Journal of Advances and Scholarly Research In Allied Education
Vol.23, Issue No. 3 June 2026, ISSN 2230-7540

Predictive Analytics and Machine Learning-Based Models for E-Commerce Fraud Prevention

Sachin Bagoria[1]*, Dr. Kavita[2]

1 Research Scholar, SKD University, Hanumangarh, Rajasthan, India

radheykrishnalalita@gmail.com

2 Professor, SKD University, Hanumangarh, Rajasthan, India

Abstract: The e-commerce market has grown but so have online fraud and criminality. Online marketplaces can be very complex and customers are becoming more adept at new methods of fraud, which make traditional fraud detection methods ineffective. The aim of this study is to design a machine learning and predictive analysis system for e-commerce fraud prevention. It takes into account the structure of the URL, the content of the HTML, the technology profiles, SSL certificates, HTTP headers and external reputation indicators. 2,031 ecommerce sites were used for model creation and evaluation, with 739 of them being fraudulent and 1,292 authentic. XGBoost, Odd Forest, Support Vector Machine, Logistic Regression, k-Nearest Neighbour, AdaBoost, & Naïve Bayes were used to extract and evaluate 50 features. Experimentally, XGBoost outperformed the baseline with all characteristics at 0.9688 F1-Score & 97.78 accuracy rate while the baseline is 0.9653 & 97.49%. The comparative investigation revealed the superiority of the proposed system over existing fraud detection systems. The results suggest that machine learning-based predictive analytics could be a scalable and powerful tool to protect online transactions and detect fraudulent ecommerce sites.

Keywords: E-commerce Fraud Detection, Predictive Analytics, Machine Learning, XGBoost, Cybersecurity, Fraud Prevention, Classification Models, Data Mining, E-commerce Security, Artificial Intelligence.

1. INTRODUCTION

Internet technology has influenced consumers in their purchasing and accelerated global e-commerce. Consumers increasingly utilised Amazon, eBay, Alibaba, and Facebook Marketplace to buy products and services from home during the COVID-19 epidemic [1]. All the possibilities that growth has provided to companies and consumers have come with a price, namely higher cybercrime and online fraud. Fraudulent online shopfronts, phishing attacks, identity theft, account takeovers, bogus product listings, money laundering, and unauthorised financial transactions include e-commerce fraud [2,4]. Fraudulent activities can be expensive to companies and create a loss of trust on platforms. The losses from online payment fraud are increasing, thus the importance of fraud detection and prevention [5]. Cybercriminals' continually developing techniques make rule-based systems, human verification, and

authentication processes unsuitable for fraud prevention [6]. Therefore, organisations are using modern technologies like machine learning (ML), artificial intelligence (AI), predictive analytics, and data mining to identify fraud in huge and complicated datasets. These techniques show hidden patterns and abnormalities that can be used to distinguish between fraudulent and legitimate transactions [6–11].

Fraud detection methods based on network analysis to detect suspicious transactions and interactions between users, transactions, and digital entities have also been investigated [12]. Due to its capacity to quickly analyse massive amounts of structured and unstructured data, machine learning-based predictive analytics remains one of the most successful and scalable ways for contemporary e-commerce. In this study, an e-commerce fraud avoidance technique based on predictive analytics and machine learning is used. This method relies on 50 predictive factors included in the URL structure, HTML content, technology profile, SSL certificate, HTTP headers, and external reputation signals. The project aims to develop a robust and scalable solution to detect fraudulent e-commerce sites and enhance commercial security online, evaluating several machine learning algorithms for their effectiveness.

2. OBJECTIVES

- To create and assess machine learning-based prediction models that use website-related information to reliably identify fraudulent e-commerce websites.
- To determine the best model for preventing e-commerce fraud by comparing the performance of several machine learning algorithms.

3. RESEARCH METHODOLOGY

The quantitative and predictive analytics-based research approach is used to develop and evaluate machine learning models to detect and prevent fraudulent online stores. The technique involves collecting data, feature extraction, feature engineering, model training, hyperparameter optimisation, and performance evaluation. The aim is to develop a system that can identify fake online stores and pinpoint patterns that distinguish them from legitimate ones so that they can be detected..

3.1 Dataset Description

In the sample of 1,292 real e-commerce websites, there were 739 fraudulent websites. The websites were categorized into 13 different industries: apparel, sports, technology, health, housing, food, education, entertainment, pets, toys, & office and industrial supplies.

The distribution of real websites, on the other hand, was more even, while the distribution of fraudulent websites was the greatest in the categories of fashion (41.95%), marketplace (31.27%), & sport (14.21%). Scammers often aim for high-demand consumer sectors, as shown by this distribution, which is reflective of actual fraud tendencies.

3.2 Feature Extraction

The third version of Python was used to create a thorough feature extraction framework. To create data from multiple online sources, six separate modules were created:

1. URL Features.

2. HTML Features.

3. Technology Features.

4. SSL Certificate Features.

5. HTTP Header Features.

6. External Reputation and Social Media Features.

For each module extracted, features were extracted and saved in a JSON structure, where the keys are feature names and the values are features. 50 qualities were grouped into 6 categories. The following criteria were used to identify e-commerce websites: structure, behavior, technology and reputation.

3.3 Feature Engineering.

Using the six feature groups, we created unique feature vectors for every website.

URL Feature Vector: $Uv = [U1, U2.1, U2.2, ..., U3.5]$

HTML Feature Vector: $Hv = [H1, NH2, NH3, ..., NH8.3]$

Technology Feature Vector: $Tv = [NT1, NT2.1, NT2.2, ..., NT3.3]$

SSL Feature Vector: $Sv = [S1, S2]$

HTTP Header Feature Vector: $Pv = [NP1, NP2, ..., P6]$

External Feature Vector: $Ev = [NM1, NM2, ..., NM8]$

These vectors were concatenated to create a complete feature vector:

$$Fv = [Uv, Hv, Tv, Sv, Pv, Ev]$$

After processing all websites, feature vectors were ordered into a M × N matrix X, where M represents extracted features & N represents the total website samples.

3.4 Link Structure Analysis

HTML link architecture was studied to extract the features. Six categories were used to group the links:

- External Links.
- Internal Connections.
- Links for Internal References.
- No Links to Content.
- Links that contain no text. Links with no text attached.
- Links that don't have a reference (href).

Their distribution and frequency were used to indicate the structural quality and legitimacy of the website, for these link types.

3.5 Data Pre-processing

The feature values were standardised with the help of StandardScaler function of scikit learn prior to training the models. Feature scaling was performed on both training and testing sets to have all the models agree on the scaling.

3.6 Machine Learning Models

Eight popular machine learning algorithms in fraud detection research were used:

- x2 extreme increasing boosting (XGBoost).
- Random Forest(RF) Classifier.
- Random Forest (RF).
- Support Vector Machine (SVM).
- Logistic Regression (LR).
- k-Nearest Neighbour (kNN).
- AdaBoost.
- Naïve Bayes (NB).

The best hyperparameters were searched for each of the classifiers.

3.7 Model Validation

These parameters were used in a 5-fold Cross-Validation technique to ensure the robustness of the model and reduce model bias:

- Number of folds (k) = 5.
- Shuffle = True.
- Random State = 42.

The five parts of the data set were divided. The data set was divided into five equal parts. Four training subsets and one testing subset were used in each cycle. All folds were used to calculate performance measures and averaged results were used.

3.8 Performance Evaluation Metrics

Four common classification measures were used to assess the performance of each of the classifiers:

- **Precision:** Proportionate rate of correct identification of fraudulent sites out of total number of fraudulent sites.
- **Recall:** Calculates the percentage of correct identification of fraudulent sites by the model.
- **Accuracy:** Calculates the overall accuracy of the websites being classified.
- **F1-Score:** It is the harmonic average of Precision and Recall and gives a fair balance for the performance of the classifiers.

Confusion matrix was made up of:

- **True Positives (TP):** Correctly classified fraudulent websites.
- **False Positives (FP):** Not spam websites that end up being classified as spam.
- **True Negatives (TN):** Valid websites correctly identified.
- **False Negatives (FN):** Fraudulent websites misidentified as legitimate.

3.9 Experimental Setup

A preliminary evaluation of two experimental approaches was made:

- **Complete Feature Set Model:** All retrieved features were used in this model, such as social media cues and outside reputation.
- **Independent Model:** This method only used certain website elements that were available locally; elements from external reputation were not used. The aim was to

develop an independent fraud detection programme which does not rely on other services.

3.10 Comparative Analysis

The framework recommended was discussed with Wu and Wadleigh's work done regarding fraud detection. All benchmarks used the same experimental techniques and machine learning optimisation to maintain the impartiality of the comparisons. Elements that were not available for the comparison models due to privacy, GDPR and third-party API restrictions were not included. To assess the effectiveness of the predictive analytics system in preventing online purchasing fraud, we used the metrics of accuracy, precision, recall, or F1-Score.

Table 1: Comparison of the proportion and number of categories on the authentic and counterfeit websites

Category	Fraud (n)	Fraud (%)	Legit (n)	Legit (%)
Automotive	5	0.68	45	3.48
Education	0	0.00	75	5.80
Entertainment	3	0.41	62	4.80
Fashion	310	41.95	179	13.86
Food	2	0.27	100	7.74
Health	11	1.49	142	10.99
Home	19	2.57	208	16.10
Marketplace	231	31.27	142	10.99
Office and Industrial Material	6	0.81	56	4.33
Pets	1	0.14	19	1.47
Sport	105	14.21	65	5.03
Technology	17	2.30	176	13.62
Toys	29	3.92	23	1.78
Total	**739**	**100.00**	**1292**	**100.00**

Total Transactions = 2,031 (Fraud = 739; Legit = 1,292).

Table 2: Link types based on the href tag's origin or destination

Type of Link	Example
External	
Internal	
Internal Reference	
No Content	
Empty	
No Reference	<a>

4. RESULT

We used a 3.6 GHz Intel Core i3 9100F & 16 GB DDR4 RAM. For several experiments and machine learning model creation, we utilised scikit-learn13 & Python 3. Nine state-of-the-art classification techniques were employed to assess and compare design aspects [13, 14]. Algorithms such as XGBoost, GBC, RF, kNN, SVM, LR, NB, and Adaboost are used. Classifiers were trained using the best hyper-parameters from a 5-fold cross-validated grid search. Table 3 lists the finished projected model hyperparameters. We scaled the features vector across all features as well as instruction data using scikit-learn's StandardScaler, then applied it to test samples.

Table 3: An overview of the features that have been implemented and the group that corresponds to them

Feature ID	Group	Feature Name	Value Type	Description
U1	URL	domain_digit_count	D	Number of digits in the domain name
U2.1	URL	domain_length	D	Number of characters in the domain name
U2.2	URL	subdomain_length	D	Number of characters in the subdomain
U3.1	URL	raw_word_count	D	Number of words in the URL

U3.2	URL	average_word_length	C	Average length of words in the URL
U3.3	URL	longest_word_length	D	Length of the longest word in the URL
U3.4	URL	shortest_word_length	D	Length of the shortest word in the URL
U3.5	URL	std_word_length	C	Standard deviation of word lengths
H1	HTML	text_length	D	Number of characters in the HTML text
NH2	HTML	domain_title	B	Whether the domain appears in the title
NH3	HTML	domain_in_html	D	Number of times the domain appears in HTML text
NH4	HTML	base64	B	Website loads resources encoded in Base64
H5.1	HTML	link_int	D	Number of internal links
H5.2	HTML	link_ext	D	Number of external links
H5.3	HTML	link_#	D	Number of empty (“#”) links
H5.4	HTML	link_emp	D	Number of empty links
H5.5	HTML	link_null	D	Number of links without href attribute
H6	HTML	currencies	D	Number of currencies detected on the website
NH7.1	HTML	prices	D	Total number of prices detected
H7.2	HTML	most_times	D	Repetitions of the most frequent price
NH7.3	HTML	avg_times	C	Average repetitions of prices
NH7.4	HTML	avg_discount	C	Average discount percentage
NH8.1	HTML	num_social_html	D	Number of social media links
NH8.2	HTML	fake_fb	B	Facebook share link detected
NH8.3	HTML	fake_tw	B	Twitter share link detected

NT1	Tech	n_tech	D	Number of technologies detected
NT2.1	Tech	e-commerce	D	Number of e-commerce technologies used
NT2.2	Tech	live-chat	D	Number of live-chat technologies used
NT2.3	Tech	cookie-compliance	D	Number of cookie-compliance technologies used
NT2.4	Tech	analytics	D	Number of analytics technologies detected
NT2.5	Tech	payment-processors	D	Number of payment-processing platforms detected
NT3.1	Tech	google-analytics	B	Website uses Google Analytics
NT3.2	Tech	google-analytics-enh	B	Website uses enhanced Google Analytics for e-commerce
NT3.3	Tech	recaptcha	B	Website uses reCAPTCHA
S1	SSL	has_cert	B	Domain uses a valid SSL certificate
S2	SSL	n_name	D	Number of domain names registered in SSL certificate
NP1	HTTP	content-security-policy	B	Website defines a Content Security Policy (CSP) header
NP2	HTTP	strict-transport-security	B	HSTS is implemented
NP3	HTTP	x-content-type-options	B	Nosniff directive is enabled
NP4	HTTP	x-frame-options	B	Uses DENY or SAMEORIGIN directives
P5	HTTP	cache-control	B	Website avoids outdated post-check directive
P6	HTTP	expect-ct	B	Expect-CT header is configured
NM1	External	total_followers	D	Total social media followers
NM2	External	total_following	D	Total accounts followed on Instagram and Twitter

NM3	External	total_posts	D	Total posts on Instagram and Twitter
NM4	External	fb_likes	D	Facebook page likes
NM5	External	fb_visits	D	Facebook page visits in the last 24 hours
NM6	External	tw_age	D	Months since Twitter account registration
NM7	External	trustpilot_score	C	Trustpilot review score
NM8	External	trustpilot_reviews	D	Number of Trustpilot reviews

Table 4: Machine learning classifier evaluation for the suggested techniques

Classifier	Hyper-parameter	Value
XGBoost	eval_metric	error
	n_estimators	120
	objective	binary
	scale_pos_weight	2
Gradient Boosting Classifier (GBC)	learning_rate	0.1
	max_depth	3
	max_features	sqrt
	n_estimators	242
Random Forest (RF)	max_features	auto
	n_estimators	127
k-Nearest Neighbors (kNN)	metric	manhattan
	n_neighbors	2
	weights	uniform
Support Vector Machine (SVM)	C	100
	gamma	0.001
	kernel	rbf

Logistic Regression (LR)	C	0.1
	penalty	l2
AdaBoost	learning_rate	0.1
	n_estimators	43
Naïve Bayes	kind	BernoulliNB

Finally, we used k-fold cross-validation with k = 5, shuffle = T rue, & random_state = 42 to evaluate classifier performance. We used averaged 5-fold cross-validation data to provide accuracy, precision, recall, and F1-Score [15, 16]. The number of fraudulent websites recognised properly is called true positives (TP). The FP is the number of legitimate samples misclassified as fraudulent. A properly classified sample is a true negative (TN). Finally, false negatives (FN) are fake websites misclassified as genuine.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

This work develops a machine-learning system to notify clients about fraudulent e-commerce websites. This study investigated two strategies. The initial optimised performance with all intended features, including external services. The second technique produces a local-only model. Table 5 reveals that XGBoost had the highest F1-Score (0.9688) of all features, then GBC (0.9684) & Random Forest (0.9661). At 97.78% accuracy, the XGBoost algorithm can classify data, making it suitable for bogus website identification. Random Forest raised accuracy by 0.0069 & lowered recall by 0.0118, while the top two stars did not change. This is suggested for systems the requirement to detect particular attacks.

Table 5: Machine learning classifier evaluation for the suggested techniques

Algorithm	Full Set Precision	Full Set Recall	Full Set F1-Score	Full Set Accuracy (%)	Standalone Precision	Standalone Recall	Standalone F1-Score	Standalone Accuracy (%)
XGBoost	0.9778	0.9601	**0.9688**	**97.78**	0.9647	0.9660	**0.9653**	**97.49**
GBC	0.9751	0.9619	0.9684	97.73	0.9590	0.9686	0.9637	97.39
Random Forest	0.9847	0.9483	0.9661	97.59	0.9765	0.9342	0.9546	96.80
SVM	0.9622	0.9602	0.9611	97.19	0.9619	0.9618	0.9618	97.24
Logistic Regression (LR)	0.9566	0.9633	0.9599	97.09	0.9535	0.9576	0.9555	96.80
kNN	0.9564	0.9366	0.9463	96.21	0.9337	0.9481	0.9407	95.72
AdaBoost	0.9373	0.9534	0.9452	96.01	0.9444	0.9454	0.9448	96.01
Naïve Bayes (NB)	0.9231	0.9412	0.9320	94.84	0.9286	0.9346	0.9316	94.84

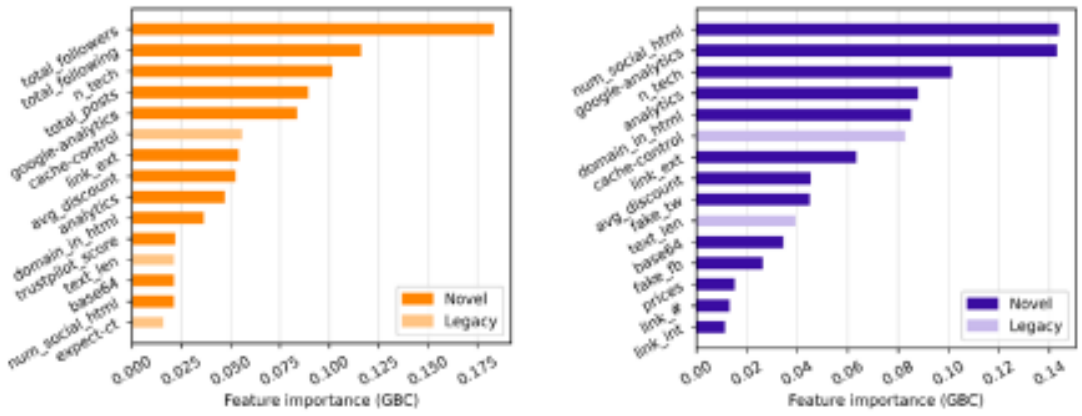

Figure 1: Full set and solo model feature important. In light colours, heritage traits from prior works; in deeper colours, unique features presented in this work.

Results of the experiment showed that XGBoost algorithm achieved the highest classification performance with the F1-Score value of 0.9688 and the accuracy of 97.78% when all the features were used. Random Forest had a slight improvement in the precision score, but a lower recall score, which makes XGBoost a better option for fraud detection. The standalone version also had a very good performance with an F1-Score of 0.9653, meaning that the performance of external reputation features did not make a large impact on the general performance. The feature importance analysis showed that the social media indicators, technology related features, HTML characteristics, and pricing information are the most important features in predicting fraudulent websites, while URL based features play a relatively small role. Additional experiments revealed that, if the HTML and technology aspects were removed, the performance of the models would be drastically diminished, indicating that they do play a significant role in fraud detection. The proposed framework was shown to be efficient and independent from language-specific, brand-specific, and expensive third-party resources, achieving a better performance than the current approaches, and being practical and scalable for e-commerce fraud prevention.

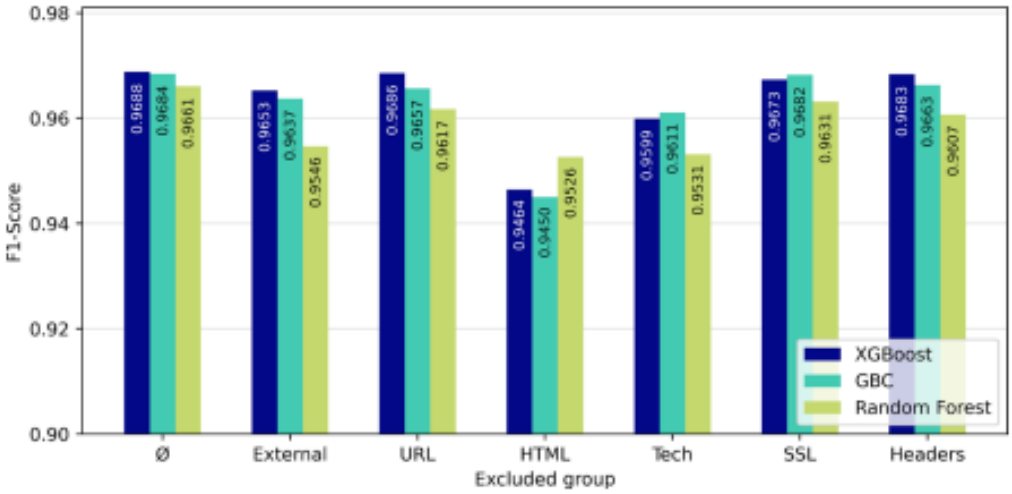

Figure 2: Findings for the top-performing algorithms when one of the resources is removed

The second experiment shows how well suggested features work alone. The 17 HTML features had the highest XGBoost and Random Forest F1-Scores, 0.9541 and 0.9564. URLs are needed for 7 of 17 HTML functions. Thus, low-resource systems should use it. The eight external feature model placed second with 0.8667 and 0.8662 GBC and Random Forest F1-Scores. The findings support this group's eight aims. Major issue: it relies on others and cannot run without specific services. The nine Wappalyzer technology report features for XGBoost and GBC have 0.8329 and 0.8333 F1-Scores, suitable for general-purpose systems. This experiment's URL, SSL, and HTTP Headers setting failed for several reasons. Due to domain name similarities, the URL set cannot distinguish fake and legitimate websites. Features were designed without keywords or brand lists since they may develop language-dependent models. Low SSL set results were due to a lack of model input as it had two features. The latest three HTTP Header sets were above-average (0.7740 GBC F1-Score). Its biggest drawbacks are its high sensitivity and false positive rate (0.9094 recall and 0.6746 accuracy on GBC) Figure 3. Compare the proposed methods to Wu et al. (2018) and Wadleigh (2015) [17,18]. Compare to other publications is impossible since none published data or restricted their methodology. Current works lack feature and method implementation information, thus we generated our own extraction, which may be different but accurate. These works cannot use third-party features in Europe under GDPR. Table 6 lists unimplemented features. For fairness, we will compare these methods to our solo version without third-party data. We employed the same

experimental methodology to discover the optimum hyper-parameters as these investigations were not disclosed.

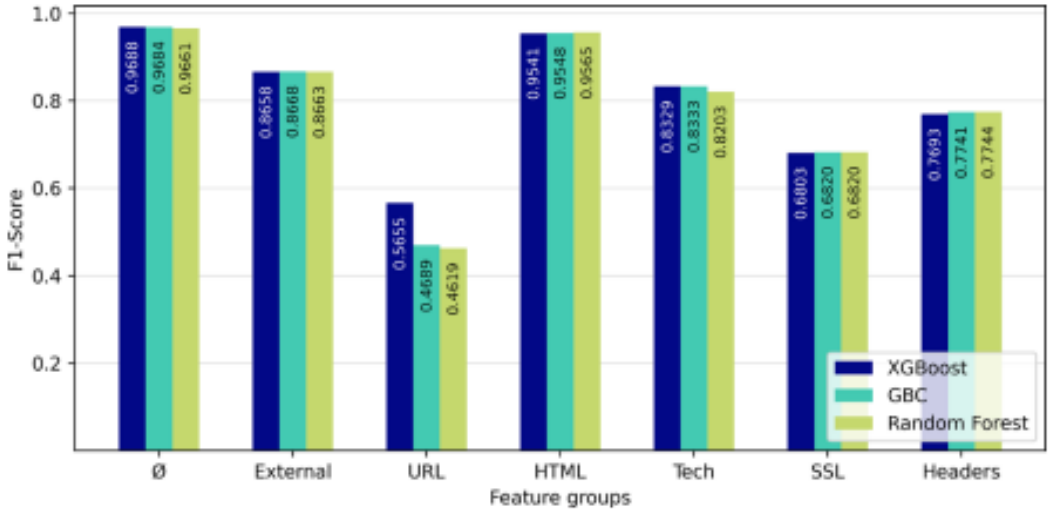

Figure 3: Findings for the top-performing algorithms when using a separate set of characteristics that match the resources used in this study

Table 6: Comparison Between Our Method and Existing Works

Work	Classifier	Precision	Recall	F1-Score	Accuracy (%)
Our Standalone Method	XGBoost	0.9647	0.9660	0.9653	95.49
Wu et al. [7]	Random Forest (RF)	0.9224	0.8795	0.9003	92.91
Wadleigh et al. [20]	XGBoost	0.6599	0.7715	0.7111	77.25

Note: Values representing the highest performance for each metric are shown in bold.

Table 7: Features Not Implemented in the Comparison Works

Work	Feature	Reason for Exclusion
Wadleigh et al. (2015)	Private or China WHOIS	No WHOIS data is publicly available for most EU websites
Wadleigh et al. (2015)	WHOIS Registration < 1 Year	No WHOIS data is publicly available for most EU websites

Wadleigh et al. (2015)	Website on Takedown Page	Our dataset contains no seized websites, only working ones
Wadleigh et al. (2015)	Website in Alexa Top 100K	Costly API requirement
Wu et al. (2018)	in_top_one_million	Costly API requirement
Wu et al. (2018)	china_registered	No WHOIS data is publicly available for most EU websites
Wu et al. (2018)	under_a_year	No WHOIS data is publicly available for most EU websites

Table 7 shows that our technique is better than the state-of-the-art currently available. Not only that, the suggested features don't rely on any outside data, thus they should work reliably across all countries and languages.

5. CONCLUSION

With online shopping, there has been a rise in online fraud, and there is a need to have sophisticated and effective fraud detection systems. In this study, 50 URL structure, HTML, technological profile, SSL certificate, HTTP header, and external reputation database features were used to identify fake e-commerce websites. The system was powered by machine learning and predictive analytics. A number of machine learning techniques were compared such as XGBoost, Gradient Boosting Classifier, Random Forest, SVM, Logistic Regression, k-Nearest Neighbour, AdaBoost and Naïve Bayes. XGBoost was the most successful with 96.78% feature set accuracy and 0.9688 F1-Score. The solo model achieved a good result in terms of accuracy of 97.49% and an F1-Score of 0.9653. The feature importance analysis revealed it was important HTML, technology and external reputation aspects. The framework offers a dependable, scalable, and accurate means of fighting e-commerce fraud, which helps to make online transactions safer.

References

1. Monteith, S., Bauer, M., Alda, M., Geddes, J., Whybrow, P. C., & Glenn, T. (2021). Increasing cybercrime since the pandemic: Concerns for psychiatry. *Current Psychiatry Reports, 23*(4), 18.

2. Kodate, S., Chiba, R., Kimura, S., & Masuda, N. (2020). Detecting problematic transactions in a consumer-to-consumer e-commerce network. *Applied Network Science, 5*(1), 90.

3. Samani, R., & Davis, G. (2019). *McAfee mobile threat report*. McAfee. https://www.mcafee.com/enterprise/en-us/assets/reports/rp-mobile-threat-report-2019.pdf

4. Smith, S., & Juniper Research. (2024). *Online payment fraud: Market forecasts, emerging threats & segment analysis 2022–2027*. Juniper Research. https://www.juniperresearch.com/press/losses-online-payment-fraud-exceed-362-billion/

5. Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559–569.

6. Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications, 68*, 90–113.

7. Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–255.

8. Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). *A comprehensive survey of data mining-based fraud detection research* (arXiv:1009.6119) . arXiv. https://arxiv.org/abs/1009.6119

9. Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery, 29*(3), 626–688.

10. Irani, D., Webb, S., & Pu, C. (2010). Study of static classification of social spam profiles in MySpace. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 82–89.

11. Bhowmick, S., & Hazarika, S. M. (2016). *Machine learning for E-mail spam filtering: Review, techniques and trends* (arXiv:1606.01042) . arXiv. https://arxiv.org/abs/1606.01042

12. Savage, D., Zhang, X., Yu, X., Chou, P., & Wang, Q. (2014). Anomaly detection in online social networks. *Social Networks*, *39*, 62–70.

13. Mostard, W., Zijlema, B., & Wiering, M. (2019). Combining visual and contextual information for fraudulent online store classification. In *Proceedings of the International Conference* (pp. 84–90). https://doi.org/10.1145/3350546.3352504

14. Beltzung, L., Lindley, A., Dinica, O., Hermann, N., & Lindner, R. (2020). Real-time detection of fake-shops through machine learning. In *2020 IEEE International Conference on Big Data* (pp. 2254–2263). https://doi.org/10.1109/BigData50022.2020.9378204

15. Maktabar, M., Zainal, A., Maarof, M. A., & Kassim, M. N. (2018). Content based fraudulent website detection using supervised machine learning techniques. *Advances in Intelligent Systems and Computing*, *734*, 294–304. https://doi.org/10.1007/978-3-319-76351-4_30

16. Khoo, E., Zainal, A., Ariffin, N., Kassim, M. N., Maarof, M. A., & Bakhtiari, M. (2021). Fraudulent e-commerce website detection model using HTML, text and image features. *Advances in Intelligent Systems and Computing*, *1182*, 177–186. https://doi.org/10.1007/978-3-030-49345-5_19

17. Wu, K., Chou, S., Chen, S., Tsai, C., & Yuan, S. (2018). Application of machine learning to identify counterfeit websites. In *Proceedings of the International Conference* (pp. 321–324). https://doi.org/10.1145/3282373.3282407

18. Wadleigh, J., Drew, J., & Moore, T. (2015). The e-commerce market for “lemons”: Identification and analysis of websites selling counterfeit goods. In *Proceedings of the*

24th International Conference on World Wide Web (pp. 1188–1197). https://doi.org/10.1145/2736277.2741677