Scam Shield- A Hybrid ML–NLP–XAI Framework for Real-Time Phishing Detection and Explainable Cybersecurity Awareness

Reshita Chaudhary[1], Snigdha Som[2], Rishika Yadav[3], Parth Goel[4]
[1,2,3,4]*Department of Computer Science & Engineering, Meerut Institute of Engineering and Technology, Meerut, India*

***Abstract*-Phishing and online scams are nowadays becoming the face of a cybersecurity threat which is breaking people's trust as well as they are now tending to steal the personal and sensitive data of any particular person or a client of any company. Despite of being progressing in the area of Machine Learning, deep learning and Natural Language Processing, still in today's world current solutions do not have the ability to adapt, to explain and to user centered design.**

This paper offers Scam Shield- which is a mixture of Machine Learning (ML), Natural Language Processing (NLP), Explainable Artificial Intelligence (XAI). It works on the browser and also as a web platform which is made up by integrating the mentioned three technologies. It mostly works for real time scan and for any type of phishing detection. This whole made up system analyses all the given Uniform Resource Locator, any content on the web page, any changes in behavior, any website which is classified to be put into a category safe, suspicious and malicious. By using Explainable Artificial Intelligence, Scam Shield provides early alerts which will work as an alarm and then explain why this particular website or Uniform Resource Locator is not safe to visit.

Results showed improve in accuracy, in a smaller number of false positives and it also has increased the whole user engagement if we compare it to before from now. Therefore, this model's feedback reloads continuously and update its dataset and ensure for the adaptation against emerging scams.

***Index Terms*-Cybersecurity, Phishing detection, Machine Learning, Natural Language Processing, Explainable Artificial Intelligence, Browser Extension, Adapting Conditions.**

I. INTRODUCTION

Online scams and phishing are still remaining among the most difficult cybersecurity danger threats in the modern society. Online attacker uses fake websites, fake login portals, and they also use social engineering to extract all the user credentials and their financial data. If we go by the statistics according to the anti-phishing working group phishing attacks reached all time maximum in the year 2023 and all the websites detected with approximately 4.7 million unique phishing cases with their sites. This figure is rising continuously and spreading wide with the addition and adoption of the generative Artificial Intelligence by malicious actors.

The results of any successful phishing attack is beyond any financial losses of any particular being. They are literally compromising the security systems and their organization and breaking trust of the users in the digital services which they are using or they are about to use in the future. Despite being all of this, many users are still falling for these schemes and these online scams because of the cleanliness in their attack which includes everything from start to the end which looks so sophisticated from start to end, so the user cannot determine what is real and what is fake, they tend to believe that this is the real and fall for the scam. These attacks can be of many types including homographic attacks IDN spoofing or ai generated phishing pages. These pages all are visually true and is near to impossible for distinguish them from the real ones.

Although many people among researchers have created a very advanced version of the modules which can detect by using the technology of Machine Learning and deep learning, but many of them lack simple understanding of any context and their ability to communicate with user back and forth. The systems which are working right now either flags a Uniform

Resource Locator as malicious without giving any explanation and also do not require any technical help to interpret their outputs by leaving many average users empty handed to understand the threat themselves only.

Scam Shield has an aim to address these problems through real time intelligent detection system which will not only identify these games but also will educate the users about the online risks. When we combine Machine Learning for its classification with Natural Language Processing for its most efficient semantic analysis and Explainable Artificial Intelligence for its good transparency, then Scam Shield becomes a bridge and reduce the gap between the theory of any algorithm precision and the understandability of the human for its security and for its guidance.

The key contributions of this paper are (1) A hybrid Machine Learning-Natural Language Processing Detection model for multi signaling phishing classification, (2) Integration of both SHAPE and LIME based explainability for user facing the alerts according to their contexts, (3) A feedback model which is continuously looping and enabling for adapting any changes against the threats which are emerging day by day and at last (4) A browser extension and a web platform interface that is operating the whole system and all the capabilities for the users which are not very technical.

II. LITERATURE REVIEW

The whole landscape of phishing detection has been evolving day by day at a healthy amount from the past decade, it has seen a lot of transition from a simple blacklist-based method towards any proper hybrid frameworks. This whole section of the paper reviews foundational and recent works that has informed the design of the Scam Shield.

(a) Uniform Resource Locator based and browser level detection

S. Arvind and R.K. Sharma's PHISTOR [1] proposed a browser-based Machine Learning model for Uniform Resource Locator analysis using lexical and host-based features. It is very effective for the patterns which are known but it misses the semantic evaluation of any content on the web page which is making it vulnerable to the phishing pages which can easily show themselves is the real Uniform Resource Locators or the legitimate ones by mimicking themselves by just changing in their structures. Similarly, the random forest classifier approach [2] which is integrated as browser extension and has achieved a strong classification accuracy reported at approximately 96.5% but it is only limited to the detection for static features and it is not adapting to a new pattern of attack without training manually.

(b) Natural Language Processing and Semantic analysis approaches

There were many more advancements which were seen in the Phishing site detection using Machine Learning and Natural Language Processing by Shinde and Shinde which absorbed Natural Language Processing to analyze the pages and their features about the texts which included keyword frequency, sentiment and urgency indicators. Whereas it shows a huge step towards semantic understanding, But the system misses real time functionality and was evaluated only on the datasets which were static at that time. AI Enhanced Phishing Defense [4] by Patel and this Desai improved rates of detection through various methods but failed to provide any reason or any explanation to the user when it's come for the education purpose leaving them clueless.

Natural Language Processing for Phishing detection [5] by Rao and Nair gripped transformer-based language model to capture the hints according to the context in the phishing emails and their web pages by demonstrating high precision. But this whole nature of transformer inference was very intense and made real time deployment very challenging and tough. Celery based real time detection [6] by John brings light to a performance which distributed many tasks which were already in a queue, which was followed by achieving low latency results, where is it was excluded By any form of interpretation or any explanation faced by the user.

c) Explainability and human factors

NLP in intrusion detection [7] by Sworna et al. which gives us a review of the system for identifying the adaptability of retraining when the key is missing is a capability in the systems which are mostly deployed. The most important Explain don't just Warn! [8] by Saha Roy, Torres and Nili Zadeh Showed user studies the contextual Phishing warning which are explainable are mostly significant, which usually performs much better than binary alerts in changing user behavior and reducing click rates on malicious links. This work directly motivates Scam Shield's XAI module.

Overall, these studies show a consistent pattern, whereas the detection algorithms have become much more powerful and are increasing at high rate, But the ability to explain to adapt and the design for the user still has some gaps to fill. No systems which are currently being used or active addresses all the three problems together. Table 1 summarizes the main capabilities and the limitations of related works.

III. PROBLEM STATEMENT

Nowadays Phishing and detection systems for scams are developed so far and beyond That they show either very strong accuracy of their algorithm or good at usage, but rarely at both. The real challenge is a three-way tension between detection performance, responsiveness during real time and the output which can be easily understood by the user. Nowadays most systems go for first way and second way at the cost of third way.

Approaches which exist nowadays such as PHISTOR [1] and the random forest browser extension [2] Which often relies pretty much on URL based detection signals. While efficient in computation these approaches are not very good against minor changes to URL structure or any registration of domain Which can easily surpass detection entirely. On the other hand, NLP based hybrid model [3-5] improve semantic understanding, but need very strong resources, which are very heavy to use, Therefore, they require a significant system to evaluate transformer embeddings at the inference time, making Realtime browser deployment impossible without being firstly optimized. In addition to that, not any of the surveyed systems can implement a feedback loop which is continuous and used for learning. Evolution of Phishing campaigns is now a days on a high peak generally, by attackers being rotating domains, by refreshing visual templates and by adapting Data collection and preprocessing-The system pullouts Uniform Resource Locators, raw Machine Learning, text content and metadata from the websites which we have visited. The data is cleaned through tokenization, stopwatch removal and generalization. Phish Tank, Open Phish and APWG crime datasets often gives us an assurance that the training will be a reflection of the current attack patterns. Uniform Resource Locators are further being separated into lexical tokens, WHOIS registration age, TLD risk scores and redirect chain depth.

(a) Feature extraction-

A multiple signal feature has been made for each analyzed site. In Uniform Resource Locator, they structure features, capture length, presence of the IP address and use of the Uniform Resource Locator shorteners. Behavioral features also record chains, which are being redirected by many new narratives of social engineering within hours of a campaign launch. A model trained only for static fade quickly and without much more operations or processes to adapt or retrain on newly confirmed samples their detection rates become very less within over a certain time period.

Finally, and most importantly, the systems which are active nowadays often feel the human factors test. A “dangerous or safe” flag does not arms users to recognize and to resist for future attacks. As showed by Sahar Roy et al. [8], explanation according to the context which means why any website is suspicious, which means this process is more effective after it is measured, at change in the behavior of a user than most simple warnings.

In summary, not a single framework presently is a combination of a real time, explainable and adaptive detection in a user-friendly form for deployment. Scam Shield is designed that way to reduce the gap by combining machine learning, NLP and explainable AI into one continuous learning detection and become a ecosystem also for raising awareness.

IV. PROPOSED SOLUTION

Scam Shield is designed as a modular cybersecurity framework within many layers in it. It operates both as a browser extension and stand-alone also as a web platform. Its architecture combines multiple technologies which are complementary, to achieve the goal of accuracy, speed and to interpret all at the same time. The six main components are as follows-

(a) Detection engine-

A hybrid Machine Learning, Natural Language Processing classification process that joined together the feature vector. A XG boost which handles all the structures, tabular features. Whereas a proper Distil BERT model processes the whole Natural Language Processing embedding vector. The outputs of the both are combined through a Meta learner which gives us a final three class dimension distribution, according to their probability- Safe, suspicious or malicious.

(b) Explainable AI (Explainable Artificial Intelligence) module-

After the classification, the Explainable Artificial Intelligence model gives us the explanation which are readable by any human, using SHAP for the tabular Machine Learning branch and LIME for the Natural Language Processing branch. SHAP identifies which Uniform Resource Locator or which behavioral features have contributed the most to the classification of any type of risks, Meanwhile LIME provides which phrases or words in the text have triggered the semantic alert. All the outputs are verified into a natural language contextual warning displayed to the user. For example, "This site uses urgency language {'Verify now or lose access'}. This kind of credential attacks or this domain was registered only two days ago."

(c) User awareness interface-

A browser extension and web dashboard, which display detection results in the visual formats. Green means "safe", amber means "suspicious" and red means "malicious". All the three indicators are being followed up by the Explainable Artificial Intelligence generated explanation. This interface includes a one click event reporting button and links to contextual cybersecurity education modules set to the detected type of threat. For example, ("How to spot a login Phishing page").

(d) Feedback Learning Loop-

In this particular module the user submits all the corrections it can either be a false positive or a false negative and they also confirm any threat intelligence from the outside field which are included into a continuous retraining system. All the new samples are properly verified against various multiple sources for being added into the training model. And also, any updates on the model are being deployed on a regular basis, giving us an assurance that Scam Shield remains effective against Phishing campaigns.

V. SYSTEM ARCHITECTURE

The Scam Shield system follows a separated three tier architecture which includes a client layer, a processing layer and a data layer. The separation ensures that each component will be independent on and scale on their own. It also enables the updates on the model continuously without bringing any change to the client.

(a) Client layer

In this layer the browser extension is being implemented using the various web extensions API, which runs smoothly with the chromium-based browsers and Firefox. It blocks the Uniform Resource Locators which are navigated and also page load events, then it extracts page DOM and the content written in the form of text through a content script and then forward Every feature vector to the processing layer through a REST API. The web platform provides some similar interface for the manual Uniform Resource Locator and the analysis of the text, which can be easily achievable without browser extension installation.

(b) Processing layer

The backend API server receives feature extraction payloads and guides them through the Machine Learning, Natural Language Processing system. Deduction is being given by a model server Torch Serve for the particular Distil BERT component. The Explainable Artificial Intelligence module runs SHAPE or LIME computations asynchronously after classification to avoid the impact of latency on the primary risk score response. There is a total end to end deduction latency which target always below 1.5 seconds for the primary classification.

(c) Data layer

The intelligence database of threat increases labelled samples from Phis Tank, Open Phish and report submitted by user. A feature store caches recently computed feature vectors for repeating Uniform Resource Locator queries again and again. The registration of the model finds the various version of model, evaluates the metric and deploys the time into various timestamps, this gives the option to rollback if a new version of the model downgrades in its production.

(d) Security and privacy

All the Uniform Resource Locator and page content data is being transferred to the backend is very encrypted through TLs 1.3. Any information which is being identified as personal and detected in the form field is being taken down before sending. Users also have been provided an option for privacy first local inference mode, in which where a compressed-on device model will handle all the classification without leaving any data on the browser.

VI. WORKFLOW MODEL

This Scam Shield workflow is a closed loop system that contains detections, explanation, user interaction and model upgradation. Each and every part supports the real time decision making while maintaining the adaptability of the system. The total complete workflow proceeds through the following steps given below-

(a) URL interception-

The browser extension stops each and every page navigation events and captures the full URL, title of the page and all the visible content in the text available.

(b) Feature computation-

This module of feature extraction computes the various signal feature vector which includes URL structural, behavioral and NLP semantic features and transfers it to the backend.

(c) Hybrid classification-

The ML, NLP engine classifies the site and come back with the risk label such as safe or suspicious or malicious with an associated confidence score.

(d) XAI explanation generation-

The XAI module calculates SHAPE or LIME features and combines a natural language explanation which will identify the risk factors that are contributing very much at the top.

(e) User alert delivery-

The extension cover displays the risk tier badge and explanation according to context. High risk sites openly trigger a full-page warning with an option to proceed or to go back.

(f) User feedback collection:

Many users report for facility but they also confirm detections. The feedback is being recorded and is being validated and it goes into the queue if there is an issue for the retraining purpose in the system.

(g) Adaptive retraining-

This validates new samples that are merged into the training model. The automated retraining system starts retraining and then evaluate the updated models, by deploying the much more improved version to the registration of the model.

This whole workflow gives us an assurance that Scam Shield is not simply a static detection tool but it is also a continuously improving, user in the loop cyber security ecosystem. The feedback loop is a critical separator or distinguisher, as new Phishing campaigns are emerging day by day, the confirmed samples are also absorbing at a pretty much high rate, also by degrading the chance of being vulnerable against novel attack patterns.

VII. IMPLEMENTATION DETAILS

(a) Technology Stack

The Scam Shield stack contains- Python 3.11 with Fast API for backend services, scikit learn and XG boost for tabular ML, Hugging Face Transformers (Distill BERT) for NLP inference, SHAPE and LIME for explainability, React + Typescript for the web UI and Web Extensions API for browser extension. The back end is mainly bundled with Docker and run smoothly through Kubernetes for horizontal scaling.

(b) Dataset

The training model combines 250,000 URLs from Phish Tank and Open Phish, 250,000 benign URLs from the Alexa Top 1M and Common Crawl, 50,000 social engineering email texts from the Spam Assassin public model an APWG phishing email archive and samples contributed by the user which were collected during the pilot testing. This dataset is balanced using stratified sampling and being increased with a different replacement for various NLP features to improve the generalization.

(c) Evaluation Methodology

The evaluation of this model uses stratified 5-fold cross validation with 80/20 train test splits. The Metrics report includes accuracy, precision, recall, F1score and false positive rate [FPR]. XAI quality is being measured through fidelity scores and user studies after comprehension. Latency is ruled on a standard basis for a consumer laptop (Intel Core i5, 8GB RAM) running Chrome 124.

VIII. COMPARATIVE ANALYSIS

Scam shield is evaluated with comparison to the most closely related systems which are being identified in the literature review. Table II presents a feature level comparison across all the dimensions from most critical to the real-world Phishing defense which contains

detection, scope, real time capability, explainability, adaptive learning and deployment form.

As shown in the table II Scam Shield is the only model and system that shows all the seven evaluation dimensions all together. It is the combination of URL text and behavioral analysis which will eliminate the single signal blind spot that will affect our prior works. The XAI and adaptive retraining capabilities represents the most important differentiators, whereas no system which is being surveyed provides both of the features. Also, the browser extension deployment gives us the assurance that at any point of risk, the detection is performed transparently without requiring any action which will be initiated by the user.

If we look from the side of an accuracy Scam Shield targets a higher performance than the old works from multi signal fusion which contains behavior signals, catch evasion attempts that defeat URL only systems while NLP semantic analysis catches social engineering content that defeats systems which were relying only on structural features.

IX. EXPECTED RESULTS AND OUTCOMES

Based on the design of the hybrid ML, NLP, XAI system and the multi signal feature extraction approach Scam Shield is projected to deliver the following outcome-

(a) Super 95% detection accuracy Across all the fishing URLs social engineering content and behavioral indicators. The grounded projection from the established performance of the individual components such as Distil BERT achieves greater than 93% F1 on phishing text classification benchmarks, while XG boost with URL structural features consistently achieves greater than 91% accuracy on standard datasets. Multi signal fusion is expected to yield further improvements through complementary error correction.

(b) False positive rate below 3% has achieved throughout the multi signal verification approach. A page that suddenly triggers an alert but presents clean URL structure and benign behavioral signals will fall from malicious to suspicious, reducing False positives that broke user's trust in all the security tools.

(c) Explainable alerts with greater than 80% user comprehension, they are measured after interaction surveys by asking users to identify any particular risk factor which has been flagged or pointed out by Scam Shield. It is main target is precisely measured against all the findings of Saha Roy et al. who showed that the contextual warnings which they have achieved has significantly higher user understanding that binary alerts.

(d) Adaptive retraining cycle of 24 to 48 hours which will ensure that all the newly confirmed Phishing campaign are absorbed into the model within 1 to 2 days of detection. This shows the most critical state of vulnerability which creates the window for any novel Phishing pages to easily get into any static classifiers.

(e) Sub 1.5 second end to end latency for the most primary risk classification response, that will give the assurance for a smooth user experience without any further unnecessary delay during simple and normal browsing of internet. The XAI explanation which is generated asynchronously Is delivered within 3 seconds.

(f) Educational engagement through contextually targeted small lessons is delivered through the user awareness interface which is a whole different other page. Users who receive any explanatory warnings are expected to show improved Phishing recognition in any follow up assessments or quiz which will contribute to a long-term security awareness for the future purposes.

X. LIMITATIONS AND FUTURE WORK

While Scam Shield presents a comprehensive approach to Phishing detection, several limitations merit acknowledgement. First of all, the current module of NLP is trained and evaluated only on English language content which will limit the effectiveness against Phishing campaigns which will target non- English speaking population countries, this will create a significant gap Given the global distribution of overall fishing victims Future work will combine or bring exclusively different transformer models to extend coverage to minimum 20 languages.

Second of all, analysis of all the behaviors module requires JavaScript and its execution to observe the page behavior which have to be dynamic and maybe avoided by the Phishing pages that will detect extension

presence and Will subdue all the malicious behaviors. Also, robustness testing and sandboxed behavioral execution are planned for future iterations.

Third of all, the feedback learning which is designed for rapid adaptation introduces a potential for adversarial poisoning, which means malicious actors mimicking as users and submitting the wrong reviews and feedback which will come under incorrect labels to degrade or suppress the whole model performance and will downgrade the whole system on the Internet if any new user was browsing. In near future we will implement robust label validation through which we will use multi source corroboration and anomaly detection on the feedback stream which will help us to know what which are proper reviews and feedbacks and which are the malicious ones.

Lastly deployment on a large scale within an organizational environment is planned in the next module or you can say a phase of this evaluation, which will provide a very longitudinal data on this model performance and its degradation, also for user adoption rates and the real-world impact of the XAI explanations on the security incident rates.

XI. CONCLUSION

This whole paper presents Scam Shield, a comprehensive cyber security solution that combines machine learning, natural language processing and Explainable AI for a real time scam and phishing detection. If you compare from the previous systems that optimize for detection accuracy in isolation, Scam Shield addresses all the three most important dimensions of modern Phishing together on defense which includes technical precision contextual transparency and adaptive resilience.

This system is designed for practical deployment as a browser extension and a web platform, meeting real time latency requirements while remaining accessible to the users which are not that technical. The XAI module also transforms Scam Shield from a passive detector into an educational tool which will remain active at all the time that means each and every interaction with a flagged site or a malicious site will become an opportunity for every user to build lasting phishing recognition skills.

The projected results will indicate accuracy exceeding over 95%, false positive rates below 3% and sub 1.5 second inference latency which will represent all the improvements over individually if we compare all the prior works across all the three dimensional being together. Thus, Scam Shield provides a scalable, human centric foundation for next generation Phishing defense, which is also adaptable to the evolving threat landscape.

Future work will only focus on the multilingual NLP support, robustness hardening, enterprise scale deployment trials and the expansion of the educational content library to go over a much broader and to cover on a wide range of social engineering attack.

ACKNOWLEDGMENT

The authors of this paper would like to thank the Department of Computer Science and Engineering at Meerut Institute of Engineering and Technology for providing all the useful resources and institutional support required for this research paper. The authors also would like to acknowledge the contributions of the open-source cyber security community whose publicly available Phishing datasets such as (Phish Tank, Open Phish, APWG crime archive) Which made the design an evaluation of the Scam Shield possible.

REFERENCES

[1] S. Arvind and R. K. Sharma, “PHISHTOR: A phishing site detection browser extension,” International Journal of Computer Science and Mobile Computing, vol. 14, no. 6, pp. 185–190, Jun. 2025.

[2] “A phishing detection system integrated as a browser extension using random forest classifier,” International Journal of Multidisciplinary Research, vol. 7, no. 2, Mar.–Apr. 2025.

[3] P. R. Shinde and R. N. Shinde, “Phishing site detection using Machine Learning and Natural Language Processing for robust detection,” International Journal of Advanced Research in Science, Communication and Technology, vol. 6, no. 2, pp. 23–28, Mar. 2024.

[4] M. Patel and R. J. Desai, “AI-enhanced phishing defense and real-time malicious UNIFORM RESOURCE LOCATOR detection,” International Journal of Creative Research Thoughts, vol. 12, no. 1, pp. 580–586, Jan. 2024.

[5] K. G. Rao and A. P. Nair, “Natural Language Processing for phishing detection: Leveraging AI,”

International Journal of Current Science, vol. 14, no. 4, pp. 43–48, Apr. 2024.

[6] B. John, "Building a browser extension for real-time phishing detection using Celery," ResearchGate, Jan. 2025.

[7] Z. T. Sworna, Z. Mousavi, and M. A. Babar, "Natural Language Processing methods in host-based intrusion detection systems: A systematic review and future directions," arXiv preprint arXiv:2201.08066, Jan. 2022.

[8] S. Saha Roy, C. Torres, and S. Nili Zadeh, "Explain, don't just warn! A real-time framework for generating phishing warnings with contextual cues," arXiv preprint arXiv:2505.06836, May 2025

FIGURE I

Reference	Technique	Real-Time	Explainable Artificial Intelligence	Adaptive
PHISHTOR [1]	Machine Learning + Uniform Resource Locator features	Yes	No	No
Rand. Forest [2]	Random Forest	Partial	No	No
Machine Learning + Natural Language Processing [3]	Machine Learning + Natural Language Processing	No	No	No
AI-Enhanced [4]	Ensemble Machine Learning	Yes	No	No
Natural Language Processing [5]	Transformer Natural Language Processing	No	No	No
Celery [6]	Machine Learning + Queuing	Yes	No	No
Natural Language Processing -IDS [7]	Natural Language Processing + Machine Learning	No	No	No
Explain [8]	Contextual Warn.	Yes	Partial	No
Scam Shield	Machine Learning + Natural Language Processing + Explainable Artificial Intelligence	Yes	Yes	Yes

FIGURE II

System	Uniform Resource Locator	Text	Behavioral	Real-Time	Explainable Artificial Intelligence	Adaptive	Browser Extension
PHISHTOR [1]	Yes	No	No	Yes	No	No	Yes
RF-Ext. [2]	Yes	No	No	Partial	No	No	Yes
Machine Learning + Natural Language Processing [3]	Yes	Yes	No	No	No	No	No
AI-Enh. [4]	Yes	Partial	No	Yes	No	No	No
Natural Language Processing [5]	Yes	Yes	No	No	No	No	No
Celery [6]	Yes	No	No	Yes	No	No	Yes
Scam Shield	Yes	Yes	Yes	Yes	Yes	Yes	Yes

FIGURE III

Metric	Target Value	Basis
Detection Accuracy	>95%	Multi-signal fusion (Machine Learning +Natural Language Processing)
False Positive Rate	<3%	Multi-branch verification
F1-Score (Phishing)	>0.94	Xg Boost + Distil BERT stacking
Inference Latency	<1.5 sec	Fast API + optimized inference
Explainable Artificial Intelligence Explanation Time	<3.0 sec	Async SHAP/LIME computation
User Comprehension	>80%	Contextual NL explanation
Retraining Cycle	24-48 hrs.	Airflow-orchestrated pipeline