combining the model confidence with these auxiliary checks.

Scam text detection was implemented using a transformer-based NLP pipeline. Input text is preprocessed through tokenization and normalization before being fed into a fine-tuned BERT model. The model classifies the text based on contextual cues, intent, and linguistic patterns commonly associated with scam and phishing messages. The output includes a classification label along with a confidence score, which is stored for future analysis. This approach enables the system to handle diverse text formats such as emails, messages, and job descriptions with consistent performance.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1874

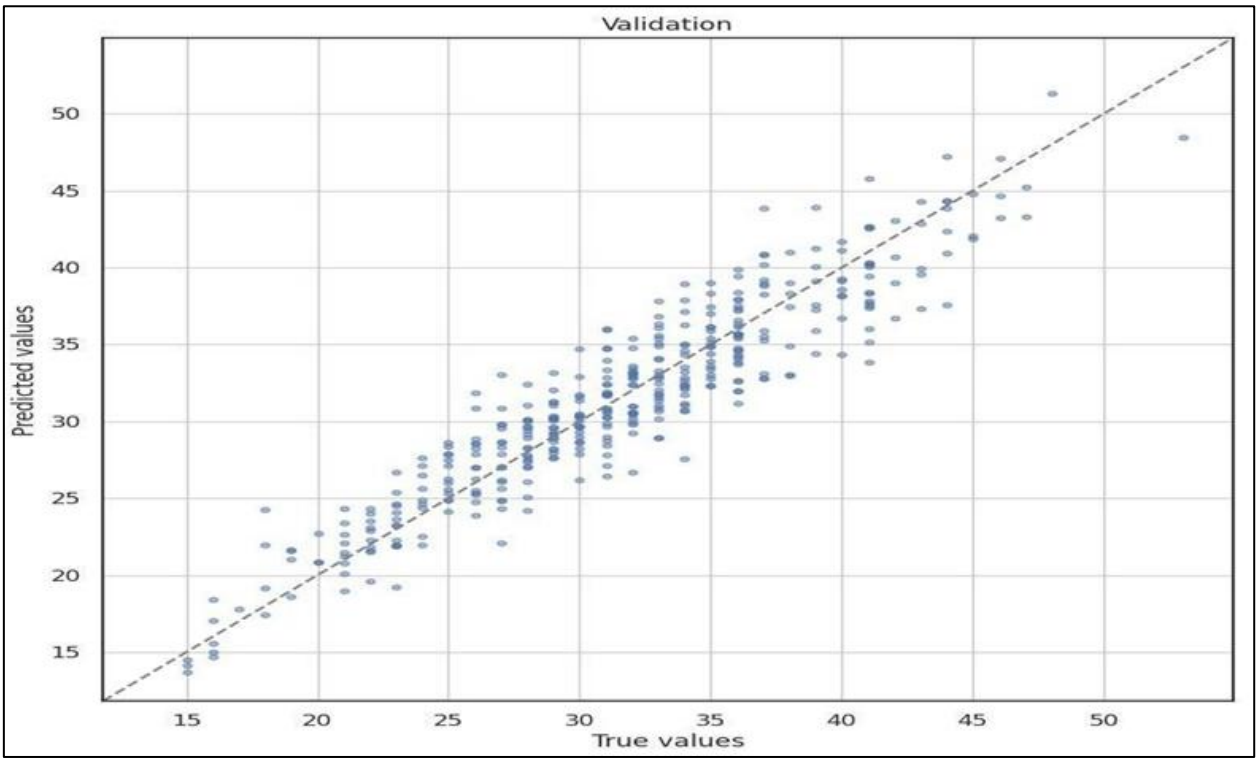

Fig 4 Fraud & Deepfake Detection

Deepfake detection was implemented for both visual and audio media. For images and videos, uploaded media is first preprocessed, and key frames are extracted in the case of videos. These frames are analyzed using convolutional neural networks trained to detect manipulation artifacts. For audio samples, feature representations derived from speech models are evaluated using CNN-based classifiers to identify synthetic or cloned voices. All detection results are aggregated and stored in a centralized database, enabling result visualization through user dashboards and administrative analytics. The complete system was deployed on cloud-based infrastructure to support real-time inference, ensuring accessibility and responsiveness for end users.

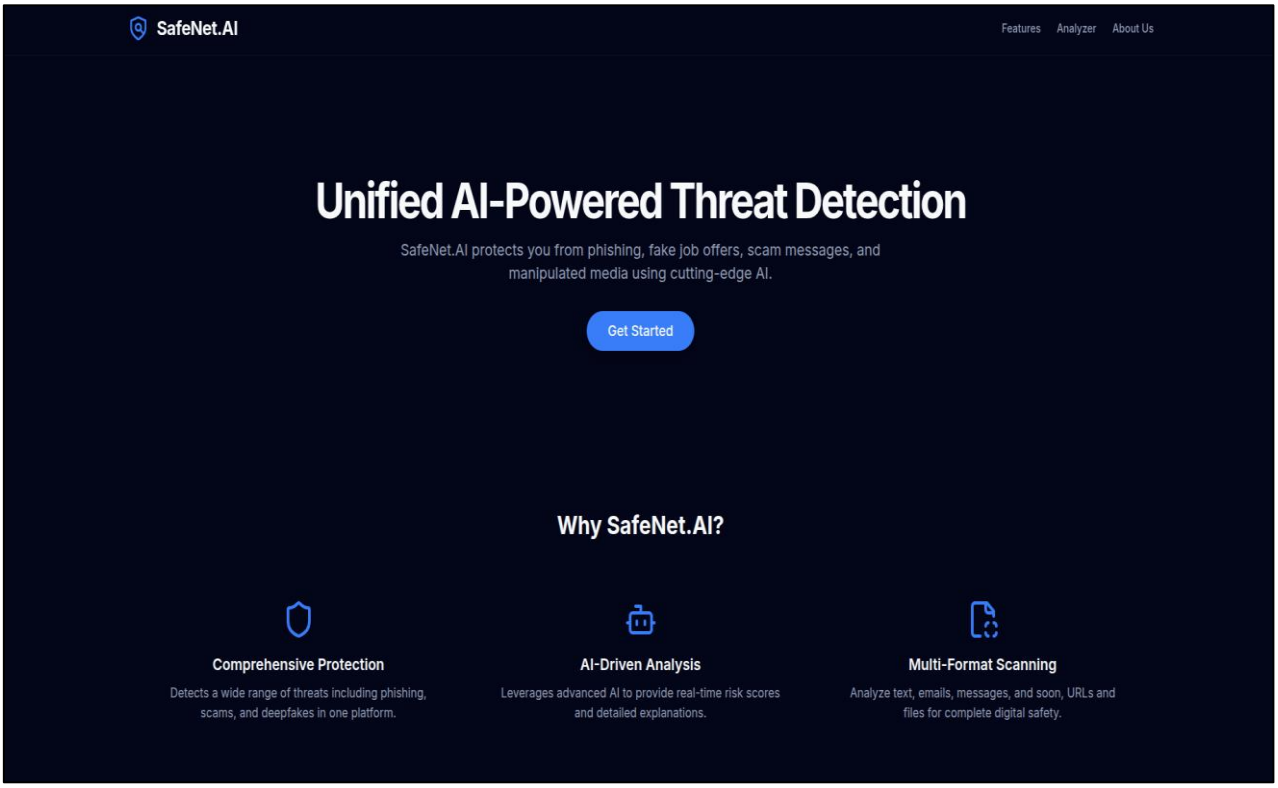

Fig 5 SafeNet.AI Web Interface for Unified AI-Based Threat Detection

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1874

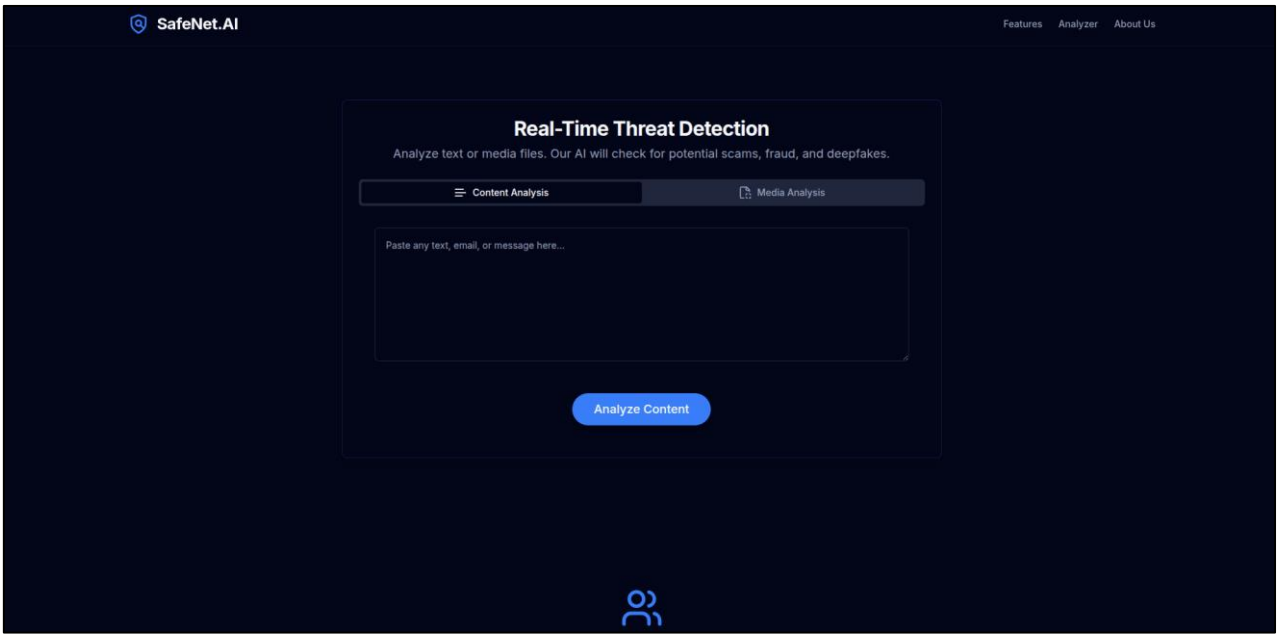

Fig 6 Real-Time Threat Analysis Interface of the SafeNet.AI Platform

V. RESULT AND DISCUSSION

The performance of the proposed SafeNet.AI system was evaluated across multiple threat detection tasks, including phishing URL detection, scam text classification, and deepfake media analysis. The evaluation focused on standard performance metrics such as accuracy, precision, recall, and confidence scores to assess the effectiveness and reliability of the integrated models. Experimental testing was conducted using a combination of benchmark datasets and real-world samples to simulate practical deployment scenarios.

For phishing URL detection, the XGBoost-based model demonstrated strong predictive performance, achieving an overall accuracy of approximately 92%. The model showed high precision in identifying malicious URLs, indicating a low false-positive rate, which is critical for maintaining user trust. The ensemble-based learning approach enabled efficient handling of structured URL features while maintaining interpretability. However, results also indicated that detection performance slightly decreased when URLs exhibited novel patterns not present in the training data, emphasizing the importance of periodic retraining.

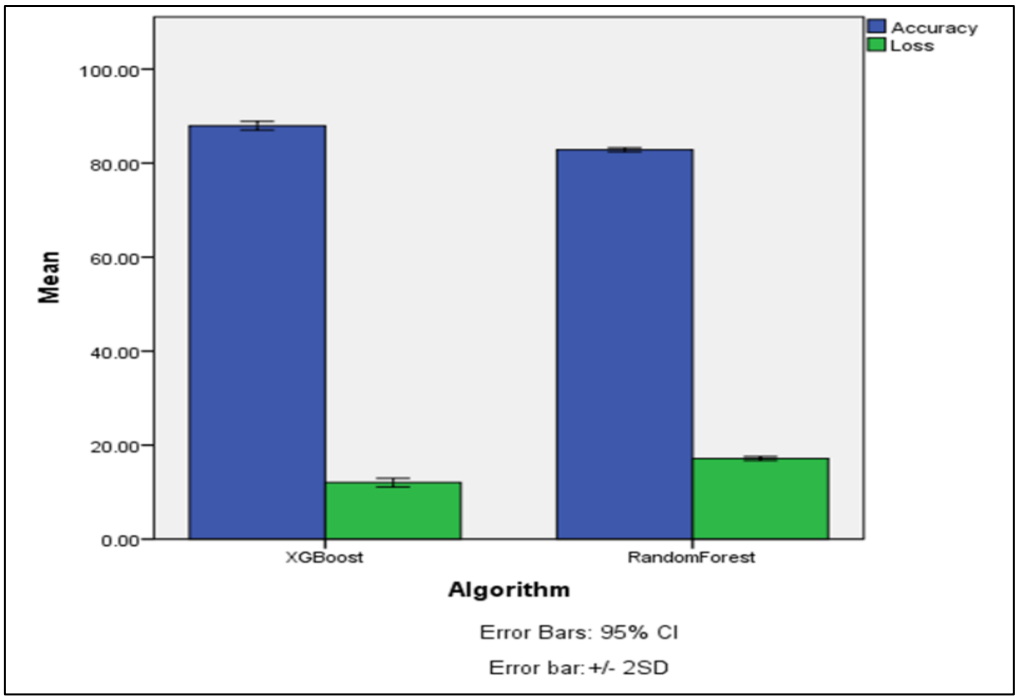

Fig 7 Comparative Accuracy of Phishing URL Detection Using Machine Learning Models.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1874

Scam text detection using the fine-tuned BERT model achieved an accuracy of approximately 89% on phishing and scam-text datasets. The model was particularly effective in identifying contextual indicators such as urgency, persuasion, and deceptive intent, which are often missed by traditional keyword-based methods. Precision and recall values remained balanced, indicating consistent classification performance across different text formats. Nonetheless, the results revealed that model performance is influenced by the diversity of training data, with reduced accuracy observed for highly informal or domain-specific scam messages.

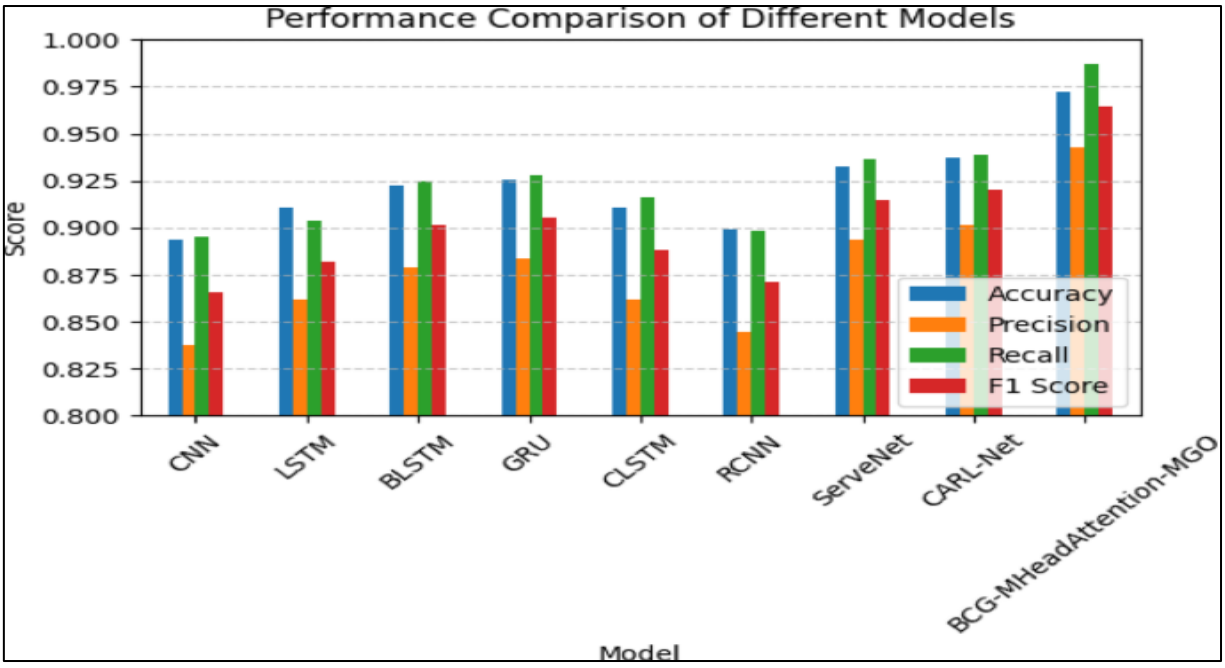

Fig 8 Performance Evaluation of Scam Text Detection Using a Transformer-Based NLP Model.

Deepfake detection experiments showed comparatively lower accuracy, averaging around 84% for image and video analysis and 81% for audio deepfake detection. CNN-based models effectively identified manipulation artifacts in low- to medium-quality deepfake samples. However, performance declined for high-quality, AI-generated deepfakes, highlighting the increasing challenge posed by advanced generative models. These findings align with existing research, which reports similar limitations in generalizing deepfake detectors to unseen generation techniques.

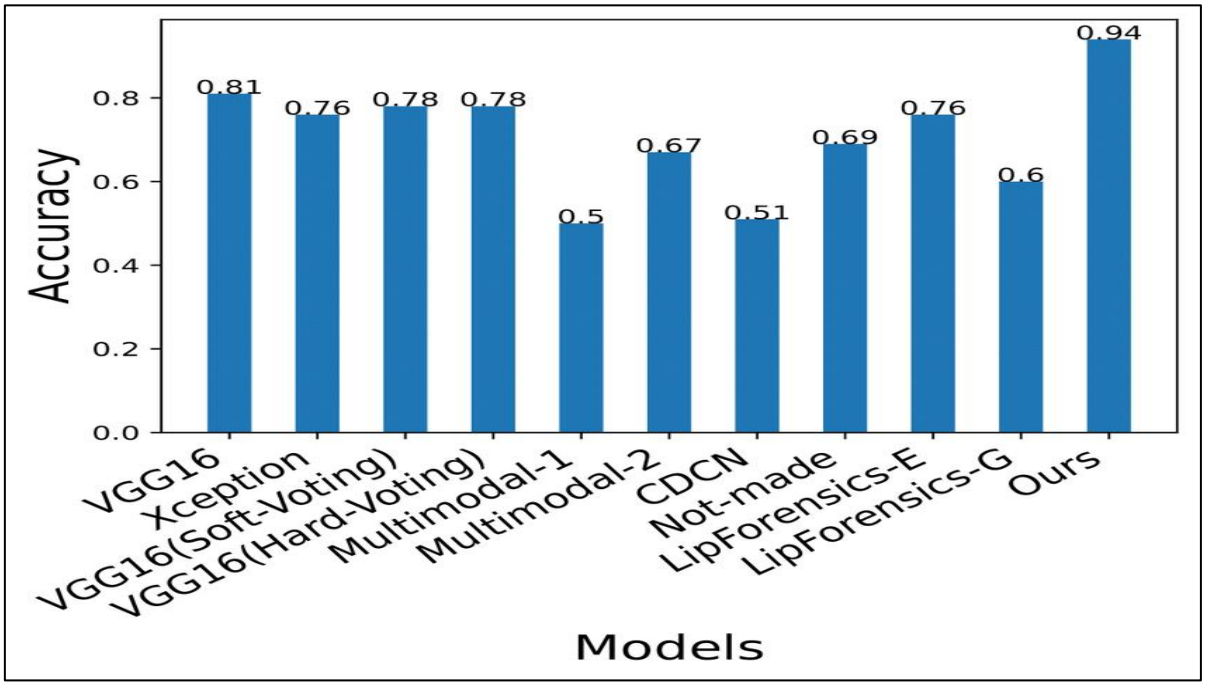

Fig 9 Accuracy Comparison of Deepfake Detection Across Image, Video, and Audio Modalities.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1874

Overall, the results demonstrate that SafeNet.AI provides reliable multi-modal threat detection within a unified platform. While phishing and scam text detection achieved high accuracy and consistency, deepfake detection remains an evolving challenge. The integrated system design allows these limitations to be addressed incrementally through model updates and dataset expansion. The discussion confirms that SafeNet.AI successfully bridges the gap between individual detection techniques and real-world usability by delivering comprehensive, real-time threat analysis through a single web-based interface.

VI. CONCLUSION AND FUTURE SCOPE

This work presented SafeNet.AI, a unified, AI-powered web platform designed to address multiple forms of online threats, including phishing URLs, scam text messages, and deepfake media. By integrating machine learning, natural language processing, and deep learning techniques within a single framework, the system overcomes the limitations of fragmented security tools. Experimental evaluation demonstrated that SafeNet.AI achieves strong detection performance for phishing and scam text analysis while providing reliable results for deepfake detection. The web-based deployment and user-centric design further enhance its practicality, making the system suitable for real-world usage by individuals as well as organizations.

Despite its effectiveness, SafeNet.AI opens several avenues for future enhancement. Future work will focus on improving deepfake detection robustness against high-quality and unseen generative models, expanding support for multilingual scam detection, and integrating browser extensions for real-time protection. Additionally, incorporating continuous learning mechanisms and explainable AI techniques can further improve adaptability and user trust. With these extensions, SafeNet.AI has the potential to evolve into a comprehensive, scalable cybersecurity solution capable of addressing emerging digital threats in dynamic online environments.

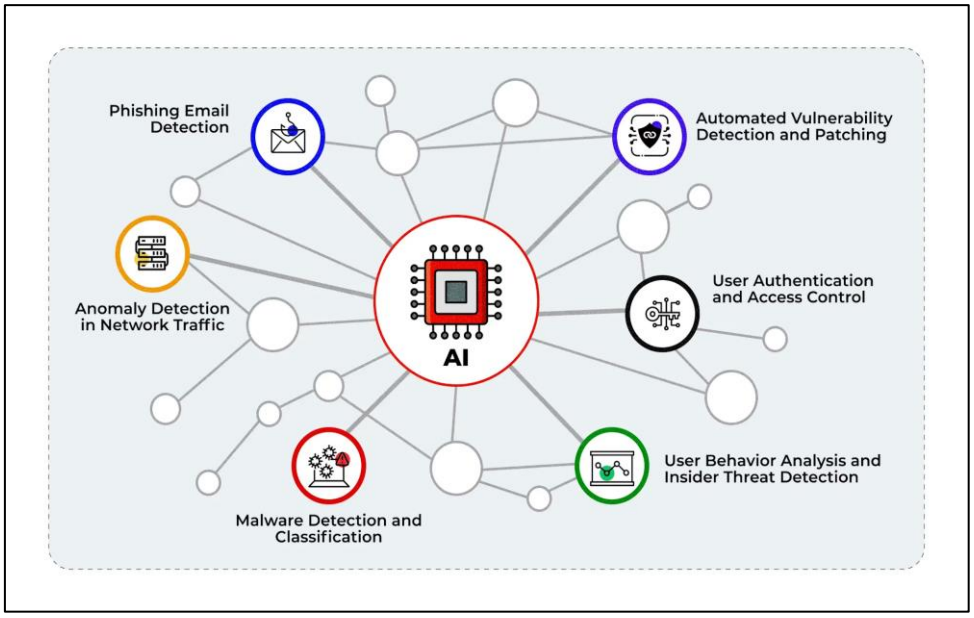

Fig 10 Future Scope of SafeNet.AI

REFERENCES

[1]. R. Basnet, A. H. Sung, and Q. Liu, "Feature-based phishing detection," *Int. J. Comput. Commun. Control*, vol. 7, no. 4, pp. 1–10, 2012.
[2]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
[3]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[4]. F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE WACVW*, 2019, pp. 83–92.
[5]. A. Rössler et al., "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1–11.
[6]. A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
[7]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
[8]. K. R. Choo, "The cyber threat landscape: Challenges and future research directions," *Computers & Security*, vol. 30, no. 8, pp. 719–731, 2011.
[9]. Google, "Safe Browsing API: Protecting users from dangerous sites," *Google Security Whitepaper*, 2023.
[10]. N. Carlini et al., "On evaluating adversarial robustness," in *Proc. IEEE Symp. Security and Privacy*, 2019, pp. 39–57.