

## MACHINE LEARNING

# Comparative Analysis of Different Machine Learning Models to Recognise Spam and Benign Mail Using URL

Debdutta Banerjee<sup>1</sup>, Soumendu Banerjee<sup>2</sup>

<sup>1</sup>Department of Computer Science, St. Xavier's College (Autonomous), Burdwan, India

<sup>2</sup>Department of Computer Application, Academy of Technology, Hooghly, India

Published October 2026

## Abstract

Phishing attacks are the most prevalent and concerning cybersecurity threats. Phishing typically involves exploiting malicious URLs to deceive users into revealing sensitive information or accessing fraudulent websites. The increasing rates of phishing techniques require the development of accurate and automated detection mechanisms. This study proposes a machine learning-based phishing URL detection framework that utilises Term Frequency-Inverse Document Frequency (TF-IDF) used here for feature extraction from URL strings. A comprehensive comparative analysis was conducted using four widely adopted supervised learning algorithms: Logistic Regression, Linear Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost). Experiments were performed on a large-scale phishing URL dataset comprising 871,590 labelled URL samples. The models were evaluated using standard performance metrics, including accuracy, precision, recall, and F1-score. The experimental results indicate that XGBoost, when combined with TF-IDF features, consistently outperformed the other classifiers, achieving the highest detection accuracy and superior classification performance. The findings demonstrate the effectiveness of integrating TF-IDF-based lexical feature extraction with ensemble learning techniques for phishing URL classification. The proposed framework offers a scalable and reliable solution for real-time phishing detection and can be integrated into web browsers, email filtering systems, and cybersecurity applications to enhance protection against evolving phishing attacks.

**Keywords:** phishing URL detection; machine learning; TF-IDF; XGBoost; cybersecurity; malicious URL classification

## 1 Introduction

The rapid growth of Internet services and the widespread adoption of digital technologies have fundamentally changed the way individuals and organisations access online resources with unprecedented convenience. This digital transformation has also expanded the cyber threat landscape, with phishing emerging as one of the most prevalent and financially damaging forms of cybercrime. Through phishing, cybercriminals may exploit deceptive emails, fraudulent websites, and malicious Uniform Resource Locators (URLs) to mislead users into disclosing sensitive information, including login credentials, financial details, and personal data, which is a major threat to society. Such attacks result in substantial financial losses, identity theft, reputational damage, and disruption of business operations worldwide. As phishing techniques continue to evolve through sophisticated URL obfuscation, domain spoofing, and dynamically generated web pages, traditional security mechanisms often struggle to identify newly emerging threats. Consequently, the development of accurate, scalable, and

intelligent phishing URL detection systems has become a critical research challenge for enhancing cybersecurity and protecting users from increasingly sophisticated online attacks.

Conventional phishing detection mechanisms were based on the blacklist databases, signature-based detection, and heuristic analysis to identify malicious websites and URLs. To overcome the limitations of conventional phishing detection techniques, machine learning (ML) has emerged as a promising solution due to its ability to automatically learn discriminative patterns from historical data and accurately classify previously unseen URLs. ML algorithms can generalise from learned characteristics and effectively detect zero-day phishing attacks without requiring frequent manual updates. The performance of phishing detection models largely depends on the quality of feature representation. Existing studies have extensively explored both lexical features, derived directly from URL strings, and structural features, such as URL length, the number of dots, digits, special characters, subdomains, HTTPS availability, and domain-specific attributes, to distinguish malicious URLs from legitimate ones. Among various feature extraction techniques, Term Frequency–Inverse Document Frequency (TF-IDF) has demonstrated remarkable effectiveness in capturing lexical patterns by transforming URL strings into weighted numerical vectors suitable for machine learning algorithms. Several supervised learning techniques have been investigated for phishing URL classification, including Logistic Regression, which provides an efficient linear baseline; Support Vector Machine (SVM), known for its robust high-dimensional classification capability; Random Forest, which improves prediction through ensemble decision trees; and Extreme Gradient Boosting (XGBoost), a powerful boosting algorithm recognized for its superior predictive performance and computational efficiency. Recent studies consistently report that ensemble learning approaches, particularly XGBoost, achieve higher detection accuracy than conventional classifiers when trained on appropriately engineered lexical and structural features. These findings highlight the importance of combining effective feature extraction techniques with advanced machine learning algorithms to develop accurate, scalable, and robust phishing URL detection systems.

A big number of existing approaches have been validated on relatively small or imbalanced datasets, making it difficult to assess their effectiveness in real-world deployment scenarios. Motivated by these limitations, this study proposes an enhanced phishing URL detection framework that combines character-level TF-IDF lexical features with handcrafted structural URL features while ensuring leakage-free model training through proper train–test splitting before feature extraction. The proposed framework is evaluated on a large-scale dataset containing 871,590 labelled URLs, providing a comprehensive and realistic assessment of model performance for practical phishing detection applications.

Motivated by the research gaps, this paper proposes an enhanced machine learning framework for phishing URL detection that integrates character-level TF-IDF lexical features with handcrafted structural URL attributes to improve classification accuracy and robustness. The framework employs proper train–test data partitioning before feature extraction to eliminate data leakage and ensure unbiased model evaluation. A comparative study is first conducted using four widely adopted supervised learning algorithms, namely Logistic Regression, Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost), to identify the most effective baseline classifier. Subsequently, the best-performing XGBoost model is enhanced through comprehensive feature engineering and optimised preprocessing techniques. The principal contributions of this work are discussed as follows:

In the first part, we have done a comparative performance analysis of four machine learning classifiers for phishing URL detection using a large-scale dataset comprising 871,590 labelled URLs. The second part consists of the integration of character-level TF-IDF representations with handcrafted structural URL features to capture both lexical and structural characteristics of phishing URLs. The implementation of a leakage-free training framework by performing train-test splitting before feature extraction is shown in the third part, and finally, a comprehensive performance evaluation using accuracy, precision, recall, F1-score, Receiver Operating Characteristic–Area Under the Curve (ROC-AUC), and confusion matrix analysis. The proposed framework demonstrates superior detection capability while maintaining computational efficiency, making it suitable for deployment in web browsers, email security gateways, enterprise network security solutions, and other real-time cybersecurity applications.

## 2 Literature review

Researchers have explored machine learning models that classify URLs using lexical and structural characteristics. Some recent studies have shown that character-level TF-IDF combined with machine learning classifiers significantly improves phishing detection accuracy by capturing hidden textual patterns within URLs [9].

Deep learning models [14, 15] such as URLNet automatically learn semantic representations of URLs without requiring extensive manual feature engineering. Although these models achieve high detection accuracy, they generally require large computational resources and training datasets [10]. Feature engineering also plays a crucial role in phishing detection, where URL length, number of subdomains, presence of special characters, HTTPS status, and domain-related attributes have been shown to improve classification performance [11].

Among ensemble learning techniques, XGBoost has become one of the most effective classifiers due to its scalability, robustness against overfitting, and ability to handle heterogeneous feature sets efficiently [12]. Recent studies have also highlighted that combining handcrafted URL features with textual representations yields better performance than using either feature type alone [13].

Therefore, this work adopts a hybrid feature extraction strategy by combining character-level TF-IDF with handcrafted URL features and evaluates multiple machine learning classifiers, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost.

## 3 Proposed methodology

The proposed phishing URL detection framework consists of five major phases.

### 3.1 Dataset collection

A benchmark phishing URL dataset containing 261,477 URLs was used for experimentation. The dataset includes 140,299 legitimate URLs and 121,178 phishing URLs. The dataset was divided into training and testing subsets using an 80:20 ratio before feature extraction to prevent train-test leakage.

---

### 3.2 Data pre-processing

The pre-processing stage included:

- Removing duplicate URLs.
- Handling missing values.
- Splitting the dataset before TF-IDF fitting.
- Parsing URLs using Python's `urlparse()` library.
- Encoding class labels into binary values.

Proper preprocessing minimizes bias and improves model generalization [11].

### 3.3 Feature extraction

Two complementary feature sets were extracted.

#### 3.3.1 Handcrafted URL features

- URL length
- Domain length
- Path length
- HTTPS availability
- Number of digits
- Number of dots
- Number of hyphens
- Number of special characters
- URL shortener detection
- Presence of IP address

#### 3.3.2 Character-level TF-IDF features

Character n-gram TF-IDF representation captures lexical patterns commonly observed in phishing URLs and has demonstrated excellent performance in previous phishing detection studies [9, 10].

The handcrafted and TF-IDF features were combined using Scipy Sparse HStack to create the final feature vector.

### 3.4 Machine learning models

The following classifiers were implemented and compared:

- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)
- XGBoost

XGBoost was selected as the primary model because of its high predictive capability, efficient gradient boosting mechanism, and excellent performance on structured cybersecurity datasets [12] We have also enhanced XGBoost to get better accuracy.

### 3.5 Performance evaluation

Performance was evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC
- Confusion Matrix

These metrics provide a comprehensive evaluation of phishing detection systems by measuring both classification accuracy and discrimination capability [13].

## 4 Experimental results and discussion

The accuracy, recall value, and F1-score of the above-mentioned four models have been shown using the diagram, which is given in the annexure section. Also, we have shown the confusion matrix analysis and displayed the ROC-AUC graph for all models. The experimental comparison conducted in the study shows that Logistic Regression provides a competitive linear baseline, but the ensemble models achieve slightly higher accuracy in the initial comparison. Logistic Regression obtained 79.81% accuracy, Random Forest obtained 80.42%, SVM obtained 79.47%, and the initial XGBoost model obtained 90.02%.

**Table 1.** Accuracy of the classification models.

| Model                        | Accuracy |
|------------------------------|----------|
| Support Vector Machine (SVM) | 79.47%   |
| Logistic Regression          | 79.81%   |
| Random Forest                | 80.42%   |
| Initial XGBoost              | 90.02%   |
| Enhanced XGBoost             | 93.00%   |

The difference between Logistic Regression and the initial XGBoost result is 1.97 percentage points, indicating that the initial improvement from changing the classifier alone was relatively modest. The substantially larger improvement was obtained after enhancing the feature engineering and preprocessing pipeline. This supports the observation that representation quality and leakage-free preprocessing are critical factors in phishing URL classification.

### 4.1 Enhanced XGBoost

Following the baseline comparison, the XGBoost pipeline was improved using the changes documented in the experimental implementation. First, the dataset was split into training and testing subsets before fitting the TF-IDF vectorizer, thereby preventing information from the test set from influencing feature extraction. Second, URL parsing was performed using `urlparse()`. Third, stronger handcrafted URL features were incorporated, including HTTPS status, long-path indicators, counts of dots and digits, and URL-shortener-domain indicators. Fourth, the earlier character-ngram indicators

were replaced with genuine character-level TF-IDF features extracted from URL strings. The handcrafted numerical features and character-level TF-IDF features were then combined using `scipy.sparse.hstack`. Finally, precision, recall, F1-score, ROC-AUC, and confusion-matrix analysis were incorporated into the evaluation, and the deprecated `use_label_encoder=False` setting was removed.

With these improvements, the enhanced XGBoost model achieved an overall accuracy of 93%. The reported classification results show precision, recall, and F1-score values of 0.94, 0.93, and 0.93 for the legitimate class, and 0.92, 0.93, and 0.92 for the phishing class, respectively. The macro-average and weighted-average precision, recall, and F1-score were each reported as 0.93. These results demonstrate a substantial improvement over the baseline classifiers and indicate that combining character-level lexical information with handcrafted structural URL features provides a more effective representation for phishing detection.

The comparative findings demonstrate that the enhanced XGBoost model substantially outperforms Logistic Regression, Random Forest, SVM, and the initial XGBoost configuration. More importantly, the improvement is associated with the combined effect of leakage-free preprocessing, richer structural URL features, and real character-level TF-IDF representations rather than simply selecting a more complex classifier.

The Logistic Regression results establish a useful and interpretable baseline for the phishing URL classification task. Its 79.81% accuracy and 0.8790 ROC-AUC show that URL lexical information contains sufficient discriminatory patterns for a linear classifier to learn meaningful decision boundaries. Nevertheless, the comparatively lower phishing recall of 0.75 indicates that a baseline linear model may fail to capture complex combinations of URL structure and lexical characteristics. Random Forest and SVM provide similar performance, while the initial XGBoost configuration improves accuracy to 90.02%. The enhanced XGBoost configuration increases accuracy to 93%, demonstrating the practical value of combining character-level TF-IDF with carefully engineered structural features and rigorous train-test separation.

From a deployment perspective, the results suggest that a phishing detector should not be assessed using accuracy alone. Precision, recall, F1-score, ROC-AUC, and confusion-matrix analysis provide complementary evidence about the model's ability to identify malicious URLs while limiting false alarms. Phishing recall is an important operational measure because undetected phishing URLs can directly expose users to fraudulent websites and credential theft. The enhanced model's more balanced class-wise precision and recall therefore make it a stronger candidate for integration into browser protection, email filtering, and other cybersecurity applications.

## 5 Conclusion

This study proposed a phishing URL detection system based on machine learning using handcrafted URL features and character-level TF-IDF representations. Four classifiers—Logistic Regression, Random Forest, SVM, and XGBoost—were evaluated. Experimental analysis showed that the enhanced XGBoost classifier achieved the highest accuracy of 93%, outperforming the remaining models. The proposed hybrid feature extraction approach successfully captures both structural and lexical characteristics of URLs, enabling accurate identification of phishing websites. The developed model can be integrated into web browsers, email security systems, and cybersecurity applications to provide real-time phishing detection.

## 6 Future work

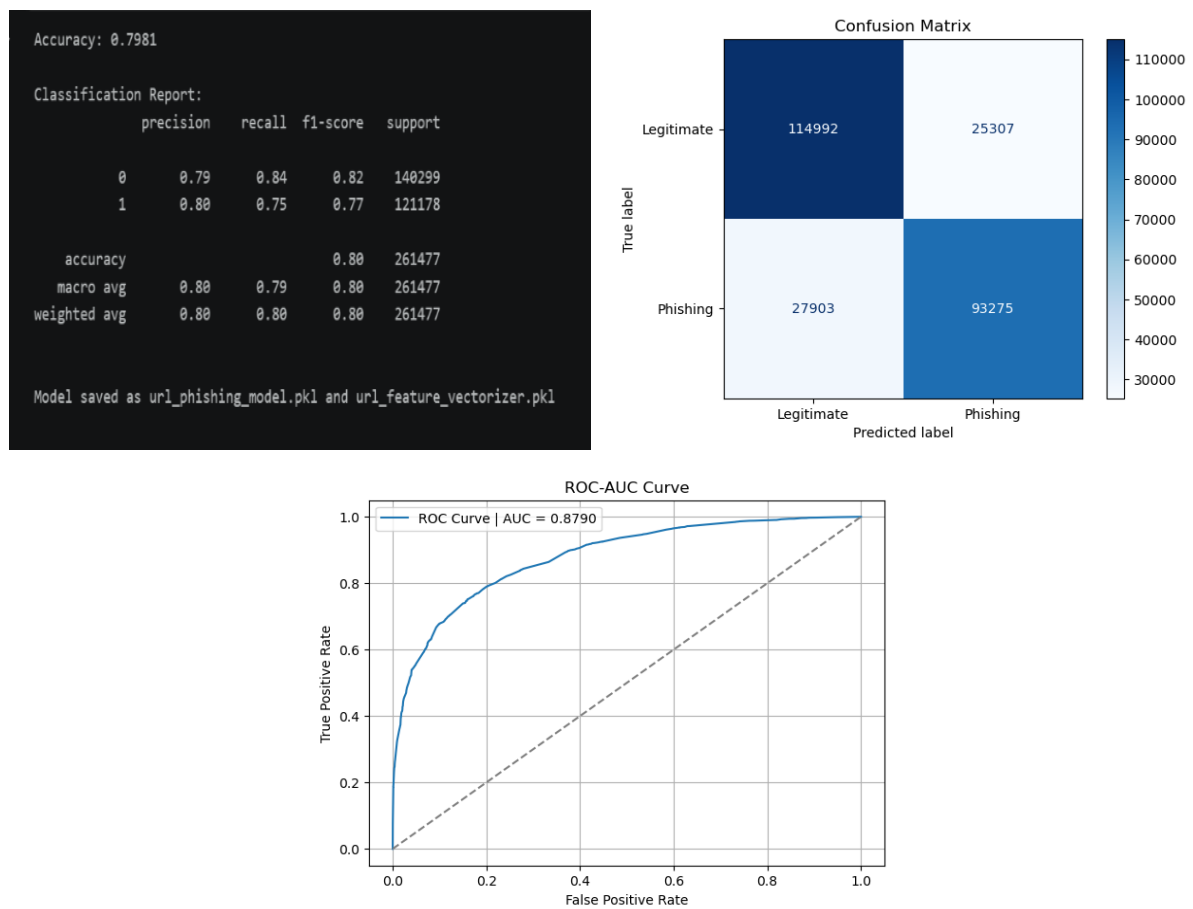
Future work may focus on:

- Incorporating webpage HTML and JavaScript features.
- Applying deep learning models such as CNNs, LSTMs, and Transformers.
- Developing lightweight real-time phishing detection systems for browsers and mobile devices.
- Using explainable AI techniques such as SHAP and LIME to improve model interpretability.
- Continuously updating the model using newly emerging phishing datasets to improve robustness against evolving cyber threats.

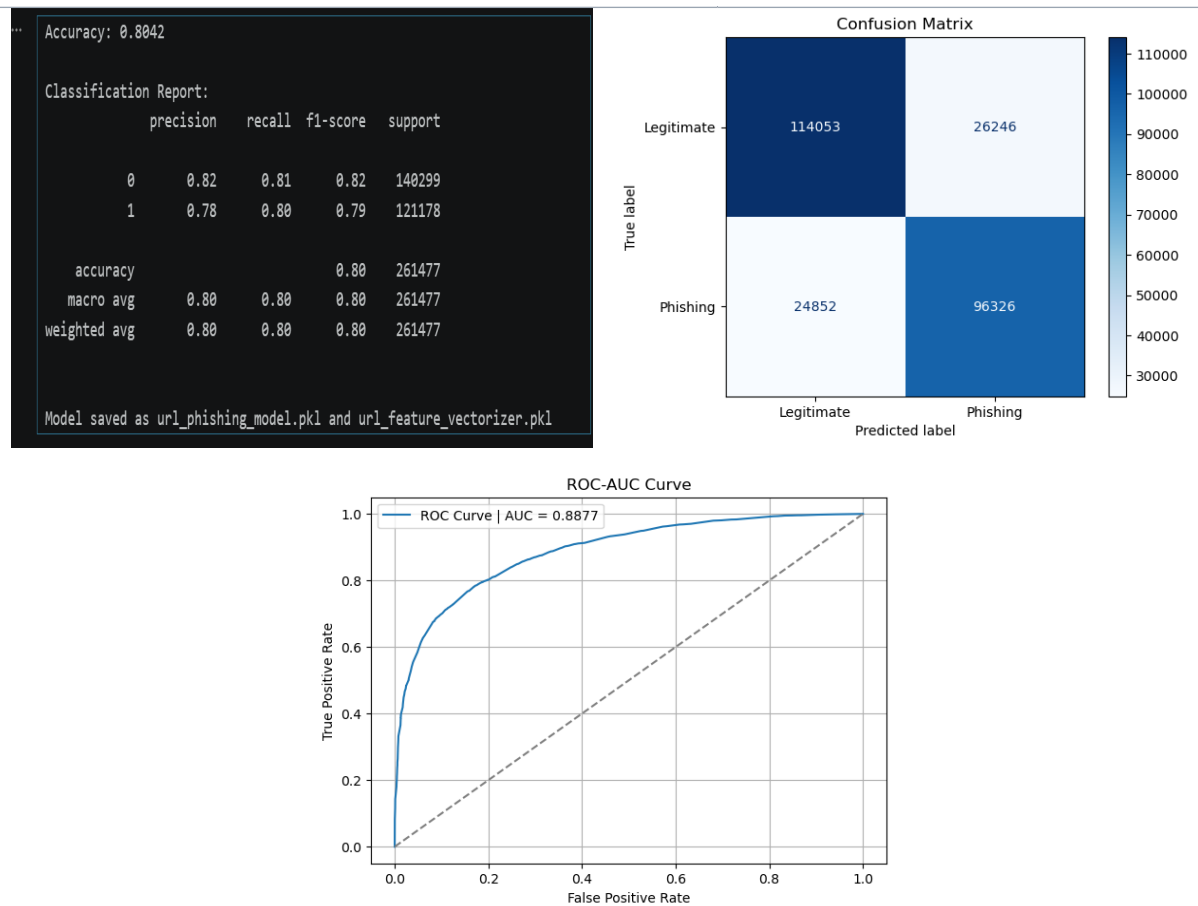
## References

- [1] APWG. *Phishing activity trends report*. Anti-Phishing Working Group; 2024.
- [2] Verma R, Dyer K. On the character of phishing URLs: Accurate and robust statistical learning classifiers. In: *Proceedings of the ACM CODASPY*. 2015.
- [3] Khonji M, Iraqi Y, Jones A. Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials*. 2013;15(4):2091–2121.
- [4] Marchal S, François J, State R, Engel T. PhishStorm: Detecting phishing with streaming analytics. *IEEE Transactions on Network and Service Management*. 2014;11(4):458–471.
- [5] Sahingoz Y, Buber E, Demir O, Diri B. Machine learning based phishing detection from URLs. *Expert Systems with Applications*. 2019;117:345–357.
- [6] James C, Joseph A, Balakrishnan V. *Phishing detection using machine learning techniques: A review*. Journal of Information Security and Applications. 2021.
- [7] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. p. 785–794.
- [8] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*. 1988;24(5):513–523.
- [9] Sanchez-Paniagua M, Fidalgo E, Alegre E, Al-Nabki W, Gonzalez-Castro V. Phishing URL detection: A real-case scenario through login URLs. *IEEE Access*. 2022;10:42949–42960.
- [10] Le H, Pham Q, Sahoo D, Hoi SCH. URLNet: Learning a URL representation with deep learning for malicious URL detection [preprint]. *arXiv*. 2018. arXiv:1802.03162
- [11] Kustiawan YA, Ghauth KI. Evaluating the impact of feature engineering in phishing URL detection: A comparative study of URL, HTML, and derived features. *IEEE Access*. 2025;13:126756–126768.
- [12] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2016. p. 785–794.
- [13] Asiri S, Xiao Y, Alzahrani S, Li S, Li T. A survey of intelligent detection designs of HTML URL phishing attacks. *IEEE Access*. 2023;11:6421–6443.
- [14] Chanda T, Mondal B, Das S, Banerjee S, Mondal B. An explainable AI-based hybrid SVM–CNN ensemble model implementation for intelligent email spam detection. In: *2026 International Conference on Emerging Systems and Intelligent Computing (ESIC)*; Bhubaneswar, India. 2026. p. 240–245. <https://doi.org/10.1109/ESIC68176.2026.11496294>
- [15] Chanda T, Mondal B, Das S, Banerjee S, Mondal B. Temporal analysis based exploratory data analysis for phishing email detection via machine learning. In: Chatterjee B, Kothapalli K, Mittal N, Natarajan AM, Singh D, editors. *Distributed computing and intelligent technology (ICDCIT 2026)*. Cham: Springer; 2026. (Lecture Notes in Computer Science; vol. 16420). [https://doi.org/10.1007/978-3-032-16632-6\\_21](https://doi.org/10.1007/978-3-032-16632-6_21)

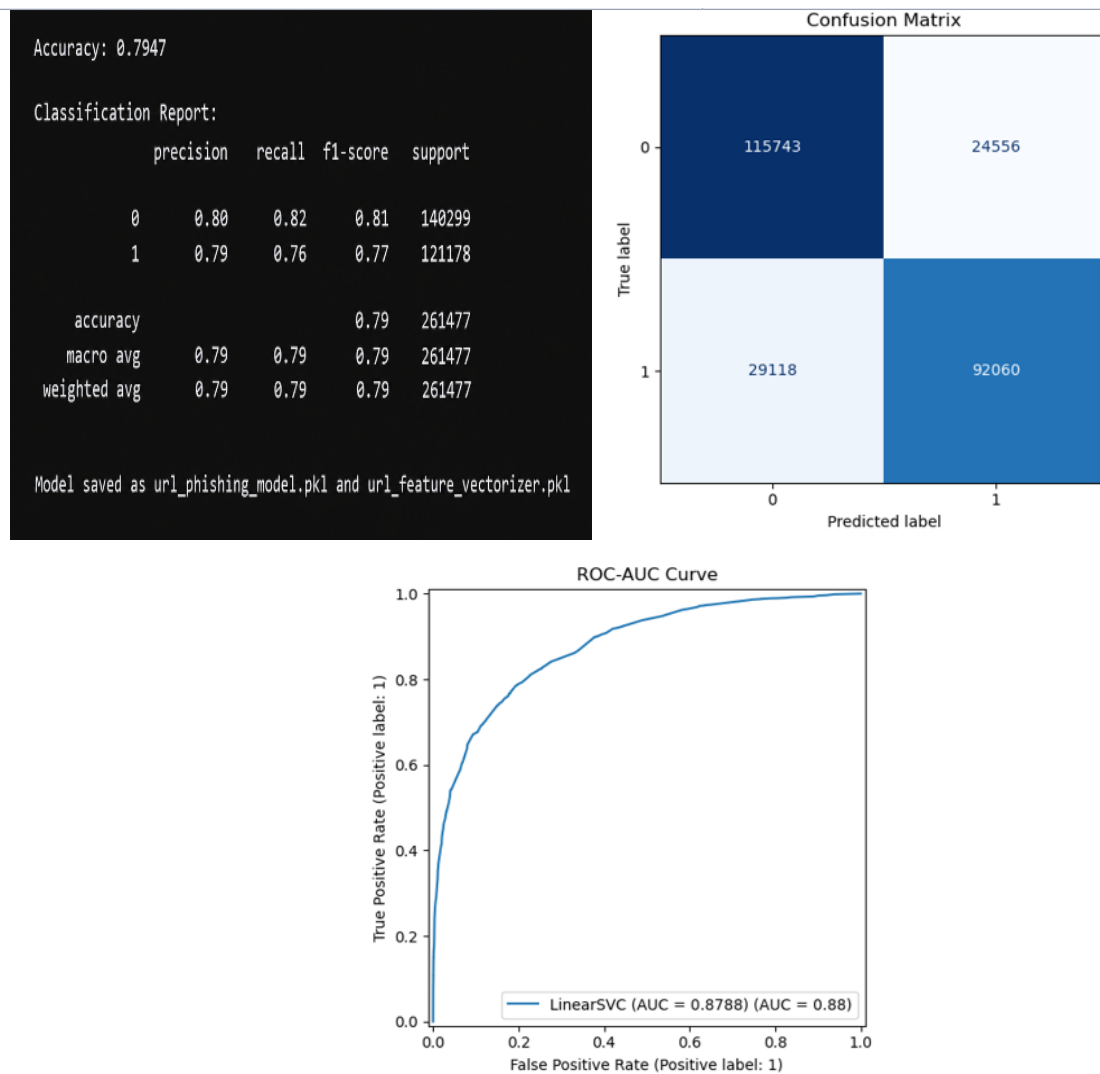
## Appendix A. Model results



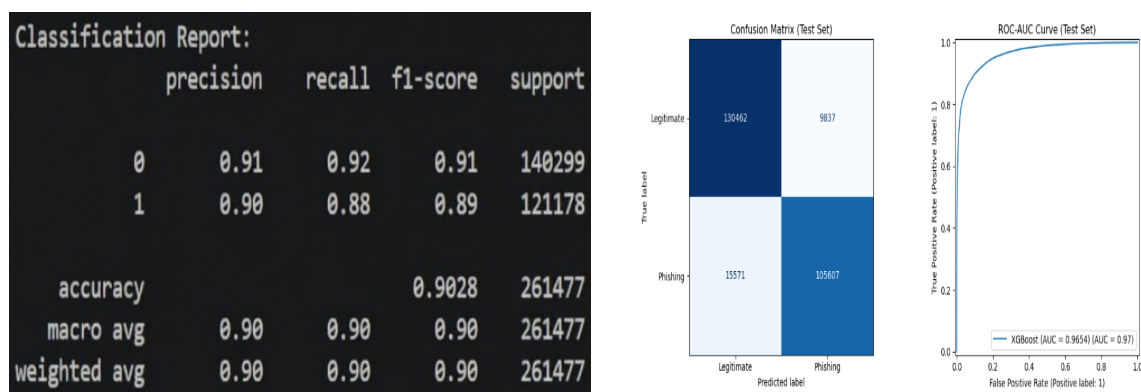
**Figure 1.** Logistic regression: (a) classification report; (b) confusion matrix; (c) ROC-AUC curve.



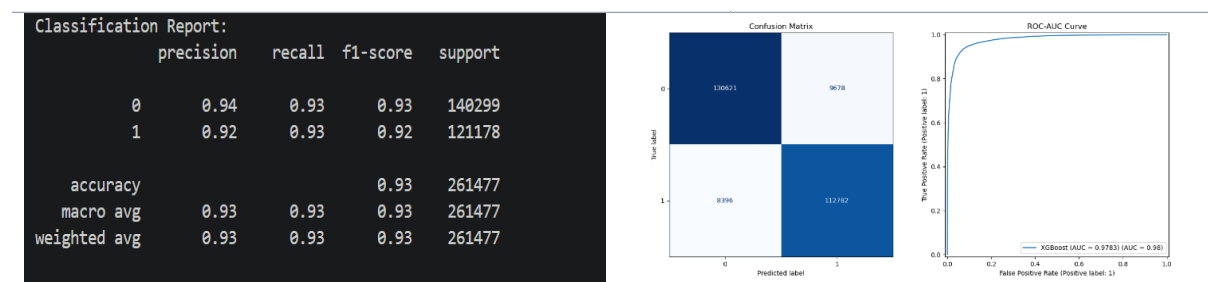
**Figure 2.** Random forest: (a) classification report; (b) confusion matrix; (c) ROC-AUC curve.



**Figure 3.** Support vector machine (SVM): (a) classification report; (b) confusion matrix; (c) ROC-AUC curve.



**Figure 4.** XGBoost: (a) classification report; (b) confusion matrix and ROC-AUC curve.



**Figure 5.** Enhanced XGBoost: (a) classification report; (b) confusion matrix and ROC-AUC curve.