ISSN:.....Print

Journal of Computing, Science &Technology
https://focjournal.unidel.edu.ng/
editorjcst@unidel.edu.ng

ISSN: ..... Online

Detection of Cues in Malicious Web-Content using a Sentiment-targeted Extreme Gradient Boosting Tree-based Ensemble

Rume Elizabeth Yoro[1], Okpako Abugor Ejaita[2], & Edun Ogechi Peace[3]
[1,3]Department of Ccybersecurity, Dennis Osadebey University, Asaba, Delta State, Nigeria
Department of Cybersecurity, University of Delta, Agbor, Delta State, Nigeria
Email: [1]elizabeth.yoro@dou.edu.ng, [2]ejaita.okpako@unidel.edu.ng, [3]ogechi.edun@dou.edu.ng
Corresponding Author's Email: elizabeth.yoro@dou.edu.ng

Article Info

Date Received: 17-03-2024
Date Accepted: 13-05-2024

Keywords:

Cyber security,Fraud
Personality traits
Phishing, Spam
Transactions

ABSTRACT

Mobility, ease of accessibility, and portability have continued to grant ease in the adoption rise of smartphones; while, also proliferating the vulnerability of users that are often susceptible to phishing. With some users classified to be more susceptible than others resulting from media presence and personality traits, many studies seek to unveil lures and cues as employed by these attacks that make them more successful. Web content has been often classified as genuine and malicious. Our study seeks to effectively identify cues and lures using the sentiment analysis targeted tree-based gradient boosting algorithm on dataset divided into train/test sets that are scraped from client/user online presence and activity over social networking sites. The dataset is scraped using the Python Google Scrapper. The essence of which is to effectively help users to classify contents from social networking sites as either malicious phishing attacks, or as genuine contents for use using sentiment analysis. The machine learning of choice is the XGBoost. Results show that the ensemble yields a prediction accuracy of 97-percent with an F1-score of 98.19% that effectively correctly classified 2089-instances with 85-incorrectly classified instances for the test-dataset.

1.0 INTRODUCTION

With the Internet advancing as an efficient and effective means of data sharing and dissemination, many adversaries have since begun to use the medium as a tool for the propagation of malicious content [1]. Access to malicious content online [2] has since become a multi-billion-dollar challenge that plagues a variety of users daily [3]. Despite the plethora of continued studies that sought to improve detection via filtering and classification schemes, users continue to fall prey to the scam [4]. This can be attributed to the fact that websites are rippled with malware that presents themselves as unsolicited insecure adverts and/or hides in third-party legitimate software [5]. Various researchers have begun to investigate how various aspects of psychology seek to compromise data – even with a plethora of cyber-security measures in place [6]. One such concern is how the Internet is gradually replacing normal social activities as users now engage themselves with web content – as tools to compensate for their inherent loneliness and social seclusion [7]–[9]. Digital transformation seeks to integrate informatics and its enabling technologies into every facet of our society[10]. Thus, changing our mode of service delivery to clients in lieu of the value they get from the services rendered [11]. It is also a cultural change that requires organizations to continually challenge the status quo [12], experiment [13], and get comfortable with failure [14]. Thus, as more individuals become connected to the Internet via enabling, support devices [15] – it consequently, also opens up many of such persons on a larger scale, to avenues of exploitation that can be harnessed by adversaries via socially-engineered threats and attacks [16].The socially-engineered attack is an old paradigm that continues to steadily grow, with no end in sight. Its continued growth hinges on the human trust instincts and insatiable wants that an attacker exploits to steal the data of a compromised user [17]–[19].

Socially-engineered attacks use technical subterfuge to defraud an unsuspecting victim of their data by posing as a trusted identity [20], [21]. Common methods employed by these adversaries (and not limited to) includes phishing, pharming, spamming, vishing, etc. This provides the attacker with an attractive entry point of contact with a compromised victim's device as well as becomes a pilot/pivot point for attack spread [22], [23]. With such attacks targeted at Internet-connected user devices as well as with over 200 percent adoption of smartphones, many users have become vulnerable and compromised victims [24] – alongside complications of work/business issues on the exposure of sensitive user-data to [25].

Phishing often employs multiple means such as spoofed emails, weblink forgeries, phone calls, man-in-middle chat, covert redirect, etc – to convince a user to divulge confidential data or indulge in fraudulent transactions [26], [27]. An effective, favored variant is spear phishing, which uses targeted mail with access links to cleverly persuade potential victims, and redirect them to spoofed malicious web contents containing malware that aims to compromise user data. Its variant (SMS-phishing) tricks a user into downloading the malware onto a user's device [28]. Phishing basically, redirects user traffic to a fake site, by either

changing the host's file on a victim's device or by exploiting the vulnerability in the domain name service server software. Thus, it allows an adversary to install malware on a user's device and redirects the user to a fraudulent site without their consent or knowledge [29]–[31]. Phishing involves an attacker redirecting a user's access to malicious content shared from spoofed websites from a viewpoint that such sites are legitimate and trustworthy sources. Typical phishing threat and/or attack consists of 3-elements namely explained as thus [31]–[33]: (a) lure message is received by the potential victim as originating from a legitimate source. Its reliability is strengthened by exploiting user curiosity, fear, and empathy, (b) hook is a compromised link or attachment included in the message, and (c) the catch involves an attacker obtaining the user's private data.

This may appear or seem simple enough; But, the technique/procedure(s) constantly evolve, to reflect new social trends, that use new methods to bypass security, and evade detection. Its continued spread allowed attacks to vary in frequency and diversity, enhancing their likelihood of success [34]. Thus, phishing is often positioned as trusted entities to defraud a victim (via mail, SMS, etc). Its characteristics include: (a) the message often makes unrealistic demands via various forms of intimidation targeted at a user's psych, (b) there is always a catch, (c) there is often missing data with spelling errors and poor grammars, (d) there is often a mismatch in URL (uniform resource locator) to redirect users to a faked website, and (e) messages often demands sensitive, confidential user data [35], [36]. Umarani et al., [37] used victimization features to characterize the design impact of websites on both the structure of the content and the probability content will victimize a user. They used 2-feats to help users understand and identify malicious contents, and eliminate the awareness gaps: (a) believability to identify cue sophistication which increases the possibility a user will believe a message, and (b) insidiousness to measure the potency in degradation lures and its success rate while remaining undetectable to users.

Ezpeleta et al., [38] investigated spam attacks with millions of malicious files sent daily via spam. They posited that for many users – it is about control rather than an issue of prevention and mitigation of spam via filters and other schemes as technical measures. Also, the users' level of suspicion, emotional control [39], [40], and attack awareness must become a critical component in either the success or failure of an attack. This is because – emotion becomes personality traits and behavior that culminates as cues/factors that drive the desire to help, to seek gain via exploitation, and to be liked. These, all suggestions make some persons more susceptible to attacks [41], [42]; And such victims, may fall repeatedly into a scam.

1.1. Machine Learning Approaches: Review of Literatures

The rise in phishing attack cases has raised concerns, making phishing detection a crucial and urgent task for businesses. Its adoption in cyber-fraud can be grouped into the following classes: (a) the outright theft of user personal details and information, (b) the theft of confidential details via malware intrusive means, and (c) surreptitiously attainment during an online transaction without the compromised user's awareness [43], [44]. The loss in cost associated with card fraud has since become staggering, with the payment card industry consequently, incurring losses in billions of dollars annually. Users and businesses must remain committed and vigilant towards the continued improvements with phishing detection and prevention systems. Though, despite these efforts, adversaries continue to invent new techniques to circumvent these security measures as well as avoid detection, making it a constant battle [45], [46]. To curb and minimize the effect therein in the society of phishing attacks over web-content, machine learning approaches have been successfully trained and adapted to effectively recognize phishing patterns within web-contents as cues and lures. These, they learn through features classification either from the normal behavior cum signature in transactions, or the quick detection of an unusual activity in the transaction pattern indicative of fraudulent profile.A variety of such machine learning (ML) models that have been successfully used or implemented includes: Logistic Regression [47]–[49], Deep Learning [50], [51], Bayesian model [52], Naive Bayes [53], Support Vector Machine [54], Random Forest [55], and other models [56] that have been effectively used to detect credit-card fraud. Many of these, have drawbacks with their flexibility in feature selection, importance, and accuracy. Our study adopts extreme gradient boost tree-based ensemble. This choice is due to its capability to reduce overfitting, address imbalanced datasets, and yield a vigorous accuracy [57]–[59]. Table 1 is a list of contributions of machine learning approaches to phishing schemes so far:

Table 1: Related Literatures Contributions

Authors	Efficient Selected Algorithms/Heuristics	Accuracy
Akazue et al. [60]	Hybrid feature selection technique using info gain, chi-square and recursive elimination with Random Forest Tree algorithm	95.83%
Ojugo et al. [61]	Deep learning modular memetic algorithm	99.6%
Btoush et al. [62]	Deep Learning	95.76%
Roseline et al. [63]	Long Term Short Memory (LSTM)	99.58%
Sinayobye et al. [64]	KNN, LR, SVM, DT and RF	82.60%
Ali et al. [65]	LR, KNN, SVM, PCA, QDA, ANN	98.45%
Rytali and Enncya [66]	LR, LSTM, XGBoost	97.23%

The inherent gaps in previous studies includes thus [67]–[71].

1. **Lack of Datasets:** Finding the right-format dataset – is crucial to machine learning task. Access to high-quality datasets is needed in training and performance evaluation [72] – as there is limited data, which often yield significant false positives [73].
2. **Imbalanced Datasets:** A critical hurdle is the challenge with imbalanced datasets with cases of phishing lagging behind. Studies must seek explore intricate sampling techniques, or harness the robust power of ensemble methods tailored explicitly to mitigating the challenges with imbalanced dataset [74], [75].
3. **Cross-Channel Detection:** With increased multiple channel [76]–[78] – newer models must integrate the varying channel data to enhance the overall accuracy. Cross-channel phishing detection has now become a critical area of research and business focus [79] as traditional phishing detection modes are limited in adapting then emergent fraud patterns as well as keeping up with novel tactics.

1.2. Tree-Based Algorithms and Ensembles

A very common approach in ML are tree-based methods which descend from single decision trees [80]. Adopting a tree-structure, each tree generates a series of if-else rules used in majority voting scheme that allows it to predict observed classes [81]. In classification/regression tasks, each tree is a recursive top-down model in which a binary tree partitions a predictor space with variables grouped into subsets for which the distribution of dependent variable y is successively more homogeneous [82]. Each decision tree has the merit of being easily understood [83]; But, its use alone often leads to model overfit in a prediction task as the model seeks to learn and identify features of interest during training [84]. Thus, it yields degraded performance in its classifying of unknown labels [85]. These drawbacks have birthed ensembles with improved predictive norms and are more expressive [86]. Tree-based ensembles learn by constructing many individually trained decision trees [87], and combines/aggregates their results into a single and stronger model, whose output outperforms the results of any single tree [88]. It achieves this via bagging [89]–[91] and/or boosting [92]–[94] approaches.

In the case of boosting – the tree(s) converts weak learners (i.e., achieve accuracy just above random guess) onto a strong learner with enhanced predictive capacity by sequentially training each weak learner to correct the inherent weaknesses of its predecessor[95]–[97]. Each tree yields a feedback from previous trees [98], [99]. Popular boosting ensembles include adaptive boosting (AdaBoost) [94], gradient boost (GB) [100], boosted logistic regression (LogitBoost) [101], and stochastic gradient boosting (SGB) [102]. They are expressed as Equation 1 – to yield its prediction by combining outcome of its weak learners with its weighted sum to yield a higher weight for incorrectly classified instances as thus:

$$L^t = \sum_{i=1}^{n} l\left(Y_i^t, \widehat{Y}_i^{t-1} + f_k(x_i)\right) + \Omega(f_t) \quad (1)$$

Conversely, bagging grow successive trees independently from earlier trees – such that each tree is constructed using a bootstrap aggregation mode to sample the data using majority vote during its prediction [103]. The Random Forest add extra layer of randomness to the bagging scheme, which in turn – changes how the trees constructed. While, standard decision trees has that each node is split using the best split among all predictor variables – the Random Forests allows its nodes to be split using the best among a subset of predictors randomly chosen at that node [104]. Its recursive structure helps it to capture interaction effects between the variables [8], [46], [105].

In all, tree-based ensembles have successfully proven to be better than other established approaches across a variety of different tasks [106] ranging from traffic flow classification [107], customer churn prediction [108], and prediction of online purchase intention [109]. They have been known to be suited to reduce both bias and variance in single learning schemes. While individual models may get stuck in local minima [110], a weighted combination of several different local minima – produced by ensemble methods [111] – are able to minimize the risk of choosing the wrong local minimum [112].

2.0: MATERIALS AND METHODS

2.1. Data Gathering / Sample Demographics

Data were collected using Google Play Scraper Library for Python, and a total of 8,693 records were gathered for the period June to December 2022. The scrapped records consist of personal data, compromised contents (links, images, and texts), user emails, posts, likes and shares, and replies as suggested by [113], [114].

2.2. The Detailed Proposed Experimental Ensemble

The access ease in web connectivity by many users has continued to see a rise in data shared between various users. With such popularity especially with the birth of smartphones, phishing attacks have been on the rise with lessened user trust in shared data [86], [115], [116]. Generally, a user's opinion of an idea or topic of interest is simply his/her belief, centered on his perception or feeling towards the issue at hand. The belief and opinions represent the user's disposition of emotion. This emotion correlates with his/her behavior concerning the situation and is referred to as sentiment. Thus, sentiment analysis deals with a class of language that seeks to trace and track a user's or community's behavior towards a topic of interest [37], [117]. In natural language processing – its data is often unstructured and thus, rippled with ambiguities, noise, and imprecisions. The proposed ensemble is herein seen in Figure 1.

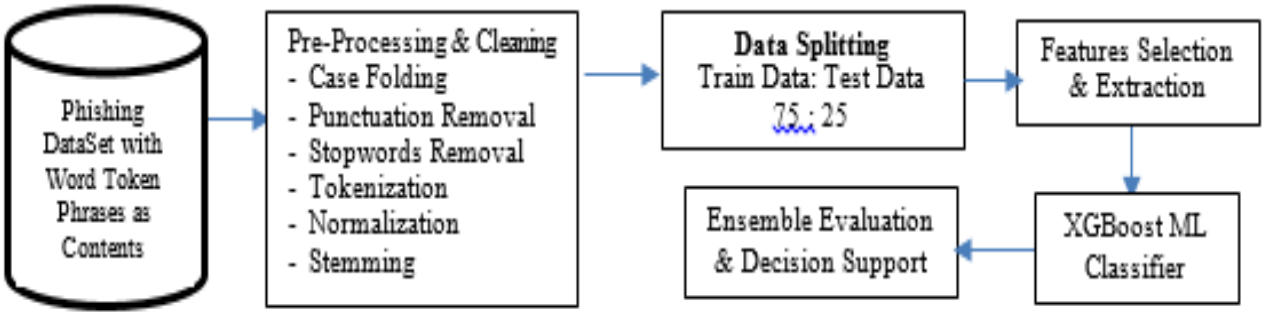

Figure 1. Sentiment Analysis Process with Decision Support System

The sentiment analysis process is carried out in several steps[41], [118]–[121] including:

1. **Step 1: Data Collection/Cleaning**involves the processes of collecting data and performing data cleaning as well as preprocessing. Once the data is collected, it is then pre-processed/cleaned. This step implies the dataset is restructured from its natural unstructured state(s), normalized, removal of some phrases or stopwords, tokenization, and word stemming. The steps are also explained thus [14], [88], [122], [123]:

- **Case Folding** attempts to convert all letters, strings, and concatenated word tokens into lowercase or uppercase. It does this to avoid two-or-more word tokens ending up with the same meaning but being treated differently by the machine due to writing in different forms; lowercase and uppercase.
- **Punctuation removal** is simply the removal of all symbols of punctuation marks from word phrases and tokens. We note also that punctuation mark in natural language processing (NLP) does not add extra information. It, however, reduces the dimensionality in our dataset, to be resolved.
- **Stopword removal** seeks to remove some common tokens/words across all the documents. Stopwords like punctuations, do not add much information to the scenario; And, their removal also only reduces the dimensionality in our dataset to be resolved. We also note that pronouns, articles, conjunctions, and prepositions – are often and are, classified as stopwords.
- **Tokenization** simply breaks down a sentence into smaller elements. Tokens help us to interpret the implicit meaning of a sentence by analyzing their order of placement in the text corpus. Thus, they act as input for various NLPs algorithms – with normalized texts that are broken into individual word elements, stopwords, and punctuation characters (that are equally removed in this unit) [124], [125].
- **Normalization** is converting/expanding a token/word/slang back to its original form. This process removes abridged versions of a token (slang/word) from a text to preserve its basic form. It also expands abbreviations into their complete forms. E.g. the term "notin" is changed to "nothing"; while, the word "welccoomme" is transformed to its base "welcome." For normalization, we used the abbreviation dictionary by Ojugo and Eboka [41] and Afifah et al. [119].
- **Word Stemming** is a step to remove affixes in a word, both affixes that appear before and after the word. Stemming converts each word to its root word without affixes.

2. **Step 2: Feats Extraction and Selection** – is the second step. It involves the extraction of the underlying features to be considered as well as formatting these features into parameters of estimation to aid the effective and efficient classification of the texts. Thus, this stage helps the framework ensemble to select the most relevant text features from an expert perspective and/or opinion. Feature extraction and feature selection methods are used for this purpose. These methods are used to reduce the number of input variables, avoid overfitting, decrease computational complexity or training time, and improve model accuracy. Vectorization and word embedding methods are used for feature extraction [118], [119].

- **Term Frequency Inverse Document Frequency (TF-IDF)** – In exploring many machine learning heuristics, which seek to discover the relative probability scores of parameters of interest – to yield optimal solutions – we note that these algorithms do not understand characters cum word tokens. But, they very well understand and accept as input numbers. With feature selection and extraction, since it is impossible for the textual nature of the dataset to interact directly with the machine and/or algorithm – we employ vectorization methods. For this study, we use the TF-IDF which seeks to compute the frequency of the occurrence of certain word tokens in a document – so that the more a word appears – the greater its TF-value; while IDF simply aggregates the weight of the words against its appearances throughout the document. Conversely, the more certain words appear, the smaller its IDF-value with the transposed TF-IDF computed as in Equation 2 and 3 respectively:

$$IDF = \log\left(\frac{N}{DF}\right) \quad (2)$$

$$TF - IDF(d,k) = TF(d,k) * IDF(k) \quad (3)$$

3. **Step 3: Machine Learning Heuristic** – We implement the proposed heuristics to help effectively classify or categorize text as positive, negative, or neutral (words) based on the sentiment polarity of opinions. Machine learning techniques classify the sentiments based on training and test dataset [126].

- **Extreme Boosting (XGBoost)** is a decision tree ensemble that leverages Gradient Boosting and is designed to be scalable. By combining weak learners, a Gradient Boost becomes stronger via iteration to find an approximate fit. It achieves this via an additional expansion to its objective function by minimizing the loss function (creating a variation) used to control the trees' complexity. XGBoost offers a better optimal solution by combining the predictive power of multiple weak base learners. Each learner contributes data about the task to the ensemble and is used for prediction – enabling the XGboost to yield a stronger learner [127]. Given the training sub-dataset to be trained x_i and its corresponding labels y_i – XGBoost predicts the optimal outcome using Equation 4:

$$\hat{Y}_i^t = \sum_{k=1}^{t} f_k(x_i) = \hat{Y}_i^t + f_k(x_i) \qquad (4)$$

For a better outcome, the XGBoost minimizes its objective function, which contains the loss function $l(Y_i^t, \hat{Y}_i^t)$ and its regularization term $\Omega(f_t)$. The loss function ensures that model overtraining does not occur and that the training data are fitted well enough to the model; while, the regularization term ensures the complexity fitness of the trees. Tuning the loss function ensures the model yields higher accuracy; while, tuning the regularization terms ensures a generalized simpler ensemble as well as helps the ensemble avoid model parameter overfitting as in Equation 5 [93], [127].

$$L^t = \sum_{i=1}^{n} l\left(Y_i^t, \hat{Y}_i^{t-1} + f_k(x_i)\right) + \Omega(f_t) \qquad (5)$$

- **Hyper-Parameter Tuning** controls how much of the tree complexity and its corresponding nodal weights need to be adjusted in place of gradient loss. The lower the value, the slower we travel on a downward slope. It also ensures how quickly a tree abandons old beliefs for new ones during the training. Thus, as the tree learns – it quickly begins to differentiate between important features cum parameters, and otherwise. A higher learning rate implies that the tree can change, learn newer features as well as adapts flexibly, and more easily. The ensemble uses the regularization term to ensure the model changes quickly, only to values that are within the lower and upper bounds. The ensemble does this to ensure that it adequately adjusts its learning rate to avoid over-fitting and overtraining. Other feats that can be adjusted include max_depth, sub_sample, and n_estimators. The n_estimators indicate the number of decision trees in XGBoost, which when set as 1 – will make the algorithm generate only a single tree. For best performance, the XGBoost ensemble must carefully tune these parameters [126].
- **Cross-Validation/Retraining** is an applied ML scheme that estimates the learned skills of a heuristic technique on unseen data. It is a procedure also, that seeks to evaluate the model's performance about its accuracy on how well it has learned the underlying features of interest via the resampling technique. Thus, in cross-validation – the modelers choose several data folds (or partitions) which helps the model ensure it is devoid of overfitting. Here, we use stratified k-fold (it rearranges the data to ensure that each fold is a good representation of the entire dataset) [128] as in algorithm listing 1.

Algorithm 1: Stratified k-fold cross-validation

shuffle the dataset

split or partition training dataset into k-folds

For k-iterations

    re-arrange data in partition: return k-folds = true

    return **k-folds = true**

end for

evaluate model

End

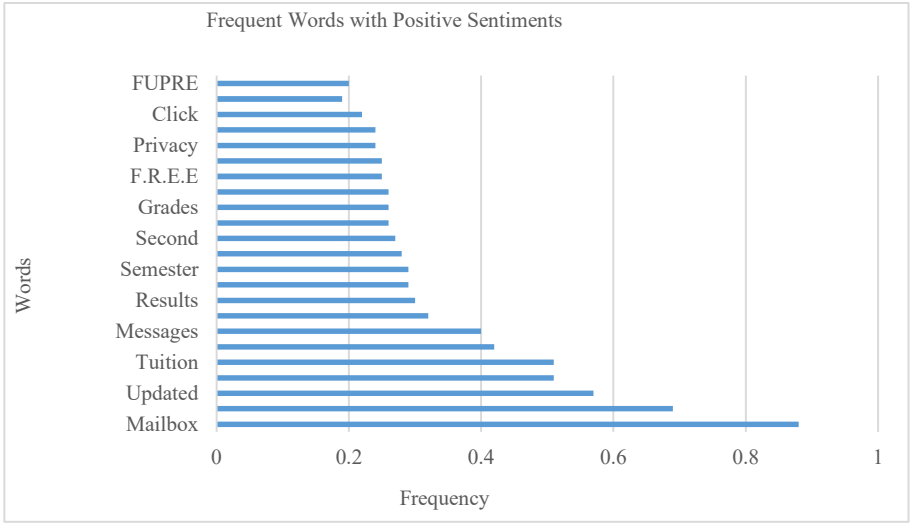

Figure 2: Frequency Chart of Positive Sentiment Words

3. 0 FINDINGS AND DISCUSSION

3.1. Findings and Discussion

We explain how we performed data preprocessing, feature extraction, model training, and evaluation.

3.1.1. Data Preprocessing

We have already applied the preprocessing steps to the dataset as earlier mentioned in the previous section. We then visualize the data with positive sentiments. Datasets of such nature can be used to (a) rank images, (b) for natural language assessment using ML schemes, and (c) mine user assessment via the use of metadata to infer semblance, company characteristics, and feelings. Thus, in general – the dataset objective is the mining of the dataset to seek the relationship of various forms through the use of these effective cues and lures [129]. These will attempt to seduce the ensemble to navigate compromised links, images, and other embedded objects. This is as seen in figure 2.

3.1.2. Training Phase

Here, we partition the retrieved dataset into 75 percent trainingdata, and 25 percent test data. For the training dataset, we used 6,520 rows, and a test dataset of 2,173 rows. We then perform feature extraction using the TF-IDF vectorization method – which helps the ensemble to effectively convert our retrieved text contents into vectors. Furthermore, we use Python's ScikitLearn ***TfidfVectorizer*** function to extract the desired features of interest – as defined in our ensemble. We then train the model using our train dataset. It is also worthy of note that we employ the trial-and-error method to tune the hyper-parameters and find the weight that yields the optimal solution. This is aimed at improving the ensemble's fitness and protects the ensemble from model overtraining and overfitting of parameters as in Table 2.

Table 2. Hyper-Parameters Trial-n-Error (and Best) Values

Hyper-Parameters	Definition	Trial-and-error values	Best Values
Learning Rate	Step-size for learning	[0.05, 0.1, 0.2, 0.3, 0.5, 0.75]	0.2
N_Estimators	Number of trees in the ensemble	[100, 200, 300, 500, 700, 800]	500
Max-Depths	Max. number of trees depth	[1, 2, 4, 5, 6, 8, 10]	6

Using the trial-n-error mode for the hyper-parameters, we observe during the training phase, that the best-fit values for learning_rate, n_estimators, and max_depth are 0.2, 500, and 6 respectively.

3.1.3. XG-Boost Classifier Evaluation and Discussion

We evaluate the sentiments of the words scrapped as shown in Table 3:

The ensemble yields the metrics in Table 3 – which notes that the cues and lures for the negative sentiments were detected and effectively classified with an 87-percent accuracy (0.87); while the cues and lures for positive sentiments were also detected with a prediction accuracy of 97-percent (0.97). Such disparities in the accuracy of prediction may have been expected and are normal – due to errors of false-positives, true-negatives, false-negatives, and true-negatives in agreement with [130]–[132].

3.1.4. Stratified K-Fold Retraining Evaluation and Discussion

The stratified k-fold retraining yields Table 4.

Table 4. Stratified k-fold Evaluation Metrics

Iteration	1	2	3	4	5	6	7	8	9	10	11	12
F1	0.972	0.981	0.979	0.978	0.983	0.991	0.989	0.984	0.990	0.989	0.986	0.987

The ensemble during the retraining or cross-validation phase – over a series of iterations (movement) yields an accuracy prediction of 99.1 percent (i.e. 0.991) in detecting the cues/lures for both sophistication and degradation of positive and negative sentiments.

3.2. Ensemble Evaluation and Performance

To compute the sensitivity, specificity, and accuracy of the ensemble – we evaluate its performance using Eq. 5 to yield Figure 3 as thus:

$$F1-scores\ or\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \rightarrow 2\,x\,\frac{Precision\ x\ Recall}{Precision+Recall} \quad (5)$$

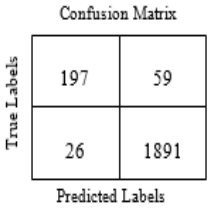

Figure 3. Confusion Matrix Evaluation Metrics

Figure 2 confusion matrix clearly shows the proposed XGBoost ensemble efficiently and correctly classified 2089-instances of the test-dataset; and incorrectly classified 89-instances as compared to the studies by [133], [134]. Underlying each user interface is often a *trust decision box* that lets a user either trust (*blue for accept*) or not trust (*red for reject*) using content-specific decisions.

3.3. Discussion of Findings

Table 3. XGBoost Classifier Evaluation Metrics

Sentiments	Precision	Recall	F1-Score
Negative	0.81	0.82	0.87
Positive	0.97	0.98	0.97

Table 5 shows sample sophistication and degradation (cues to detecting malicious content embedded

Table 5. Website Malicious Contents for Sample Lures / Cues

ID	Sophistication Lures and Cues	ID	Degradation Lures and Cues
S01	Legitimate logos on the website	D01	Suspicious URL identifies Site or Sender
S02	Duplicates look/feel of a legitimate website	D02	Poor spelling and/or grammar issues
S03	Provides contextual or personal data	D03	Includes suspicious attachments in email
S04	Legitimate links hiding malicious data	D04	Contains unnecessary warning messages
S05	Provides a sense of previous trust	D05	Directly requests the input of personal data
S06	No typos in Grammar and Writing style	D06	References item prices too good to be true
S07	Uses official account usernames	D07	Missing security designators, e.g. https padlock
S08	Identifies a known group of recipients	D08	Appeals to emotion, e.g., urgency and greed
S09	Recognizes file types as attached	D09	Unrecognized file types attached

Table 5 lists sophistication cues that make content harder to identify. Sample list is described below [75], [135]:

in posts, photo likes, and shares, targeted emails, etc.

S06: Free Grammar and Style in Writing: Uses generic greetings instead of receiver names

→*Context-Language-Tone-Professional*

S09: Unrecognize file types as downloads/attachments: File extension is unknown

→*Content-URL Links-Obfuscated*

Sample degradations, which may be known to a user are listed below [3], [19]:

D06: Information or item prices are too good to be true

→*Contract-Offer-Monetary-Products*

D08: Appeals to an emotion such as urgency and greed

→*Context-Language-Tone-Professional*

The goal of the experiment is to understand how users make trust decisions, identify their deficiencies, and adapt training/awareness capabilities to prevent victimization of the user and the associated organization where they work (in this case, the university) which agrees. The experiment is presented as a mixture of malicious and normal content to simulate real-time interactions with an email client, web browser, and social network [136], [137]. The experiment follows a scene where a participant must respond to phishing and malicious insider tactics to keep them quite interested and engaged online (increased online presence). Simulation provide the participants with rich interaction capabilities that allow them to hover over links and attachments and see natural browser-like behavior [47], [138].

4.0 CONCLUSION

Social media networks have safeguards and rules to educate users as well as protect them against phishing attempts. These often involve the capability to investigate and blacklist phishers if such cases are reported. Both the media and users are held accountable for preventing phishing attacks and their awareness. The social media platform is in charge of informing users about phishing and giving controls to prevent them. Conversely, users must stay ahead of the curve with and about preventing these attacks as well as implementing safety controls to limit such accidents.

REFERENCES

[1] Y. Zhang, S. Egelman, L. F. Cranor, and J. Hong, "Phinding Phish: Evaluating Anti-Phishing Tools," *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS 2007)*, no. March, pp. 1–16, 2007.

[2] E. D. Ananga, "Gender Responsive Pedagogy for Teaching and Learning: The Practice in Ghana's Initial Teacher Education Programme," *Creat. Educ.*, vol. 12, no. 04, pp. 848–864, 2021, doi: 10.4236/ce.2021.124061.

[3] M. I. Akazue, A. A. Ojugo, R. E. Yoro, B. O. Malasowe, and O. Nwankwo, "Empirical evidence of phishing menace among undergraduate smartphone users in selected universities in Nigeria," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 28, no. 3, pp. 1756–1765, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1756-1765.

[4] M. Callen, C. C. Gibson, D. F. Jung, and J. D. Long, "Improving Electoral Integrity with Information and Communications Technology," *J. Exp. Polit. Sci.*, vol. 3, no. 1, pp. 4–17, Oct. 2016, doi: 10.1017/XPS.2015.14.

[5] R. E. Yoro, F. O. Aghware, B. O. Malasowe, O. Nwankwo, and A. A. Ojugo, "Assessing contributor

features to phishing susceptibility amongst students of petroleum resources varsity in Nigeria," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 2, p. 1922, Apr. 2023, doi: 10.11591/ijece.v13i2.pp1922-1931.

[6] E. Altman, "Synthesizing credit card transactions," in *Proceedings of the Second ACM International Conference on AI in Finance*, New York, NY, USA: ACM, Nov. 2021, pp. 1–9. doi: 10.1145/3490354.3494378.

[7] D. Nallaperuma *et al.*, "Online Incremental Machine Learning Platform for Big Data-Driven Smart Traffic Management," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4679–4690, 2019, doi: 10.1109/TITS.2019.2924883.

[8] A. A. Ojugo, C. O. Obruche, and A. O. Eboka, "Quest For Convergence Solution Using Hybrid Genetic Algorithm Trained Neural Network Model For Metamorphic Malware Detection," *ARRUS J. Eng. Technol.*, vol. 2, no. 1, pp. 12–23, Nov. 2021, doi: 10.35877/jetech613.

[9] A. A. Ojugo, M. I. Akazue, P. O. Ejeh, C. Odiakaose, and F. U. Emordi, "DeGATraMoNN: Deep Learning Memetic Ensemble to Detect Spam Threats via a Content-Based Processing," *Kongzhi yu Juece/Control Decis.*, vol. 38, no. 01, pp. 667–678, 2023.

[10] A. O. Eboka and A. A. Ojugo, "Mitigating technical challenges via redesigning campus network for greater efficiency, scalability and robustness: A logical view," *Int. J. Mod. Educ. Comput. Sci.*, vol. 12, no. 6, pp. 29–45, 2020, doi: 10.5815/ijmecs.2020.06.03.

[11] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram, "The design of phishing studies: Challenges for researchers," *Comput. Secur.*, vol. 52, pp. 194–206, Jul. 2015, doi: 10.1016/j.cose.2015.02.008.

[12] A. A. Ojugo and D. O. Otakore, "Redesigning Academic Website for Better Visibility and Footprint: A Case of the Federal University of Petroleum Resources Effurun Website," *Netw. Commun. Technol.*, vol. 3, no. 1, p. 33, Jul. 2018, doi: 10.5539/nct.v3n1p33.

[13] T. Sahmoud and D. M. Mikki, "Spam Detection Using BERT," *Front. Soc. Sci. Technol.*, vol. 14, no. 2, pp. 23–35, Jun. 2022, doi: 10.48550/arXiv.2206.02443.

[14] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "A review of spam email detection: analysis of spammer strategies and the dataset shift problem," *Artif. Intell. Rev.*, May 2022, doi: 10.1007/s10462-022-10195-4.

[15] A. A. Ojugo and A. O. Eboka, "A Social Engineering Detection Model for the Mobile Smartphone Clients," *African J. Comput. ICT*, vol. 7, no. 3, pp. 91–100, 2014, [Online]. Available: www.ajocict.net

[16] A. A. Ojugo and R. E. Yoro, "Predicting Futures Price And Contract Portfolios Using The ARIMA Model: A Case of Nigeria's Bonny Light and Forcados," *Quant. Econ. Manag. Stud.*, vol. 1, no. 4, pp. 237–248, 2020, doi: 10.35877/454ri.qems139.

[17] D. Huang, Y. Lin, Z. Weng, and J. Xiong, "Decision Analysis and Prediction Based on Credit Card Fraud Data," in *The 2nd European Symposium on Computer and Communications*, New York, NY, USA, NY, USA: ACM, Apr. 2021, pp. 20–26. doi: 10.1145/3478301.3478305.

[18] Y. Lucas *et al.*, "Multiple perspectives HMM-based feature engineering for credit card fraud detection," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, New York, NY, USA: ACM, Apr. 2019, pp. 1359–1361. doi: 10.1145/3297280.3297586.

[19] R. E. Yoro, F. O. Aghware, M. I. Akazue, A. E. Ibor, and A. A. Ojugo, "Evidence of personality traits on phishing attack menace among selected university undergraduates in Nigerian," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 2, p. 1943, Apr. 2023, doi: 10.11591/ijece.v13i2.pp1943-1953.

[20] M. I. Akazue, R. E. Yoro, B. O. Malasowe, O. Nwankwo, and A. A. Ojugo, "Improved services traceability and management of a food value chain using block-chain network : a case of Nigeria," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 29, no. 3, pp. 1623–1633, 2023, doi: 10.11591/ijeecs.v29.i3.pp1623-1633.

[21] C. L. Rash and S. M. Gainsbury, "Disconnect between intentions and outcomes: A comparison of regretted text and photo social networking site posts," *Hum. Behav. Emerg. Technol.*, vol. 1, no. 3, pp. 229–239, Jul. 2019, doi: 10.1002/hbe2.165.

[22] I. A. Anderson and W. Wood, "Habits and the electronic herd: The psychology behind social media's successes and failures," *Consum. Psychol. Rev.*, vol. 4, no. 1, pp. 83–99, Jan. 2021, doi: 10.1002/arcp.1063.

[23] Y. Kang, M. Ozdogan, X. Zhu, Z. Ye, C. Hain, and M. Anderson, "Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest," *Environ. Res. Lett.*, vol. 15, no. 6, p. 064005, Jun. 2020, doi: 10.1088/1748-9326/ab7df9.

[24] S. M. Albladi and G. R. S. Weir, "User characteristics that influence judgment of social engineering attacks in social networks," *Human-centric Comput. Inf. Sci.*, vol. 8, no. 1, p. 5, Dec. 2018, doi: 10.1186/s13673-018-0128-7.

[25] A. A. Ojugo and A. O. Eboka, "Inventory prediction and management in Nigeria using market basket analysis associative rule mining: memetic algorithm based approach," *Int. J. Informatics Commun. Technol.*, vol. 8, no. 3, p. 128, 2019, doi: 10.11591/ijict.v8i3.pp128-138.

[26] J. Jung, M. Maeda, A. Chang, M. Bhandari, A. Ashapure, and J. Landivar-Bowles, "The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems," *Curr. Opin. Biotechnol.*, vol. 70, pp. 15–

22, Aug. 2021, doi: 10.1016/j.copbio.2020.09.003.
[27] M. Gratian, S. Bandi, M. Cukier, J. Dykstra, and A. Ginther, "Correlating human traits and cyber security behavior intentions," *Comput. Secur.*, vol. 73, pp. 345–358, Mar. 2018, doi: 10.1016/j.cose.2017.11.015.
[28] H. Tingfei, C. Guangquan, and H. Kuihua, "Using Variational Auto Encoding in Credit Card Fraud Detection," *IEEE Access*, vol. 8, pp. 149841–149853, 2020, doi: 10.1109/ACCESS.2020.3015600.
[29] W. Rocha Flores, H. Holm, M. Nohlberg, and M. Ekstedt, "Investigating personal determinants of phishing and the effect of national culture," *Inf. Comput. Secur.*, vol. 23, no. 2, pp. 178–199, Jun. 2015, doi: 10.1108/ICS-05-2014-0029.
[30] M. Barlaud, A. Chambolle, and J.-B. Caillau, "Robust supervised classification and feature selection using a primal-dual method," Feb. 2019.
[31] G. Sasikala *et al.*, "An Innovative Sensing Machine Learning Technique to Detect Credit Card Frauds in Wireless Communications," *Wirel. Commun. Mob. Comput.*, vol. 2022, pp. 1–12, Jun. 2022, doi: 10.1155/2022/2439205.
[32] M. Laavanya and V. Vijayaraghavan, "Real Time Fake Currency Note Detection using Deep Learning," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1S5, pp. 95–98, 2019, doi: 10.35940/ijeat.a1007.1291s52019.
[33] P. Moodley, D. C. S. Rorke, and E. B. Gueguim Kana, "Development of artificial neural network tools for predicting sugar yields from inorganic salt-based pretreatment of lignocellulosic biomass," *Bioresour. Technol.*, vol. 273, pp. 682–686, Feb. 2019, doi: 10.1016/j.biortech.2018.11.034.
[34] A. Algarni, Y. Xu, and T. Chan, "An empirical study on the susceptibility to social engineering in social networking sites: the case of Facebook," *Eur. J. Inf. Syst.*, vol. 26, no. 6, pp. 661–687, Nov. 2017, doi: 10.1057/s41303-017-0057-y.
[35] A. A. Ojugo and O. D. Otakore, "Improved Early Detection of Gestational Diabetes via Intelligent Classification Models: A Case of the Niger Delta Region in Nigeria," *J. Comput. Sci. Appl.*, vol. 6, no. 2, pp. 82–90, 2018, doi: 10.12691/jcsa-6-2-5.
[36] J. Yao, C. Wang, C. Hu, and X. Huang, "Chinese Spam Detection Using a Hybrid BiGRU-CNN Network with Joint Textual and Phonetic Embedding," *Electronics*, vol. 11, no. 15, p. 2418, Aug. 2022, doi: 10.3390/electronics11152418.
[37] V. Umarani, A. Julian, and J. Deepa, "Sentiment Analysis using various Machine Learning and Deep Learning Techniques," *J. Niger. Soc. Phys. Sci.*, vol. 3, no. 4, pp. 385–394, 2021, doi: 10.46481/jnsps.2021.308.
[38] E. Ezpeleta, I. V. de Mendizabal, J. M. Gómez Hidalgo, and U. Zurutuza, "Novel email spam detection method using sentiment analysis and personality recognition," *Log. J. IGPL*, vol. 28, no. 1, pp. 83–94, 2020, doi: 10.1093/jigpal/jzz073.
[39] A. A. Ojugo and A. O. Eboka, "Modeling the Computational Solution of Market Basket Associative Rule Mining Approaches Using Deep Neural Network," *Digit. Technol.*, vol. 3, no. 1, pp. 1–8, 2018, doi: 10.12691/dt-3-1-1.
[40] A. A. Ojugo *et al.*, "CoSoGMIR: A Social Graph Contagion Diffusion Framework using the Movement-Interaction-Return Technique," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 37–47, 2023, doi: 10.33633/jcta.v1i2.9355.
[41] A. A. Ojugo and A. O. Eboka, "Memetic algorithm for short messaging service spam filter using text normalization and semantic approach," *Int. J. Informatics Commun. Technol.*, vol. 9, no. 1, p. 9, 2020, doi: 10.11591/ijict.v9i1.pp9-18.
[42] E. O. Okonta, U. R. Wemembu, A. A. Ojugo, and D. Ajani, "Deploying Java Platform to Design A Framework of Protective Shield for Anti– Reversing Engineering," *West African J. Ind. Acad. Res.*, vol. 10, no. 1, pp. 50–64, 2014.
[43] A. Borucka, "Logistic regression in modeling and assessment of transport services," *Open Eng.*, vol. 10, no. 1, pp. 26–34, Jan. 2020, doi: 10.1515/eng-2020-0029.
[44] I. Sagdali, N. Sael, F. Benabbou, I. Sadgali, N. Sael, and F. Benabbou, "Performance of machine learning techniques in the detection of financial frauds," *Procedia Comput. Sci.*, vol. 148, pp. 45–54, 2019, doi: 10.1016/j.procs.2019.01.007.
[45] A. A. Ojugo, A. O. Eboka, E. O. Okonta, R. E. Yoro, and F. O. Aghware, "Predicting Behavioural Evolution on a Graph-Based Model," *Adv. Networks*, vol. 3, no. 2, p. 8, 2015, doi: 10.11648/j.net.20150302.11.
[46] A. A. Ojugo, C. O. Obruche, and A. O. Eboka, "Empirical Evaluation for Intelligent Predictive Models in Prediction of Potential Cancer Problematic Cases In Nigeria," *ARRUS J. Math. Appl. Sci.*, vol. 1, no. 2, pp. 110–120, Nov. 2021, doi: 10.35877/mathscience614.
[47] E. Ileberi, Y. Sun, and Z. Wang, "A machine learning based credit card fraud detection using GA algorithm for feature selection," *J. Big Data*, vol. 9, no. 1, p. 24, Dec. 2022, doi: 10.1186/s40537-022-00573-8.
[48] S. V. S. . Lakshimi and S. D. Kavila, "Machine Learning for Credit Card Fraud Detection System," *Int. J. Appl. Eng. Res.*, vol. 15, no. 24, pp. 16819–16824, 2018, doi: 10.1007/978-981-33-6893-4_20.
[49] C. Li, N. Ding, H. Dong, and Y. Zhai, "Application of Credit Card Fraud Detection Based on CS-SVM," *Int. J. Mach. Learn. Comput.*, vol. 11, no. 1, pp. 34–39, 2021, doi: 10.18178/ijmlc.2021.11.1.1011.
[50] I. Benchaji, S. Douzi, B. El Ouahidi, and J. Jaafari, "Enhanced credit card fraud detection based on attention mechanism and LSTM deep model," *J. Big Data*, vol. 8, no. 1, p. 151, Dec. 2021, doi: 10.1186/s40537-021-00541-8.
[51] F. O. Aghware, R. E. Yoro, P. O. Ejeh, C. C.

Odiakaose, F. U. Emordi, and A. A. Ojugo, "DeLClustE: Protecting Users from Credit-Card Fraud Transaction via the Deep-Learning Cluster Ensemble," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, pp. 94–100, 2023, doi: 10.14569/IJACSA.2023.0140610.

[52] L. E. Mukhanov, "Using bayesian belief networks for credit card fraud detection," *Proc. IASTED Int. Conf. Artif. Intell. Appl. AIA 2008*, no. February 2008, pp. 221–225, 2008.

[53] V. Filippov, L. Mukhanov, and B. Shchukin, "Credit card fraud detection system," in *2008 7th IEEE International Conference on Cybernetic Intelligent Systems*, IEEE, Sep. 2008, pp. 1–6. doi: 10.1109/UKRICIS.2008.4798919.

[54] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," in *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, IEEE, Mar. 2019, pp. 1–5. doi: 10.1109/INFOTEH.2019.8717766.

[55] A. Maureen, C. Asuai, A. Edje, E. Omede, and U. Emmanuel, "Cybershield: harnessing ensemble feature selection technique for robust distributed denial of service attacks detection," *J. Adv. Comput. Commun. Inf. Technol.*, vol. 38, no. 03, pp. 1211–1224, 2023.

[56] M. Zareapoor and P. Shamsolmoali, "Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier," *Procedia Comput. Sci.*, vol. 48, pp. 679–685, 2015, doi: 10.1016/j.procs.2015.04.201.

[57] B. Gaye and A. Wulamu, "Sentimental Analysis for Online Reviews using Machine learning Algorithms," pp. 1270–1275, 2019.

[58] D. O. Oyewola, E. G. Dada, N. J. Ngozi, A. U. Terang, and S. A. Akinwumi, "COVID-19 Risk Factors, Economic Factors, and Epidemiological Factors nexus on Economic Impact: Machine Learning and Structural Equation Modelling Approaches," *J. Niger. Soc. Phys. Sci.*, vol. 3, no. 4, pp. 395–405, 2021, doi: 10.46481/jnsps.2021.173.

[59] A. A. Ojugo and A. O. Eboka, "An Empirical Evaluation On Comparative Machine Learning Techniques For Detection of The Distributed Denial of Service (DDoS) Attacks," *J. Appl. Sci. Eng. Technol. Educ.*, vol. 2, no. 1, pp. 18–27, 2020, doi: 10.35877/454ri.asci2192.

[60] M. I. Akazue, I. A. Debekeme, A. E. Edje, C. Asuai, and U. J. Osame, "UNMASKING FRAUDSTERS : Ensemble Features Selection to Enhance Random Forest Fraud Detection," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 201–212, 2023, doi: 10.33633/jcta.v1i2.9462.

[61] A. A. Ojugo *et al.*, "Forging a User-Trust Memetic Modular Neural Network Card Fraud Detection Ensemble: A Pilot Study," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 1–11, Oct. 2023, doi: 10.33633/jcta.v1i2.9259.

[62] E. A. L. Marazqah Btoush, X. Zhou, R. Gururajan, K. C. Chan, R. Genrich, and P. Sankaran, "A systematic review of literature on credit card cyber fraud detection using machine and deep learning," *PeerJ Comput. Sci.*, vol. 9, p. e1278, Apr. 2023, doi: 10.7717/peerj-cs.1278.

[63] J. Femila Roseline, G. Naidu, V. Samuthira Pandi, S. Alamelu alias Rajasree, and D. N. Mageswari, "Autonomous credit card fraud detection using machine learning approach☆," *Comput. Electr. Eng.*, vol. 102, p. 108132, Sep. 2022, doi: 10.1016/j.compeleceng.2022.108132.

[64] O. Sinayobye, R. Musabe, A. Uwitonze, and A. Ngenzi, "A Credit Card Fraud Detection Model Using Machine Learning Methods with a Hybrid of Undersampling and Oversampling for Handling Imbalanced Datasets for High Scores," 2023, pp. 142–155. doi: 10.1007/978-3-031-34222-6_12.

[65] A. Ali *et al.*, "Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review," *Appl. Sci.*, vol. 12, no. 19, p. 9637, Sep. 2022, doi: 10.3390/app12199637.

[66] N. Rtayli and N. Enneya, "Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization," *J. Inf. Secur. Appl.*, vol. 55, p. 102596, Dec. 2020, doi: 10.1016/j.jisa.2020.102596.

[67] A. A. Ojugo and R. E. Yoro, "Computational Intelligence in Stochastic Solution for Toroidal N-Queen," *Prog. Intell. Comput. Appl.*, vol. 1, no. 2, pp. 46–56, 2013, doi: 10.4156/pica.vol2.issue1.4.

[68] A. A. Ojugo and A. O. Eboka, "Comparative Evaluation for High Intelligent Performance Adaptive Model for Spam Phishing Detection," *Digit. Technol.*, vol. 3, no. 1, pp. 9–15, 2018, doi: 10.12691/dt-3-1-2.

[69] B. N. Supriya and C. B. Akki, "Sentiment prediction using enhanced xgboost and tailored random forest," *Int. J. Comput. Digit. Syst.*, vol. 10, no. 1, pp. 191–199, 2021, doi: 10.12785/ijcds/100119.

[70] S. Meghana, B. . Charitha, S. Shashank, V. S. Sulakhe, and V. B. Gowda, "Developing An Application for Identification of Missing Children and Criminal Using Face Recognition.," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 12, no. 6, pp. 272–279, 2023, doi: 10.17148/ijarcce.2023.12648.

[71] Sharmila, R. Sharma, D. Kumar, V. Puranik, and K. Gautham, "Performance Analysis of Human Face Recognition Techniques," *Proc. - 2019 4th Int. Conf. Internet Things Smart Innov. Usages, IoT-SIU 2019*, no. May 2020, pp. 1–4, 2019, doi: 10.1109/IoT-SIU.2019.8777610.

[72] E. Omede, J. Anenechukwu, and C. Hampo, "Use of Adaptive Boosting Algorithm to Estimate User's Trust in the Utilization of Virtual Assistant Systems," *Int. J. Innov. Sci. Res. Technol.*, vol. 8, no. 1, pp. 502–509, 2023.

[73] M. K. G. Roshan, "Multiclass Medical X-ray Image Classification using Deep Learning with Explainable

AI," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 6, pp. 4518–4526, Jun. 2022, doi: 10.22214/ijraset.2022.44541.

[74] A. A. Ojugo and O. D. Otakore, "Forging An Optimized Bayesian Network Model With Selected Parameters For Detection of The Coronavirus In Delta State of Nigeria," *J. Appl. Sci. Eng. Technol. Educ.*, vol. 3, no. 1, pp. 37–45, Apr. 2021, doi: 10.35877/454RI.asci2163.

[75] A. A. Ojugo and A. O. Eboka, "Empirical Bayesian network to improve service delivery and performance dependability on a campus network," *IAES Int. J. Artif. Intell.*, vol. 10, no. 3, p. 623, Sep. 2021, doi: 10.11591/ijai.v10.i3.pp623-635.

[76] L. De Kimpe, M. Walrave, W. Hardyns, L. Pauwels, and K. Ponnet, "You've got mail! Explaining individual differences in becoming a phishing target," *Telemat. Informatics*, vol. 35, no. 5, pp. 1277–1287, Aug. 2018, doi: 10.1016/j.tele.2018.02.009.

[77] K. Deepika, M. P. S. Nagenddra, M. V. Ganesh, and N. Naresh, "Implementation of Credit Card Fraud Detection Using Random Forest Algorithm," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 3, pp. 797–804, Mar. 2022, doi: 10.22214/ijraset.2022.40702.

[78] A. A. Ojugo, P. O. Ejeh, C. C. Odiakaose, A. O. Eboka, and F. U. Emordi, "Improved distribution and food safety for beef processing and management using a blockchain-tracer support framework," *Int. J. Informatics Commun. Technol.*, vol. 12, no. 3, p. 205, Dec. 2023, doi: 10.11591/ijict.v12i3.pp205-213.

[79] P. Boulieris, J. Pavlopoulos, A. Xenos, and V. Vassalos, "Fraud detection with natural language processing," *Mach. Learn.*, Jul. 2023, doi: 10.1007/s10994-023-06354-5.

[80] F. U. Emordi *et al.*, "TiSPHiMME: Time Series Profile Hidden Markov Ensemble in Resolving Item Location on Shelf Placement in Basket Analysis," *Digit. Innov. Contemp. Res. Sci.*, vol. 12, no. 1, pp. 33–48, 2024, doi: 10.22624/AIMS/DIGITAL/v11N4P3.

[81] M. K. Elmezughi, O. Salih, T. J. Afullo, and K. J. Duffy, "Comparative Analysis of Major Machine-Learning-Based Path Loss Models for Enclosed Indoor Channels," *Sensors*, vol. 22, no. 13, p. 4967, Jun. 2022, doi: 10.3390/s22134967.

[82] D. Kilroy, G. Healy, and S. Caton, "Using Machine Learning to Improve Lead Times in the Identification of Emerging Customer Needs," *IEEE Access*, vol. 10, pp. 37774–37795, 2022, doi: 10.1109/ACCESS.2022.3165043.

[83] F. Safara, "A Computational Model to Predict Consumer Behaviour During COVID-19 Pandemic," *Comput. Econ.*, vol. 59, no. 4, pp. 1525–1538, Apr. 2022, doi: 10.1007/s10614-020-10069-3.

[84] I. P. Okobah and A. A. Ojugo, "Evolutionary Memetic Models for Malware Intrusion Detection: A Comparative Quest for Computational Solution and Convergence," *Int. J. Comput. Appl.*, vol. 179, no. 39, pp. 34–43, 2018, doi: 10.5120/ijca2018916586.

[85] D. Linh, "Insider threat detection: Where and how data science applies," *Cyber Secur. A Peer-Reviewed J.*, vol. 2, pp. 1–8, 2018, [Online]. Available: https://www.ingentaconnect.com/content/hsp/jcs/2018/00000002/00000003/art00003

[86] A. A. Ojugo and E. O. Ekurume, "Deep Learning Network Anomaly-Based Intrusion Detection Ensemble For Predictive Intelligence To Curb Malicious Connections: An Empirical Evidence," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 3, pp. 2090–2102, Jun. 2021, doi: 10.30534/ijatcse/2021/851032021.

[87] C. C. Odiakaose *et al.*, "DeLEMPaD: Pilot Study on a Deep Learning Ensemble for Energy Market Prediction of Price Volatility and Direction," *Comput. Inf. Syst. Dev. Informatics Allied Res. J.*, vol. 15, no. 1, pp. 47–62, 2024, doi: 10.22624/AIMS/CISDI/V15N1P4.

[88] F. Jáñez-Martino, E. Fidalgo, S. González-Martínez, and J. Velasco-Mata, "Classification of Spam Emails through Hierarchical Clustering and Supervised Learning," *Natl. Cybersecurity Inst.*, vol. 24, pp. 1–4, May 2020, [Online]. Available: http://arxiv.org/abs/2005.08773

[89] A. Maureen, O. Oghenefego, A. E. Edje, and C. O. Ogeh, "An Enhanced Model for the Prediction of Cataract Using Bagging Techniques," vol. 8, no. 2, 2023.

[90] D. H. Zala and M. B. Chaudhari, "Review on use of 'BAGGING' technique in agriculture crop yield prediction," *IJSRD - Int. J. Sci. Res. Dev.*, vol. 6, no. 8, pp. 675–676, 2018.

[91] R. E. Yoro and A. A. Ojugo, "Quest for Prevalence Rate of Hepatitis-B Virus Infection in the Nigeria: Comparative Study of Supervised Versus Unsupervised Models," *Am. J. Model. Optim.*, vol. 7, no. 2, pp. 42–48, 2019, doi: 10.12691/ajmo-7-2-2.

[92] A. A. Ojugo *et al.*, "Dependable Community-Cloud Framework for Smartphones," *Am. J. Networks Commun.*, vol. 4, no. 4, p. 95, 2015, doi: 10.11648/j.ajnc.20150404.13.

[93] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A Comparative Analysis of XGBoost," no. February, 2019, doi: 10.1007/s10462-020-09896-5.

[94] N. M. Shahani, X. Zheng, C. Liu, F. U. Hassan, and P. Li, "Developing an XGBoost Regression Model for Predicting Young's Modulus of Intact Sedimentary Rocks for the Stability of Surface and Subsurface Structures," *Front. Earth Sci.*, vol. 9, Oct. 2021, doi: 10.3389/feart.2021.761990.

[95] G. Cho, J. Yim, Y. Choi, J. Ko, and S. H. Lee, "Review of machine learning algorithms for diagnosing mental illness," *Psychiatry Investig.*, vol. 16, no. 4, pp. 262–269, 2019, doi: 10.30773/pi.2018.12.21.2.

[96] D. A. Al-Qudah, A. M. Al-Zoubi, P. A. Castillo-Valdivieso, and H. Faris, "Sentiment analysis for e-payment service providers using evolutionary

extreme gradient boosting," *IEEE Access*, vol. 8, pp. 189930–189944, 2020, doi: 10.1109/ACCESS.2020.3032216.

[97] A. A. Ojugo, D. A. Oyemade, D. Allenotor, O. B. Longe, and C. N. Anujeonye, "Comparative Stochastic Study for Credit-Card Fraud Detection Models," *African J. Comput. ICT*, vol. 8, no. 1, pp. 15–24, 2015, [Online]. Available: www.ajocict.net

[98] F. Omoruwou, A. A. Ojugo, and S. E. Ilodigwe, "Strategic Feature Selection for Enhanced Scorch Prediction in Flexible Polyurethane Form Manufacturing," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 126–137, 2024, doi: 10.62411/jcta.9539.

[99] R. E. Yoro and A. A. Ojugo, "An Intelligent Model Using Relationship in Weather Conditions to Predict Livestock-Fish Farming Yield and Production in Nigeria," *Am. J. Model. Optim.*, vol. 7, no. 2, pp. 35–41, 2019, doi: 10.12691/ajmo-7-2-1.

[100] T. Edirisooriya and E. Jayatunga, "Comparative Study of Face Detection Methods for Robust Face Recognition Systems," *5th SLAAI - Int. Conf. Artif. Intell. 17th Annu. Sess. SLAAI-ICAI 2021*, no. December, 2021, doi: 10.1109/SLAAI-ICAI54477.2021.9664689.

[101] M. G. Kibria and M. Sevkli, "Application of Deep Learning for Credit Card Approval: A Comparison with Two Machine Learning Techniques," *Int. J. Mach. Learn. Comput.*, vol. 11, no. 4, pp. 286–290, Aug. 2021, doi: 10.18178/ijmlc.2021.11.4.1049.

[102] A. Razaque *et al.*, "Credit Card-Not-Present Fraud Detection and Prevention Using Big Data Analytics Algorithms," *Appl. Sci.*, vol. 13, no. 1, p. 57, Dec. 2022, doi: 10.3390/app13010057.

[103] A. Satpathi *et al.*, "Comparative Analysis of Statistical and Machine Learning Techniques for Rice Yield Forecasting for Chhattisgarh, India," *Sustainability*, vol. 15, no. 3, p. 2786, Feb. 2023, doi: 10.3390/su15032786.

[104] A. Bahl *et al.*, "Recursive feature elimination in random forest classification supports nanomaterial grouping," *NanoImpact*, vol. 15, p. 100179, Mar. 2019, doi: 10.1016/j.impact.2019.100179.

[105] C. L. Udeze, I. E. Eteng, and A. E. Ibor, "Application of Machine Learning and Resampling Techniques to Credit Card Fraud Detection," *J. Niger. Soc. Phys. Sci.*, vol. 12, p. 769, Aug. 2022, doi: 10.46481/jnsps.2022.769.

[106] B. P. Bhuyan, R. Tomar, T. P. Singh, and A. R. Cherif, "Crop Type Prediction: A Statistical and Machine Learning Approach," *Sustainability*, vol. 15, no. 1, p. 481, Dec. 2022, doi: 10.3390/su15010481.

[107] E. U. Omede, A. Edje, M. I. Akazue, H. Utomwen, and A. A. Ojugo, "IMANoBAS: An Improved Multi-Mode Alert Notification IoT-based Anti-Burglar Defense System," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 43–53, 2024, doi: 10.33633/jcta.v2i1.9541.

[108] B. Ghaffari and Y. Osman, "Customer churn prediction using machine learning: A study in the B2B subscription based service context," Faculty of Computing, Blekinge Institute of Technology, Sweden, 2021. [Online]. Available: www.bth.se

[109] B. O. Malasowe, M. I. Akazue, E. A. Okpako, F. O. Aghware, D. V. Ojie, and A. A. Ojugo, "Adaptive Learner-CBT with Secured Fault-Tolerant and Resumption Capability for Nigerian Universities," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, pp. 135–142, 2023, doi: 10.14569/IJACSA.2023.0140816.

[110] J. K. Oladele *et al.*, "BEHeDaS: A Blockchain Electronic Health Data System for Secure Medical Records Exchange," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 1–12, 2024, doi: 10.33633/jcta.v2i19509.

[111] M. Srividya, S. Mohanavalli, and N. Bhalaji, "Behavioral Modeling for Mental Health using Machine Learning Algorithms," *J. Med. Syst.*, vol. 42, no. 5, 2018, doi: 10.1007/s10916-018-0934-5.

[112] C. Ren *et al.*, "Short-Term Traffic Flow Prediction: A Method of Combined Deep Learnings," *J. Adv. Transp.*, vol. 2021, pp. 1–15, Jul. 2021, doi: 10.1155/2021/9928073.

[113] A. A. Ojugo *et al.*, "Forging a learner-centric blended-learning framework via an adaptive content-based architecture," *Sci. Inf. Technol. Lett.*, vol. 4, no. 1, pp. 40–53, May 2023, doi: 10.31763/sitech.v4i1.1186.

[114] A. A. Ojugo and R. E. Yoro, "Extending the three-tier constructivist learning model for alternative delivery: ahead the COVID-19 pandemic in Nigeria," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 21, no. 3, p. 1673, Mar. 2021, doi: 10.11591/ijeecs.v21.i3.pp1673-1682.

[115] A. A. Ojugo and O. D. Otakore, "Computational solution of networks versus cluster grouping for social network contact recommender system," *Int. J. Informatics Commun. Technol.*, vol. 9, no. 3, p. 185, 2020, doi: 10.11591/ijict.v9i3.pp185-194.

[116] D. A. Oyemade *et al.*, "A Three Tier Learning Model for Universities in Nigeria," *J. Technol. Soc.*, vol. 12, no. 2, pp. 9–20, 2016, doi: 10.18848/2381-9251/CGP/v12i02/9-20.

[117] O. E. Ojo, A. Gelbukh, H. Calvo, and O. O. Adebanji, "Performance Study of N-grams in the Analysis of Sentiments," *J. Niger. Soc. Phys. Sci.*, vol. 3, no. 4, pp. 477–483, 2021, doi: 10.46481/jnsps.2021.201.

[118] R. G. Bhati, "A Survey on Sentiment Analysis Algorithms and Datasets," *Rev. Comput. Eng. Res.*, vol. 6, no. 2, pp. 84–91, 2019, doi: 10.18488/journal.76.2019.62.84.91.

[119] K. Afifah, I. N. Yulita, and I. Sarathan, "Sentiment Analysis on Telemedicine App Reviews using XGBoost Classifier," *2021 Int. Conf. Artif. Intell. Big Data Anal.*, pp. 22–27, 2022, doi: 10.1109/icaibda53487.2021.9689735.

[120] A. A. Ojugo and A. O. Eboka, "Assessing Users Satisfaction and Experience on Academic Websites: A Case of Selected Nigerian Universities Websites,"

Int. J. Inf. Technol. Comput. Sci., vol. 10, no. 10, pp. 53–61, 2018, doi: 10.5815/ijitcs.2018.10.07.

[121] L. Á. Redondo-Gutierrez, F. Jáñez-Martino, E. Fidalgo, E. Alegre, V. González-Castro, and R. Alaiz-Rodríguez, "Detecting malware using text documents extracted from spam email through machine learning," in *Proceedings of the 22nd ACM Symposium on Document Engineering*, New York, NY, USA: ACM, Sep. 2022, pp. 1–4. doi: 10.1145/3558100.3563854.

[122] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, "A Comprehensive Survey for Intelligent Spam Email Detection," *IEEE Access*, vol. 7, pp. 168261–168295, 2019, doi: 10.1109/ACCESS.2019.2954791.

[123] Z. Karimi, M. Mansour Riahi Kashani, and A. Harounabadi, "Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods," *Int. J. Comput. Appl.*, vol. 78, no. 4, pp. 21–27, Sep. 2013, doi: 10.5120/13478-1164.

[124] A. A. Ojugo and O. D. Otakore, "Investigating The Unexpected Price Plummet And Volatility Rise In Energy Market: A Comparative Study of Machine Learning Approaches," *Quant. Econ. Manag. Stud.*, vol. 1, no. 3, pp. 219–229, 2020, doi: 10.35877/454ri.qems12119.

[125] A. Jayatilaka, N. A. G. Arachchilage, and M. A. Babar, "Falling for Phishing: An Empirical Investigation into People's Email Response Behaviors," *arXiv Prepr. arXiv ...*, no. Fbi 2020, pp. 1–17, 2021.

[126] A. A. Ojugo and R. E. Yoro, "Forging a deep learning neural network intrusion detection framework to curb the distributed denial of service attack," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 2, pp. 1498–1509, 2021, doi: 10.11591/ijece.v11i2.pp1498-1509.

[127] S. Paliwal, A. K. Mishra, R. K. Mishra, N. Nawaz, and M. Senthilkumar, "XGBRS Framework Integrated with Word2Vec Sentiment Analysis for Augmented Drug Recommendation," *Comput. Mater. Contin.*, vol. 72, no. 3, pp. 5345–5362, 2022, doi: 10.32604/cmc.2022.025858.

[128] A. A. Ojugo, A. O. Eboka, R. E. Yoro, M. O. Yerokun, and F. N. Efozia, "Hybrid Model for Early Diabetes Diagnosis," *Proc. - 2015 2nd Int. Conf. Math. Comput. Sci. Ind. MCSI 2015*, pp. 55–65, 2016, doi: 10.1109/MCSI.2015.35.

[129] M. Rathi and V. Pareek, "Spam Mail Detection through Data Mining – A Comparative Performance Analysis," *Int. J. Mod. Educ. Comput. Sci.*, vol. 5, no. 12, pp. 31–39, 2013, doi: 10.5815/ijmecs.2013.12.05.

[130] Y. Gao, S. Zhang, J. Lu, Y. Gao, S. Zhang, and J. Lu, "Machine learning for credit card fraud detection," in *Proceedings of the 2021 International Conference on Control and Intelligent Robotics*, New York, USA: ACM, Jun. 2021, pp. 213–219. doi: 10.1145/3473714.3473749.

[131] P. . Maya Gopal and Bhargavi R, "Feature Selection for Yield Prediction Using BORUTA Algorithm," *Int. J. Pure Appl. Math.*, vol. 118, no. 22, pp. 139–144, 2018.

[132] S. Yuan and X. Wu, "Deep learning for insider threat detection: Review, challenges and opportunities," *Comput. Secur.*, vol. 104, 2021, doi: 10.1016/j.cose.2021.102221.

[133] F. O. Aghware *et al.*, "Enhancing the Random Forest Model via Synthetic Minority Oversampling Technique for Credit-Card Fraud Detection," *J. Comput. Theor. Appl.*, vol. 2, no. 2, pp. 190–203, 2024, doi: 10.62411/jcta.10323.

[134] B. O. Malasowe, A. E. Okpako, M. D. Okpor, P. O. Ejeh, A. A. Ojugo, and R. E. Ako, "FePARM: The Frequency-Patterned Associative Rule Mining Framework on Consumer Purchasing-Pattern for Online Shops," *Adv. Multidiscip. Sci. Res. J. Publ.*, vol. 15, no. 2, pp. 15–28, 2024, doi: 10.22624/AIMS/CISDI/V15N2P2-1.

[135] A. A. Ojugo and O. D. Otakore, "Intelligent cluster connectionist recommender system using implicit graph friendship algorithm for social networks," *IAES Int. J. Artif. Intell.*, vol. 9, no. 3, p. 497~506, 2020, doi: 10.11591/ijai.v9.i3.pp497-506.

[136] S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN Framework for Crop Yield Prediction," *Front. Plant Sci.*, vol. 10, Jan. 2020, doi: 10.3389/fpls.2019.01750.

[137] S. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks," *Front. Plant Sci.*, vol. 10, May 2019, doi: 10.3389/fpls.2019.00621.

[138] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, "Teaching Johnny not to fall for phish," *ACM Trans. Internet Technol.*, vol. 10, no. 2, pp. 1–31, May 2010, doi: 10.1145/1754393.1754396.