Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

Pilot Study on Enhanced Detection of Cues over Malicious Sites Using Data Balancing on the Random Forest Ensemble

Margaret Dumebi Okpor [1,*], Fidelis Obukohwo Aghware [2], Maureen Ifeanyi Akazue [3], Andrew Okonji Eboka [4], Rita Erhovwo Ako [5], Arnold Adimabua Ojugo [5], Chris Chukwufunaya Odiakaose [6], Amaka Patience Binitie [4], Victor Ochuko Geteloma [5], and Patrick Ogholuwarami Ejeh [6]

[1] Department of Cybersecurity, Delta State University of Science and Technology Ozoro, Nigeria; e-mail : okpormd@dsust.edu.ng
[2] Department of Computer Science, University of Delta Agbor, Nigeria; e-mail : fidelis.aghware@unidel.edu.ng
[3] Department of Computer Science, Delta State University Abraka, Nigeria; e-mail : akazue@delsu.edu.ng
[4] Department of Computer, Federal College of Education Technical Asaba, Nigeria; e-mail : ebokaandrew@gmail.com; amaka.binitie@fcetasaba.edu.ng
[5] Department of Computer Science, Federal University of Petroleum Resources Effurun, Nigeria; e-mail : ako.rita@fupre.edu.ng; ojugo.arnold@fupre.edu.ng; geteloma.victor@fupre.edu.ng
[6] Department of Computer Science, Dennis Osadebay University Asaba, Nigeria; e-mail : osegalaxy@gmail.com; patrick.ejeh@dou.edu.ng
* Corresponding Author: Margaret Dumebi Okpor

Abstract: The digital revolution frontiers have rippled across society today – with various web content shared online for users as they seek to promote monetization and asset exchange, with clients constantly seeking improved alternatives at lowered costs to meet their value demands. From item upgrades to their replacement, businesses are poised with retention strategies to help curb the challenge of customer attrition. The birth of smartphones has proliferated feats such as mobility, ease of accessibility, and portability – which, in turn, have continued to ease their rise in adoption, exposing user device vulnerability as they are quite susceptible to phishing. With users classified as more susceptible than others due to online presence and personality traits, studies have sought to reveal lures/cues as exploited by adversaries to enhance phishing success and classify web content as genuine and malicious. Our study explores the tree-based Random Forest to effectively identify phishing cues via sentiment analysis on phishing website datasets as scrapped from user accounts on social network sites. The dataset is scrapped via Python Google Scrapper and divided into train/test subsets to effectively classify contents as genuine or malicious with data balancing and feature selection techniques. With Random Forest as the machine learning of choice, the result shows 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.

Keywords: Malicious contents; Phishing; Random Forest; Social engineering; Tree-based models.

Received: June, 8th 2024
Revised: September, 5th 2024
Accepted: September, 6th 2024
Published: September, 7th 2024
Curr. Ver.: September, 7th 2024

1. Introduction

The Internet and the constant evolution in informatics – have become both the mainstay of and backbone of businesses today [1]. Its infrastructure connects businesses and helps them meet client needs [2]. With the Internet as an efficient means to disseminate and share data, many adversaries utilize it to proliferate malicious content [3]. Access to malicious content has since become a multi-billion-dollar challenge that plagues users daily [4]. Despite the plethora of continued studies that sought to improve detection via filtering and classification schemes, users continue to fall prey to scams [5]. This is attributed to websites being rippled with content that presents as insecure adverts or hides in third-party legitimate software [6], [7]. Various studies have begun investigating how various aspects seek to compromise data – even with many cyber measures in place [8]. One such concern is how the Internet is gradually

replacing normal social activities as users now engage with web content – as tools to compensate for their loneliness and social seclusion [9], [10].

The digital revolution seeks to integrate informatics and its enabling technologies into every facet of our society [11]. Thus, changing our mode of service delivery to clients in lieu of the value they get from the services rendered [12]. It is also a cultural change that requires businesses to challenge the status quo continually [13], experiment, and get comfortable with failure [14]. Thus, as more individuals connect via enabling support devices [15] – it also opens up many of such users to avenues of exploitation to be harnessed by adversaries via socially engineered attacks [16]. These attacks are an old paradigm that continues to grow, with no end in sight steadily. 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) include phishing, pharming, spamming, vishing, etc. This gives an attacker an attractive entry point to compromise a victim's device and become a pivot point for attack spread [22]. With such attacks targeted at user-connected devices and with over 200 percent adoption of smartphones, users have become more vulnerable and compromised victims [23] alongside its range of complications to work and business-related issues caused by the exposure of sensitive users to [24], [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]. Spear phishing is an effective, favored variant, which uses targeted mail with access links to cleverly persuade potential victims and redirect them to spoofed malicious web content 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 redirects user traffic to a fake site by either changing the host's file on a victim's device or exploiting the vulnerability in the domain name service server software. Thus, it allows an adversary to install malware on a user's device and redirect the user to a fraudulent site without their consent or knowledge [29], [30]. 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 consists of 3-elements, namely: (a) the potential victim receives a lure message as originating from a legitimate source, and its reliability is strengthened by exploiting a user curiosity, fear, and empathy; (b) a hook is often a compromised link/attachment included in the message, and (c) the catch involves an attacker means to obtain a user's private data [30], [31].

This may appear simple enough, but the technique(s) constantly evolves 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 [32]. Thus, phishing is often posited as a message from trusted entities to compromise a victim. Its characteristics include: (a) the message often makes unrealistic demands via various targeted intimidation of a user's psych [33], (b) there is always a catch, (c) there is often missing data with spelling errors and poor grammar, (d) there is often a mismatch in its URL (uniform resource locator) to redirect users to faked sites, and (e) messages often demands sensitive, confidential user data. Umarani et al. [34] used victimization features to characterize the websites' design impact on both the content's structure and the probability that content will victimize a user. They used 2-feats to help users identify malicious contents and eliminate awareness gaps, namely: (a) the believability to identify cues [35], which increases the possibility a user will believe a message, and (b) the insidiousness to measure the potency in degradation lures [36] and its success rate while remaining undetectable to users.

2. Preliminaries

2.1. Machine Learning Approaches

Ezpeleta et al. [37] investigated spam attacks with millions of malicious files sent daily via spam. They posited that for many users – it is about control rather than preventing and mitigating spam via filters and other schemes as technical measures. Also, the users' level of suspicion, emotional control [38], [39], and attack awareness must become a critical feat in either the success or failure of an attack – since emotion becomes personality behaviour and

traits that culminates as cues that drive the desire to help [40], to seek gain via exploitation, and to be liked. All these suggestions make some persons more susceptible to attacks [41], [42]; And such victims may fall repeatedly into a scam.

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]. 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 improving phishing detection and prevention systems. Despite these efforts, adversaries continue to invent new techniques to circumvent these security measures and avoid detection, making it a constant battle [44], [45].

To curb and minimize the effect of phishing attacks on web content, machine learning approaches have been successfully trained and adapted to effectively recognize phishing patterns within web content as cues and lures. There they learn through features classification either from the normal behavior cum signature in transactions or the quick detection of unusual activity in the transaction pattern indicative of a fraudulent profile. Various machine learning (ML) schemes have been successfully used in the detection of phishing attacks, namely Logistic Regression [46], Deep Learning [47], [48], Bayes [49], SVM [50], Random Forest [51], and others that have been effectively used to detect phishing cases. Many of these have drawbacks with their feature selection and accuracy flexibility.

The choice of our study with the adoption and adaption of the tree-based Random Forest heuristics is due to its capability to greatly reduce model overfitting, resolve data encoding conflicts, easily accept schemes and approaches to resolving dataset imbalance. Its capability to yield a vigorous accuracy, and to also yield enhanced model prediction cum malicious content detection performance [52], [53].

2.2. Tree-based Heuristic(s)

A common ML approach is the tree-based method, which descends from single decision trees. Adopting a tree structure, each tree generates a series of if-else rules used in the majority voting scheme, allowing it to predict observed classes [54]. 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 [55]. Each decision tree has the merit of being easily understood [56], But, its use alone often leads to model overfit in a prediction task as the model seeks to identify feats of interest during training. Thus, it degrades performance in classifying unknown labels [57]. Tree-based models learn by constructing many individually trained decision trees [58]. They combine/aggregate their results into a single and stronger model whose output outperforms the results of any single tree [59]. It achieves this via either bagging [60], [61], and boosting [62], [63] modes.

In the case of boosting – the tree(s) converts weak learners (i.e., achieve accuracy just above random guess) onto a strong learners with enhanced predictive capacity by sequentially training each weak learner to correct the inherent weaknesses of its predecessor [64], [65]. Each tree yields feedback from previous weaker trees [66]. Popular boosting models include gradient boost [67], LogitBoost [68], stochastic gradient boost [69], and adaptive boost [70]. They can often be expressed via Equation (1) – which makes its prediction by combining the outcome of its weak learners with its weighted sum to yield a higher weight for incorrectly classified cases as in Equation – where L^t is the objective function, $l\left(Y_i^t, \hat{Y}_i^t\right)$ is the loss function, and (Ωf_t) is its regularization term.

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

Conversely, bagging grows successive trees independently from earlier trees – such that each tree is constructed using a bootstrap aggregation mode to sample the data using a majority vote during its prediction [71]. The Random Forest adds an extra layer of randomness to the bagging scheme, which changes how the trees are constructed. While standard decision

trees have 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 [72]. Its recursive structure helps it to capture interaction effects between variables. In general, tree-based models have successfully proven to be better than other established approaches across a variety of different tasks [73], ranging from traffic flow classification [74], customer churn prediction [75], and prediction of online purchase intention [76]. 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 [77], a weighted combination of several different local minima – produced by ensemble methods [78] can minimize the risk of choosing the wrong local minimum [79].

2.3. Study Motivation

Inherent gaps from existing studies include [80]–[84]:

1. Lack of Datasets: Finding the right-format dataset – is crucial to machine learning tasks. Access to high-quality datasets is needed in training and performance evaluation – as there is limited data, which often yields significant false positives [85].
2. Imbalanced Datasets: A critical challenge with dataset imbalances is that phishing cases are often unreported. Thus, large datasets show that phishing cases often lag behind in class distribution plots. Studies must explore intricate sampling techniques or harness the robust power of ensemble methods tailored explicitly to mitigate the challenges with imbalanced datasets [86].
3. Cross-Channel Acquisition: The continued rise in both volume, veracity, and value of data generated across multiple channels [87]–[89] by a variety of businesses and users has continued to ensure the use of big data analytics as data mining methodologies and heuristics must seek to glean meaning, insightful knowledge from such huge data. Thus, the new model must account for these changes and must integrate means to harness these data-points generation (i.e., from the various channel data) to enhance the overall accuracy and performance of the models; as such, cross-channel detection has now become a critical area of research and business focus [90].

Thus, we construct known tree-based models using bagging and boosting capabilities with data balancing techniques on the dataset as retrieved from Kaggle. This aims at a comparative predictive analytic(s) and ascertain which model best fits the data balancing technique for future studies. Our study hopes to achieve these feats:

1. Model construction: To yield a more sophisticated decision support model in detecting phishing lures and cues that render users more susceptible to attacks vis-à-vis compromising network infrastructure. Using a machine learning scheme will help the system effectively capture those cues that make phishing more successful [91], [92].
2. Data balancing: The resultant model(s) will investigate the effects of data balancing on the reliability cum predictive power of the tree-based Random Forest model; whilst, analyzing its implication on the model's capability to predict phishing attack cues accurately. This will help users glean insightful knowledge on the significance of their online presence vis-à-vis enhancing a model's performance in various contexts [93].
3. Comparative analysis will evaluate diverse machine learning approaches within the constructed prediction model, aimed at comparing the performance, accuracy, and robustness of various algorithms to identify sophisticated cue and degradation lures that trick susceptible users during their time online over social networking sites.

3. Proposed Method

The ease of access to 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 increased, lessening user trust in shared data [94], [95]. Generally, a user's opinion of an idea or topic of interest is his/her belief centered on his/her perception or feeling toward the issue. The beliefs and opinions represent the user's disposition of emotion. This emotion correlates with his/her behaviour concerning the situation and is called sentiment. Thus, sentiment analysis deals with a language class that seeks to trace and track a user's or community's behaviour toward a topic of interest [96]. In natural language processing

– its data is often unstructured and, thus, rippled with ambiguities, noise, and imprecisions. The proposed model is herein seen in figure 1.

The methodology adopted for construction, training and testing of the phishing detection model using the tree-based algorithms Random Forest model will be divided into these sections: (a) data collection and preprocessing, (b) proposed Random Forest model explained, and (c) model construction and training. They are explained below as thus:

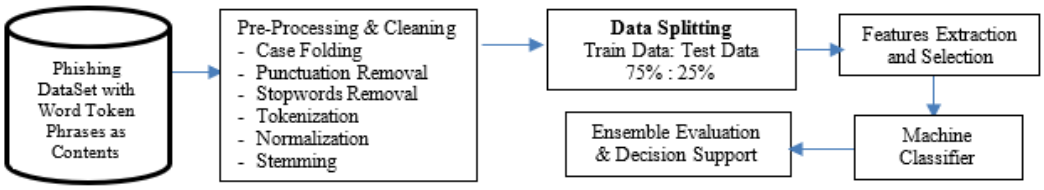

Figure 1. Sentiment-based Analysis Process and Decision Support

3.1. Dataset: Collection

Data were collected using Google Scraper Library, and 8,693 records were gathered from June to December 2023 from participants (i.e., undergraduates of the Federal University of Petroleum Resources Effurun in Nigeria). Scrapped records consist of personal user data, compromised contents (links, images, and texts), emails, and sites (posts, likes, shares, and replies). Input records were transformed via PCA [97], [98].

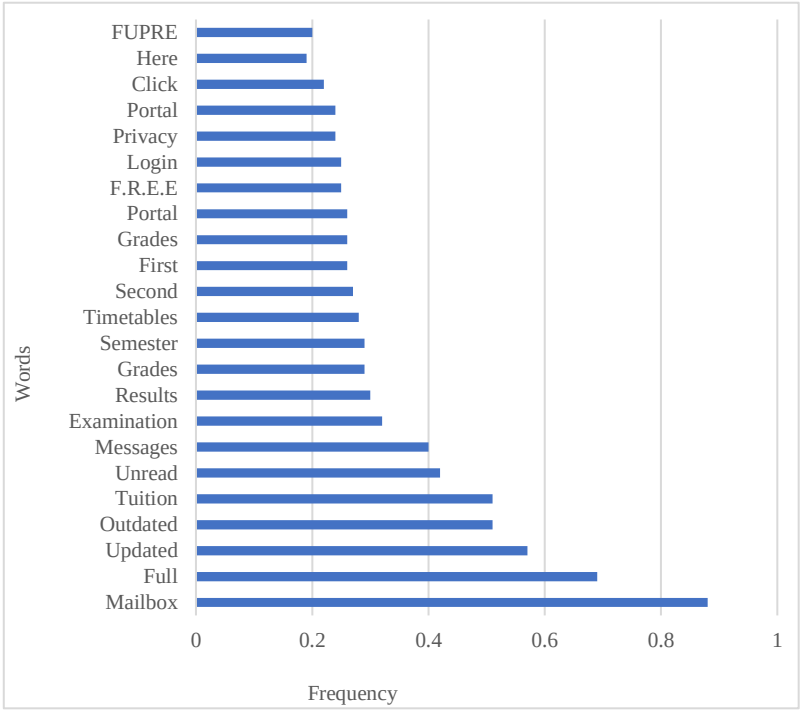

Figure 2. Frequency chart of Word Sentiments

3.2. Data PreProcessing

Some reasons for our choice of the tree-based algorithm are: (a) each tree learns and votes to decide the outcome of the classifier, (b) it can effectively handle complex, continuous, and categorical datasets, (c) it often yields improved generalization and is devoid of overfit, (d) they efficiently reflect in a heuristic, relative contribution of feature selection to performance, and (e) they are resilient to noise in the quest for ground-truth even with (un)structured dataset for real-time applications [99], [100]. Data preprocessing is performed as thus [101]:

3.2.1 Data Cleaning

Collects and cleans the dataset. It restructures the dataset from its natural unstructured state(s) to a normalized state by removing phrases and stopwords via tokenization and word stem.

1. 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 [102].
2. Punctuation removal seeks to remove all symbols/punctuation from word tokens. We also note that punctuation marks in natural language processing (NLP) do not add extra information. However, it reduces our dataset's dimensionality, which needs to be resolved.
3. Stopword removal seeks to remove some common tokens/words across all the documents. Stopwords like punctuations do not add much information to the scenario; their removal also only reduces the dimensionality in our dataset to be resolved. We note that pronouns, articles, conjunctions, and prepositions – are classified as stopwords.
4. Tokenization breaks down a sentence into smaller elements and helps interpret the implicit meaning of a sentence by analyzing its order of placement in the text. These yield input for NLPs models with normalized texts broken into individual word elements, stopwords, and punctuation characters (that are equally removed in this unit) [103].
5. 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 and expands abbreviations into their complete forms. E.g. the term "notin" is changed to "nothing"; And "welcome" is transformed to its base "welcome." We used the dictionary by Ojugo and Eboka [41] and Afifah et al. [104] for normalization.
6. Word Stemming is a step to remove affixes in a word, both appearing before and after the word. Stemming converts each word to its root word without affixes.

3.2.2 Data Balancing:

A critical feature here is to adopt a properly-format dataset. ML is applied to tasks that require (a) flexibility to adequately encode a chosen dataset irrespective of its format/structure, (b) robustness to be re-used in related task (s), and (c) adaptive to yield cost-effective alternates as optimal solution irrespective of ambiguity, noise, and partial truth as contained therein the dataset used. Learning the underlying feats of interest in an ill-formatted, imbalanced dataset – leads to poor generalization and results in imbalanced learning. A balanced generalization is a product of balanced data in a balanced learning [104]. An imbalanced dataset results when a sample class overwhelmingly dominates the dataset and yields an imbalanced class distribution. Studies have often posited that a balanced dataset enhances the overall performance of classifier evaluation. Various data balancing modes are explored in ML to help address dataset imbalance issues as in figure 3 [105], [106] including:

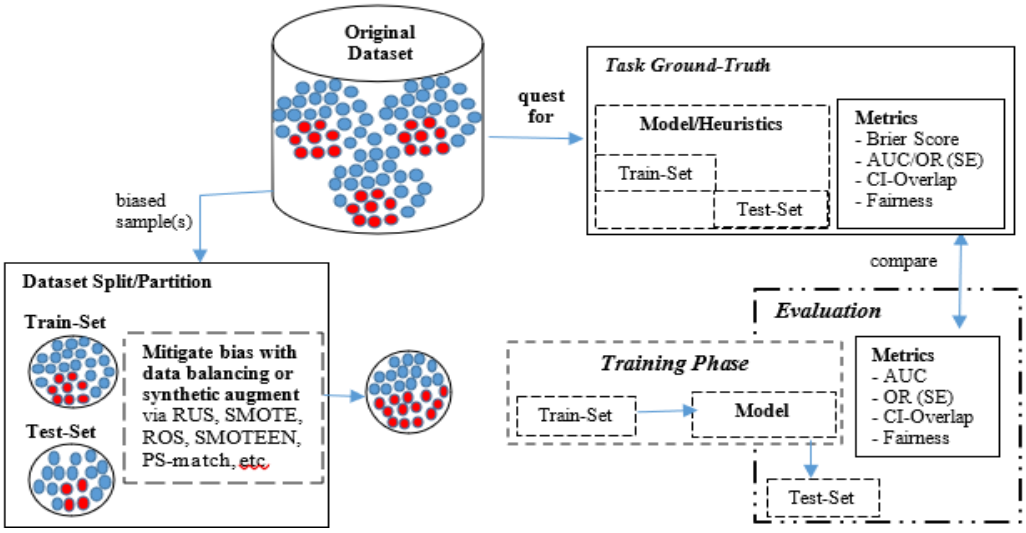

Figure 3. Data balancing technique in Machine Learning

In preprocessing, dataset splitting into train/test sub-sets occurs after dataset balancing. Test datasets often consist of hypothetical cases to enable critical examination of model's

ability to identify the churn class. Inherent benefits of balancing are: (a) prevents variance, bias, and skewness in datasets that often distort cum hamper performance, (b) enhances generalization as model can adequately learn patterns from all classes even with majority or minority voting, (c) helps the model to detect anomalies during testing effectively, and (d) its characteristics linked to the majority class often impacts significance as balancing helps a model to understand better the significance of each feature in each class to yield more insightful result(s). The 3-major modes of data balancing include:

1. Under-sampling randomly reduces a majority class till all class distributions are roughly equal. It achieves this by exploring its *k*-closest neighbor to identify points and link them to the original dataset. Thus, it cleans up the dataset's oversampled points in the majority class distribution [107], [108] as in Figure 4. See [109] for details.

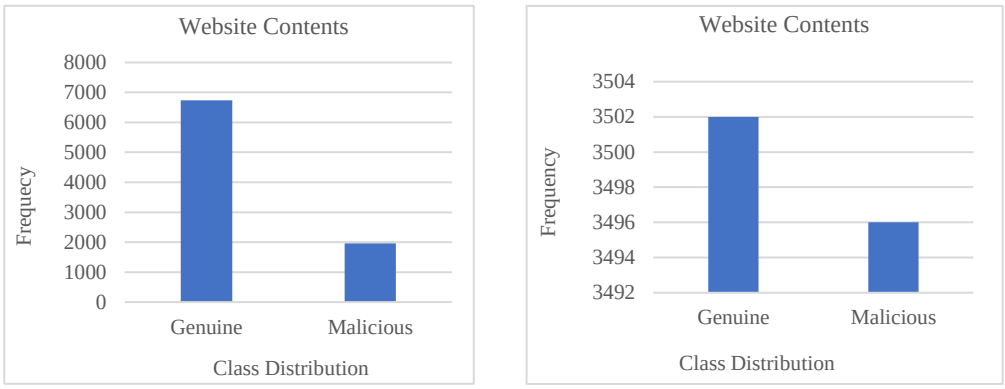

Figure 4. Class Distribution (**a**) Original Data; (**b**) Undersampling Applied to Data

2. Over-Sample Technique: We adopt the synthetic over-sample technique (SMOTE), which achieves class-distribution balance via (a) identifying the minority class, (b) adjusting instances to its closest neighbors, (c) interpolating data-point range between the minority-class instances and to its closest neighbors to create synthetic data-points, and (d) add the synthetic instances to original dataset to yield an oversampled, balanced dataset of both classes as in Figure 5. See [110] for more details.

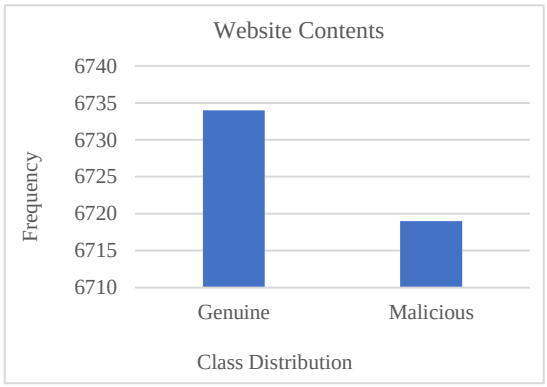

Figure 5. Dataset with SMOTE applied

3. SMOTE-Edited Nearest Neighbor (SMOTEEN) is a hybrid that combines features of over-sampling and under-sampling modes by identifying and linking data points to its closest neighbor(s) to address both issues of over/under-sampling via the actions of data cleaning [111]. SMOTEENN resample to create synthetic instances for a minority class (i.e., churn) and randomly removes from a majority class to resolve the dataset imbalance via the closest neighbor approach [112]. It generates new instances via the sampling ranges to its closest neighbor, balancing class distributions as in Figure 6. See [113] for more details.

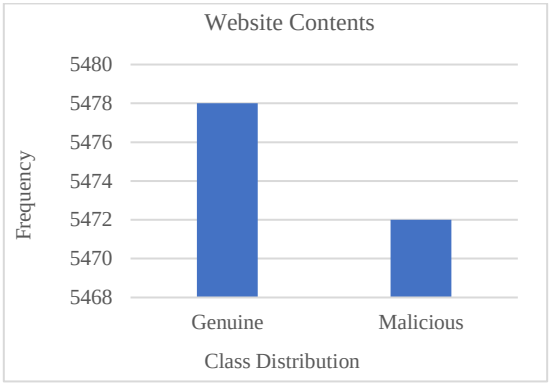

Figure 6. Dataset with SMOTEEN applied

3.2.3 Feats Extraction and Selection

It involves extracting the underlying feats to be considered and their respective formatting into estimation parameters to aid the effective classification of the texts. This helps a model to select relevant features from an expert perspective. Feature extraction: First, extract all the feats rippled across the dataset. Then, selection helps us to reduce the number of inputs, reduce the risk of overfitting, decrease computational complexity, fasten model construction time complexity, reduce training time, and improve model accuracy. We use vectorization and word embedding feature extraction schemes. See [114], [115] for details. We utilize the term frequency in reverse document frequency (TF-IDF) for feature selection. As we seek to uncover the underlying probabilities of interest feats, ML models do not understand characters/word tokens. But, they understand numbers as input. The impossibility of directly using the dataset's textual nature to interact directly with our ML causes us to use vectorization. Thus, we use TF-IDF 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. IDF aggregates the weight of the words against their appearances throughout the document. Conversely, the more a particular word appears, the smaller its IDF value with the transposed TF-IDF computed as in Equation (2) and (3), respectively

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

$$TF - IDF(d,k) = TF(d,k) * IDF(k) \tag{3}$$

where N is the number of words appearing in the document, DF is the frequency with which the word appears in the document, d is the document being considered, k is the word being considered in a particular document.

3.3. The Random Forest (RF) Model Training

Major merits of our choice of the Random Forest include: (a) each tree learns/votes to decide its outcome with equal weight, (b) it effectively handles complex datasets, (c) yields improved generalization, devoid of model overfit, (d) efficiently understand and reflect within a heuristic, relative contribution of feature selection to prediction performance, and (e) are resilient to noise in the quest for ground-truth even with (un)structured dataset for real-time case [101]. The RF is a widely used supervised model constructed from various decision trees. Its accuracy is achieved using majority voting, which combines the decisions of its weak tree into a single outcome. Its flexibility has necessitated adopting a voting scheme that assumes all its base learners have the same weight. It uses randomized bootstrap sampling to ensure that some trees will yield higher weights during iteration, though all trees have the same ability to make decisions. This helps it handle complex continuous and categorical datasets effectively, avoid overfitting, and mitigate poor generalization.

Detailed steps for adopting and adapting RF are expressed in [116] with parameters and hyper-parameter tuning as in Table 1. Afterward, we adopt data balancing and feature selection to fasten model construction and training. Data balancing helps to create

synthetic/artificial points in a minor class and cleans off unwanted data to resolve the imbalance in class distribution.

Table 1. Random Forest Model Construction with (Hyper-)Parameter Configuration

Features	Values	Details
n_estimators	150	Number of trees constructed
learning_rate	0.25	Step size learning for update
max_depth	5	Max depth of each tree
max_features	auto	Maximum number of features to construct the RF tree ensemble
min_sample_leaf	auto	Number of feats to be considered
min_sample_split	10	Minimal samples needed
min_weight_fraction_leaf	0.1	Tree's structure based on the weight assigned to each sample
random_state	25	The seeds for reproduction
eval_metric	error, logloss	Performance evaluation metrics
eval_set	x,val, y_val	Train data for evaluation
verbose	True	Determines if ensemble evaluation metric is printed at training
bootstrap	True	Ensures bootstrap aggregation use
warm_start	False	Ensure tree does not restart

The model learns from scratch using the identified data balancing techniques. It has been observed that the designated training set has all been expanded using RUS, SMOTE, and SMOTEEN data balancing techniques to include (i.e., for SMOTE and SMOTEEN) as well as exclude in RUS both the original and artificially created data points. We used iterative tree construction to create and adjust the RF trees. The model is trained via a bootstrap sample with a resampled subset for each tree to enhance training performance. This iterative process also enhances our trees' collective knowledge and helps to identify the intricate patterns in the phishing website. Thus, our training dataset is often a blend of synthetic and actual examples that guarantees a comprehensive learning experience for the proposed experimental RF model. This, will yield improved flexibility to a variety of settings for both models used on train/test datasets as well as in their inherent folds/partitions.

4. Results and Discussion

4.1. Model Performance

As mentioned in the previous section, we have already applied the preprocessing steps to the dataset. Such datasets are used to (a) rank images, (b) assess natural languages, and (c) metadata user assessment to infer semblance, characteristics, and feelings. In general, the objective is mining a dataset that provisions relations of various forms via the effective use of these cues [117], which will seduce the user to navigate compromised links, images, and other embedded objects. Table 2 shows the performance evaluation using feature selection and data balancing schemes.

Table 2. Performance Evaluation with Feature selection and 'with/without' Data Balancing

Balancing Modes	Without Data Balancing				With Data Balancing			
	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall
Default	0.7519	0.7969	0.8011	0.8011	0.9759	0.9718	0.8362	0.9282
RUS	0.8291	0.8302	0.8349	0.8492	0.9819	0.9947	0.9264	0.9557
SMOTE	0.8653	0.8763	0.8891	0.8918	0.9868	0.9970	0.9357	0.9645
SMOTEEN	0.8789	0.8802	0.8928	0.8944	0.9898	0.9973	0.9457	0.9698

Noting the impact and effects of the data balancing schemes using the 'before section' of Table 2 – it shows that the model yields F1 for the various data balancing techniques (i.e., default, RUS, SMOTE, and SMOTEEN) of 0.7519, 0.8291, 0.8653 and 0.8789 respectively;

It yields an accuracy (i.e., default, RUS, SMOTE and SMOTEEN) of 0.7969, 0.8302, 0.8763 and 0.8802 respectively; It yields a Recall (i.e., default, RUS, SMOTE and SMOTEEN) of 0.8011, 0.8492, 0.8918 and 0.8944 respectively; And yields a Precision (i.e., default, RUS, SMOTE and SMOTEEN) of 0.8011, 0.8349, 0.8891, and 0.8928 respectively.

Conversely, on the 'after section' of Table 2, the model yields F1 (i.e., default, RUS, SMOTE and SMOTEEN) of 0.9759, 0.9819, 0.9868 and 0.9898 respectively; Accuracy of 0.9718, 0.9947, 0.9970 and 0.9973 respectively; Recall of 0.9282, 0.9557, 0.9645 and 0.9698 respectively; And Precision of 0.8362, 0.9264, 0.9357 and 0.9457 (i.e., default, RUS, SMOTE and SMOTEEN) respectively. The study disagrees with [75] that SMOTE balancing outperforms others. Results show that SMOTEEN outperforms both SMOTE and RUS data balancing schemes [118]. It is also worth noting that the study agrees that the SMOTEEN data balancing scheme outperforms other modes of adopted data balancing techniques.

4.2. Discussion of Findings

The study investigates which data balancing technique has a greater influence on/to ground-truth truth and thus impacts overall performance by identifying important features that influence model prediction. It supports the effectiveness of differentiating between genuine (true) positive, true negative, genuine (false) positive [118], and false negative in an model's capability to classify test instances of the phishing dataset correctly.

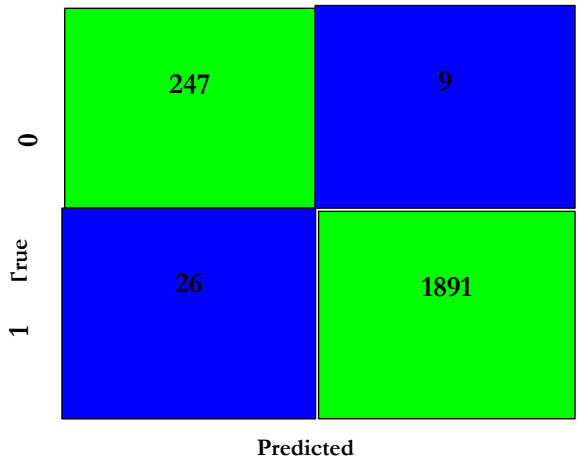

Figure 3. RF Confusion matrix using SMOTEEN

Figure 3 shows that the proposed RF model can correctly classify test data with over 99.73% accuracy, with only 18 incorrect classifications and 9,595 correctly classified test instances, which agrees with [119]. The model's best performance is with the SMOTEEN data balancing technique as a sampling method in combination with the chi-square feature selection scheme as adapted [11]. Overall, the RF model yields an F1 of 0.9898, an accuracy of 0.9973, a precision of 0.9457, and a recall of 0.9698, respectively [120].

The sample cue and lures include:

- 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

4.3. Comparison

As we explore the high performance of our proposed stacking ensemble across the domain dataset to demonstrate its flexibility, adaptability, robustness, and prediction ability – we also benchmark it against previous methods that have utilized the same or similar dataset. To this end – we found none. However, we decided to benchmark the proposed ensemble against similar design constructs on various datasets for various domain tasks, as seen in Table 3.

Table 3. Performance Evaluation with Feature selection and Data Balancing

Method/Metric	F1	Accuracy	Precision	Recall
Ref [58]	0.8728	0.8500	0.8120	0.8925
Ref [110]	1.0000	1.0000	0.9999	1.0000
Ref [116]	0.7815	0.7025	0.7372	0.7902
Ref [118]	0.7824	0.7631	0.7500	0.7732
Our Method	0.9981	0.9541	0.9881	0.9925

While some domain task datasets have proven much easier to detect/recognize and classify, others have also proven to be more painstaking [121]. Some domain task(s) such as medical and image records require their chosen ensemble design metric to be strongly impacted by the consequence of diagnostic errors within the captured dataset. Thus, the measure of both specificity and sensitivity becomes two critical feats to be evaluated since they are directly related to the patient clinical outcomes.

5. Conclusions

The goal here is to understand how users make trust decisions, identify user deficiencies, and adapt awareness capabilities to prevent victimization of a user vis-à-vis the associated network. With mixed malicious and normal web-contents scrapped from user social networking sites to simulate real-time interactions, we adopted user email, and social network accounts. Experiment scenarios a participant's response to phishing lures (i.e., clicking malicious web content) to engage a user (i.e., increase online presence). Simulation provides the participants with rich interaction capabilities that allow them to hover over links and attachments and see natural browser-like behaviour. The study employed a mix of qualitative and quantitative research design to increase the scope of understanding in light of collected and analyzed data. With data analysis, the qualitative data was used to support/authenticate quantitative conclusions. Social media networks have safeguards and rules to educate and protect users against phishing attempts. These often involve the capability to investigate and blacklist phishers if such cases are reported. The media and users are held accountable for preventing phishing attacks and their awareness. Social media platforms should inform users about phishing and give controls to prevent them. Conversely, users must stay ahead to dissuade such attacks; while implementing safety controls to limit such accidents.

Author Contributions: Conceptualization: A. Ojugo and V. Geteloma; Methodology: F. Aghware, M. Okpor., C. Odiakaose; Software: M. Akazue., A. Binitie and P. Ejeh; Validation: A. Ojugo and A. Eboka; Formal Analysis: V. Geteloma; Investigation: F. Aghware, M. Okpor and A. Binitie; Data Curation: C. Odiakaose; Writing—original draft preparation: F. Aghware and A. Ojugo Writing—review and editing: P.O. Ejeh and C.C. Odiakaose; Visualization: M.I. Akazue; Supervision: M. Okpor; Project administration: W. Adigwe; funding acquisition: All.

Funding: This research received no external funding.

Data Availability Statement: Data was retrieved from UCI available online at https://archive.ics.uci.edu/dataset/379/website+phishing.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] 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.

[2] 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.

[3] B. O. Malasowe, M. I. Akazue, A. E. 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.

[4] F. U. Emordi, C. C. Odiakaose, P. O. Ejeh, O. Attoh, and N. C. Ashioba, "Student's Perception and Assessment of the Dennis Osadebay University Asaba Website for Academic Information Retrieval, Improved Web Presence, Footprints and Usability," *FUPRE J. Sci. Ind. Res.*, vol. 7, no. 3, pp. 49–60, 2023.

[5] 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.

[6] S. F. Tan and G. C. Chung, "An Evaluation Study of User Authentication in the Malaysian FinTech Industry With uAuth Security Analytics Framework," *J. Cases Inf. Technol.*, vol. 25, no. 1, pp. 1–27, 2023, doi: 10.4018/JCIT.318703.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] S. Chiemeke and E. Omede, "Mal-typho diagnosis intelligent system (MATDIS): the auto-diagnostic rule generation algorithm," *Comput. Inf. Syst. Dev. Informatics Allied Res. J.*, vol. 5, no. 4, pp. 83–92, 2021.

[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] M. Dadkhah, T. Sutikno, J. M. Davarpanah, and D. Stiawan, "An Introduction to Journal Phishings and Their Detection Approach," *TELKOMNIKA*, vol. 13, no. 2, p. 373, Jun. 2015, doi: 10.12928/telkomnika.v13i2.1436.

[16] 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.

[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. Broadhurst, K. Skinner, N. Sifniotis, and B. Matamoros-Macias, "Cybercrime Risks in a University Student Community," *SSRN Electron. J.*, no. May, 2018, doi: 10.2139/ssrn.3176319.

[20] 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.

[21] 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.

[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] 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.

[24] S. E. Brizimor *et al.*, "WiSeCart: Sensor-based Smart-Cart with Self-Payment Mode to Improve Shopping Experience and Inventory Management," *Adv. Multidiscip. Sci. Res. J. Publ.*, vol. 10, no. 1, pp. 53–74, Mar. 2024, doi: 10.22624/AIMS/SIJ/V10N1P7.

[25] P. O. Ejeh *et al.*, "Counterfeit Drugs Detection in the Nigeria Pharma-Chain via Enhanced Blockchain-based Mobile Authentication Service," *Adv. Multidiscip. Sci. Res. J. Publ.*, vol. 12, no. 2, pp. 25–44, 2024, doi: 10.22624/AIMS/MATHS/V12N2P3.

[26] 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.

[27] P. O. Ejeh, E. Adishi, E. Okoro, and A. Jisu, "Hybrid integration of organizational honeypot to aid data integration, protection and organizational resources and dissuade insider threat," *FUPRE J. Sci. Ind. Res.*, vol. 6, no. 3, pp. 80–94, 2022.

[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] 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.

[31] 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.

[32] 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.

[33] D. R. I. M. Setiadi, A. Susanto, K. Nugroho, A. R. Muslikh, A. A. Ojugo, and H. Gan, "Rice yield forecasting using hybrid quantum deep learning model," *MDPI Comput.*, vol. 13, no. 191, pp. 1–18, 2024, doi: 10.3390/computers13080191.

[34] 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.

[35] M. A. Haque *et al.*, "Cybersecurity in Universities: An Evaluation Model," *SN Comput. Sci.*, vol. 4, no. 5, 2023, doi: 10.1007/s42979-023-01984-x.

[36] O. V. Lee *et al.*, "A malicious URLs detection system using optimization and machine learning classifiers," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 17, no. 3, p. 1210, Mar. 2020, doi: 10.11591/ijeecs.v17.i3.pp1210-1214.

[37] 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.

[38] 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.

[39] 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.

[40] D. R. I. M. Setiadi, S. Widiono, A. N. Safriandono, and S. Budi, "Phishing Website Detection Using Bidirectional Gated Recurrent Unit Model and Feature Selection," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 75–83, Jul. 2024, doi: 10.62411/faith.2024-15.

[41] R. E. Ako *et al.*, "Effects of Data Resampling on Predicting Customer Churn via a Comparative Tree-based Random Forest and XGBoost," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 86–101, Jun. 2024, doi: 10.62411/jcta.10562.

[42] E. O. Yeboah-Boateng and P. M. Amanor, "Phishing , SMiShing & Vishing : An Assessment of Threats against Mobile Devices," *J. Emerg. Trends Comput. Inf. Sci.*, vol. 5, no. 4, pp. 297–307, 2014.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] F. O. Aghware *et al.*, "BloFoPASS: A blockchain food palliatives tracer support system for resolving welfare distribution crisis in Nigeria," *Int. J. Informatics Commun. Technol.*, vol. 13, no. 2, p. 178, Aug. 2024, doi: 10.11591/ijict.v13i2.pp178-187.

[49] 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.

[50] 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.

[51] M. Akazue, C. Asuai, A. Edje, E. Omede, and E. Ufiofio, "Cybershield : Harnessing Ensemble Feature Selection Technique for Robust Distributed Denial of Service Attacks Detection," *Kongzhi yu Juece/Control Decis.*, vol. 38, no. 03, pp. 1211–1224, 2023.

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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

[60] N. C. Ashioba *et al.*, "Empirical Evidence for Rainfall Runoff in Southern Nigeria Using a Hybrid Ensemble Machine Learning Approach," *J. Adv. Math. Comput. Sci.*, vol. 12, no. 1, pp. 73–86, 2024, doi: 10.22624/AIMS/MATHS/V12N1P6.

[61] 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.

[62] 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.

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

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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

[76] M. I. Akazue *et al.*, "Handling Transactional Data Features via Associative Rule Mining for Mobile Online Shopping Platforms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 3, pp. 530–538, 2024, doi: 10.14569/IJACSA.2024.0150354.

[77] 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.

[78] 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.

[79] 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.

[80] E. U. Omede, A. E. 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. 1, no. 3, pp. 273–283, Feb. 2024, doi: 10.62411/jcta.9541.

[81] A. A. Ojugo, A. O. Eboka, R. E. Yoro, M. O. Yerokun, and F. N. Efozia, "Framework design for statistical fraud detection," *Math. Comput. Sci. Eng. Ser.*, vol. 50, pp. 176–182, 2015.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] E. A. Otorokpo *et al.*, "DaBO-BoostE: Enhanced Data Balancing via Oversampling Technique for a Boosting Ensemble in Card-Fraud Detection," *Adv. Multidiscip. Sci. Res. J. Publ.*, vol. 12, no. 2, pp. 45–66, 2024, doi: 10.22624/AIMS/MATHS/V12N2P4.

[92] M. I. Akazue *et al.*, "FiMoDeAL: pilot study on shortest path heuristics in wireless sensor network for fire detection and alert ensemble," *Bull. Electr. Eng. Informatics*, vol. 13, no. 5, pp. 3534–3543, Oct. 2024, doi: 10.11591/eei.v13i5.8084.

[93] R. R. Atuduhor *et al.*, "StreamBoostE: A Hybrid Boosting-Collaborative Filter Scheme for Adaptive User-Item Recommender for Streaming Services," *Adv. Multidiscip. Sci. Res. J. Publ.*, vol. 10, no. 2, pp. 89–106, Jun. 2024, doi: 10.22624/AIMS/V10N2P8.

[94] 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.

[95] I. Vågsholm, N. S. Arzoomand, and S. Boqvist, "Food Security, Safety, and Sustainability—Getting the Trade-Offs Right," *Front. Sustain. Food Syst.*, vol. 4, no. February, pp. 1–14, 2020, doi: 10.3389/fsufs.2020.00016.

[96] 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.

[97] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019," *Comput. Sci. Rev.*, vol. 40, p. 100402, May 2021, doi: 10.1016/j.cosrev.2021.100402.

[98] S. N. Okofu *et al.*, "Pilot Study on Consumer Preference, Intentions and Trust on Purchasing-Pattern for Online Virtual Shops," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 7, pp. 804–811, 2024, doi: 10.14569/IJACSA.2024.0150780.

[99] D. A. Obasuyi *et al.*, "NiCuSBlockIoT: Sensor-based Cargo Assets Management and Traceability Blockchain Support for Nigerian Custom Services," *Adv. Multidiscip. Sci. Res. J. Publ.*, vol. 15, no. 2, pp. 45–64, Jun. 2024, doi: 10.22624/AIMS/CISDI/V15N2P4.

[100] A. M. Ifioko *et al.*, "CoDuBoTeSS: A Pilot Study to Eradicate Counterfeit Drugs via a Blockchain Tracer Support System on the Nigerian Frontier," *J. Behav. Informatics, Digit. Humanit. Dev. Res.*, vol. 10, no. 2, pp. 53–74, 2024, doi: 10.22624/AIMS/BHI/V10N2P6.

[101] A. A. Ojugo, P. O. Ejeh, C. C. Odiakaose, A. O. Eboka, and F. U. Emordi, "Predicting rainfall runoff in Southern Nigeria using a fused hybrid deep learning ensemble," *Int. J. Informatics Commun. Technol.*, vol. 13, no. 1, p. 108, Apr. 2024, doi: 10.11591/ijict.v13i1.pp108-115.

[102] 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.

[103] 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.

[104] A. A. Ojugo and O. Nwankwo, "Spectral-Cluster Solution For Credit-Card Fraud Detection Using A Genetic Algorithm Trained Modular Deep Learning Neural Network," *JINAV J. Inf. Vis.*, vol. 2, no. 1, pp. 15–24, Jan. 2021, doi: 10.35877/454RI.jinav274.

[105] 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.

[106] 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.

[107] M. S. Sunarjo, H.-S. Gan, and D. R. I. M. Setiadi, "High-Performance Convolutional Neural Network Model to Identify COVID-19 in Medical Images," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 19–30, 2023, doi: 10.33633/jcta.v1i1.8936.

[108] B. O. Malasowe, D. V. Ojie, A. A. Ojugo, and M. D. Okpor, "Co-infection prevalence of Covid-19 underlying tuberculosis disease using a susceptible infect clustering Bayes Network," *Dutse J. Pure Appl. Sci.*, vol. 10, no. 2a, pp. 80–94, Jul. 2024, doi: 10.4314/dujopas.v10i2a.8.

[109] 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. 1, no. 4, pp. 407–420, Mar. 2024, doi: 10.62411/jcta.10323.

[110] D. R. I. M. Setiadi, K. Nugroho, A. R. Muslikh, S. W. Iriananda, and A. A. Ojugo, "Integrating SMOTE-Tomek and Fusion Learning with XGBoost Meta-Learner for Robust Diabetes Recognition," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 23–38, May 2024, doi: 10.62411/faith.2024-11.

[111] B. Pavlyshenko and M. Stasiuk, "Data augmentation in text classification with multiple categories," *Electron. Inf. Technol.*, vol. 25, p. 749, 2024, doi: 10.30970/eli.25.6.

[112] A. A. Ojugo and R. E. Yoro, "Empirical Solution For An Optimized Machine Learning Framework For Anomaly-Based Network Intrusion Detection," *Technol. Rep. Kansai Univ.*, vol. 62, no. 08, pp. 6353–6364, 2020.

[113] M. D. Okpor *et al.*, "Comparative Data Resample to Predict Subscription Services Attrition Using Tree-based Ensembles," *J. Fuzzy Syst. Control*, vol. 2, no. 2, pp. 117–128, 2024, doi: 10.59247/jfsc.v2i2.213.

[114] 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.

[115] 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.

[116] 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.

[117] 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.

[118] E. B. Wijayanti, D. R. I. M. Setiadi, and B. H. Setyoko, "Dataset Analysis and Feature Characteristics to Predict Rice Production based on eXtreme Gradient Boosting," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 79–90, 2024, doi: 10.62411/jcta.10057.

[119] Y. Bouchlaghem, Y. Akhiat, and S. Amjad, "Feature Selection: A Review and Comparative Study," *E3S Web Conf.*, vol. 351, pp. 1–6, 2022, doi: 10.1051/e3sconf/202235101046.

[120] 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.

[121] A. N. Safriandono, D. R. I. M. Setiadi, A. Dahlan, F. Z. Rahmanti, I. S. Wibisono, and A. A. Ojugo, "Analyzing Quantum Feature Engineering and Balancing Strategies Effect on Liver Disease Classification," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 51–63, Jun. 2024, doi: 10.62411/faith.2024-12.