Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

Hypertension Detection via Tree-Based Stack Ensemble with SMOTE-Tomek Data Balance and XGBoost Meta-Learner

Christopher Chukwufunaya Odiakaose [1,*], Fidelis Obukohwo Aghware [2], Margareth Dumebi Okpor [3], Andrew Okonji Eboka [4], Amaka Patience Binitie [4], Arnold Adimabua Ojugo [5], De Rosal Ignatius Moses Setiadi [6], Ayei Egu Ibor [7], Rita Erhovwo Ako [5], Victor Ochuko Geteloma [5], Eferhire Valentine Ugbotu [8], and Tabitha Chukwudi Aghaunor [9]

[1] Department of Computer Science, Dennis Osadebay University Asaba, Nigeria; e-mail : osegalaxy@gmail.com
[2] Department of Computer Science, University of Delta Agbor, Nigeria; e-mail : fidelis.aghware@unidel.edu.ng
[3] Department of Cybersecurity, Delta State University of Science and Technology Ozoro, Nigeria; e-mail : okpormd@dsust.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 : ojugo.arnold@fupre.edu.ng; ako.rita@fupre.edu.ng; geteloma.victor@fupre.edu.ng
[6] Department of Informatic Engineering, Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia; e-mail : moses@dsn.dinus.ac.id
[7] Trustworthy Digital Infrastructure for Identity Systems, The Alan Turing Institute, United Kingdom; e-mail : aibor@turing.ac.uk
[8] Department of Data Science, University of Salford, Manchester, United Kingdom; e-mail : ugbotueferirhe@gmail.com
[9] Department of Data Intelligence and Technology, Robert Morris University, Pittsburg, Pennsylvania, United States; e-mail : cxast461@rmu.edu
* Corresponding Author: Christopher Chukwufunaya Odiakaose

Abstract: High blood pressure (or hypertension) is a causative disorder to a plethora of other ailments – as it succinctly masks other ailments, making them difficult to diagnose and manage with a targeted treatment plan effectively. While some patients living with elevated high blood pressure can effectively manage their condition via adjusted lifestyle and monitoring with follow-up treatments, Others in self-denial leads to unreported instances, mishandled cases, and in now rampant cases – result in death. Even with the usage of machine learning schemes in medicine, two (2) significant issues abound, namely: (a) utilization of dataset in the construction of the model, which often yields non-perfect scores, and (b) the exploration of complex deep learning models have yielded improved accuracy, which often requires large dataset. To curb these issues, our study explores the tree-based stacking ensemble with Decision tree, Adaptive Boosting, and Random Forest (base learners) while we explore the XGBoost as a meta-learner. With the Kaggle dataset as retrieved, our stacking ensemble yields a prediction accuracy of 1.00 and an F1-score of 1.00 that effectively correctly classified all instances of the test dataset.

Keywords: Cardiovascular disease; Hypertension; Meta-Learner; Stacked Ensemble; Stroke; XGBoost.

Received: October, 15th 2024
Revised: November, 23rd 2024
Accepted: November, 30th 2024
Published: December, 1st 2024
Curr. Ver.: December, 1st 2024

1. Introduction

There is today, the inherent rise in the trend of sudden collapse that morphs onto death in Nigeria [1]. Attributed to this anomaly is the uncontrolled rise in patients' blood pressure (HBP) causing hypertension for which signs abound. Globally, hypertension has contributed to the increased morbidity and mortality rate [2], [3] – and has become a global health menace. The World Health Organization has since acknowledged HBP as the leading, health risk

DOI : 10.62411/faith.3048-3719-43

predictor for cardiovascular disease and premature death [4], [5]. Its asymptomatic nature makes it a vicious-silent killer, as it masks another ailment in over 40 age years of patients with a rise in cases of delayed diagnosis and improper management. This calls for continued monitoring of patients' vitals, with reported HBP-associated deaths in Nigeria to over 323,400 cases in 2023 [6], [7]. The early detection of HBP its accurate classification vis-à-vis the consequent effective monitoring thereof – has become imperative to help mitigate the continuous rise in cardiovascular diseases, stroke, and death. Ailments such as diabetes and HBP – are non-communicable and quite easy to diagnose. While some patients manage theirs via lifestyle adjustments, others self-deny their state, And such cases are unreported/unmonitored with treatment plans [8]. Legacy identification of HBP relies on cuffed-intermittent readings at clinical visits that yield observations, which does not provide a comprehensive patient health-state layout [9], [10]. Thus, continued monitoring is advised to aid in effective management and avoid ripple effects across other underlying ailment(s).

Proactive patients harness sensor-unit observed readings to gain valued insights into their status and with anomaly detection for an early prompt to quick medical intervention. Healthcare experts also glean valued predictive knowledge tailored towards targeted treatment for hypertensive patients [11]. Various classification schemes have been utilized to improve performance generalization accuracy. Widely used hypertension dataset includes Kaggle [12], [13], Data. world [14], DASH [15], [16], NCD Risk Factor[17], and PIMA Indian [18], [19]. Each classification heuristic yielded varying performance ranging from 0.67-to-0.89 accuracy with each of the datasets adopted. Thus, the input dataset can effectively impact the generalization of the classification method adopted. However, a related study [20] adapted fusion learning with XGBoost meta-learner using SMOTE-Tomek (synthetic minority over-sample technique) data balancing, yielding a perfect prediction accuracy. This also implies that using an increasingly sophisticated identification scheme can yield improved generalization performance. Thus, it has become imperative and crucial to develop a classification heuristic that proffers a more sensitive, adaptive and robustness to intrinsic dataset variations to yield improved/perfect performance accuracy.

The classification task is often accomplished via voting, stacking, and boosting. On these, detection tasks in a more general term can be grouped into three (3) categories: Deep learning (DL) [21], Ensemble Learning (EL)[22], and machine learning (ML)[23], [24]. ML offers a heuristic range that can be successfully trained to recognize evidence that supports ground truth in high-dimension tasks, even with complex datasets [25]. Their adaptive learning and flexibility ensure they effectively decipher intrinsic crucial parameters to be selected for model construction to ease outlier detection from behavioral norms of data labels [26]–[28]. Various ML models include Logistic Regression [29], [30], Naïve Bayes [31], Support Vector [32], [33], K-Nearest Neighbor [34]. DL heuristics leans on recurrent networks to capture high-dimension feats in time-series data sequences. It is common in many complex, chaotic, and non-linear spatiotemporal and medical datasets [35]. By default – RNN heuristics are best suited for medical datasets. Its demerit is that they often yield poor generalization in vanishing gradient tasks. To resolve this, some studies explore its variant, Long-Short-Term Memory (LSTM), via gates that control data flow so that the model can learn and easily adapt to minor changes experienced as long-term dependencies [36]. A demerit of the LSTM is that their efficiency requires longer training time and large datasets. To curb the challenges with DL, the EL mode can effectively combine both ML and DL [36] into a single classifier as its meta-learner to yield an optimal solution with lesser training time, irrespective of the volume and veracity of the dataset. It achieves this feat via a variety of schemes like stack [22], [37], bagging [38], [39], boosting [40], [41], and voting [42], [43] – to yield a richer insight into the targeted task domain with optimality for ground-truth [44].

2. Preliminaries

2.1. Tree-Based Classifiers and Algorithms

A common scheme in ML is the tree method, which descends from single decision trees. Each tree generates a set of if-else rules used in the majority voting scheme, allowing it to predict observed classes [45]. Each tree is a recursive top-down model in which a binary tree is used to partition its predictor space with variables grouped into subsets for which the distribution of dependent variable y is successively homogeneous [46]. A tree is easily understood, but its use alone leads to model overfit as it seeks to identify feats of interest during training [47]. Thus, it yields degraded performance in classifying of unknown labels. Tree-based models learn by constructing many individually trained decision trees and combining/aggregating their results into a single and more robust model whose output outperforms the results of any single tree [48]. It achieves this via either bagging [49], [50], and boosting [51], [52] modes.

With boosting – the tree(s) achieves its accuracy beyond random with enhanced predictive capability by sequentially training each weak learner to correct the weakness in its predecessor [53], [54]. Each tree yields feedback from previous weaker trees [55]. Popular boost models include gradient boost [56], LogitBoost [57], stochastic gradient [58], and adaptive boost [59]. As expressed in Equation (1) – prediction is achieved by combining the outcome of its weak learners with its weighted sum to yield a higher weight for incorrectly classified cases.

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

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 [60]. 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 each node (best) split among all predictor variables – Random forest nodes are split using the best from a subset of predictors randomly chosen at the node. Its recursive structure captures interaction effects between variables. Thus, tree-based models have proven successful for various tasks ranging from churn prediction[61] to user purchase intent prediction [62]. They are best suited to reduce bias and variance within learning schemes. While individual models may be stuck in local minima [63], its weighted combination of varied local minima will yield an ensemble method capable of minimizing the risk of choosing the wrong local minimum with such tree-based models/heuristics [64].

2.2. Stacked-Learning Ensembles

For stacking – it trains a meta/higher learner to effectively combine the predictive outcome of several learners, allowing its meta-learner to improve as it learns from the errors of its base classifiers. This flexibility ensures stacking yields better outcomes with more iterations [65]. In voting, the learners are applied independently to achieve a more stable performance with reduced overfit via predictive aggregation in its quest for ground truth [66]. Since it seeks to combine only the final output of all learners without recourse to their predictive relations – in some cases, it yields degraded performance due to its dataset complexity and diversity [42]. Bagging trains similar learners with equal voting weight. To promote variance, each learner is trained using a randomly drawn subset of the train data [67], [68]. It achieves higher accuracy by averaging all learners' predictions. It can be configured to use various learners on different datasets to reduce each learner's variance error(s) [69]. Lastly, boosting sequentially trains its learners so that each new model corrects the errors of its previous model. It yields a series of learners that focus on difficult tasks that their predecessors failed to predict [70], [71] correctly, resulting in higher generalization with improved accuracy. A typical boost is an adaptive boost. However, to improve this scheme, an ensemble can be based on a gradient scale such as the eXtreme Gradient Boost [72].

2.3. Study Motivation

Inherent gaps from existing studies include:

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 [73], [74].
2. Imbalanced Datasets: A critical issue with the dataset's imbalanced nature is that many HBP cases go unreported. In addition, medical records often yield large datasets where HBP cases lag in class distribution as it is often an indicator of ailments such as diabetes [75], [76], mental disorders [77], etc. New studies must explore intricate sampling

techniques or harness the robust power of ensemble methods tailored explicitly to mitigate the challenges with imbalanced datasets.

3. Data Acquisition: The rise in volume and integrity of datasets generated across multiple channels [78] using various methods [79] implies that new heuristics must become more adaptive, sensitive, and robust as analytic modes to glean insightful knowledge from such huge datasets. New models must account for minor shocks as heuristics that utilize and integrate means to harness these data generation points and enhance overall model accuracy and performance, as multi-channel detection has become a critical area for research purposes and with business monetization focus [80] in mind.

Thus, we construct known tree-based models using bagging and boosting capabilities with data normalization 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: (a) model construction as a decision support model via the utilization of ML scheme that will help effectively capture predictor features as factors that makes classification of HBP, more successful, (b) data balancing effect on the reliability and predictive power of tree-based models; while, analyzing its impact cum implication to predict HBP [81], [82]accurately, and (c) comparative benchmarking will yield the evaluation of diverse machine learning schemes within the constructed prediction model, aimed at comparing the performance, accuracy, and robustness of various algorithms to identify sophisticated, dynamic cases vis-à-vis underlying health factors.

3. Proposed Method

The development of ML in the identification of HBP leans on these steps: (a) extensive study via the body of knowledge to identify HBP as a problem, (b) dataset collection consisting of observed relevant physiological, lifestyle, and demographic variables as predictor factors, (c) preprocessing handles missing data, duplicates, and normalizes data to ensure consistency, and feature selection to aid practical model construction, (d) model construction and training to yield a qualitative and reliable predictive heuristic classifier, and (e) utilization of proposed scheme. Thus, the adopted methodology yields Figure 1 – explained thus:

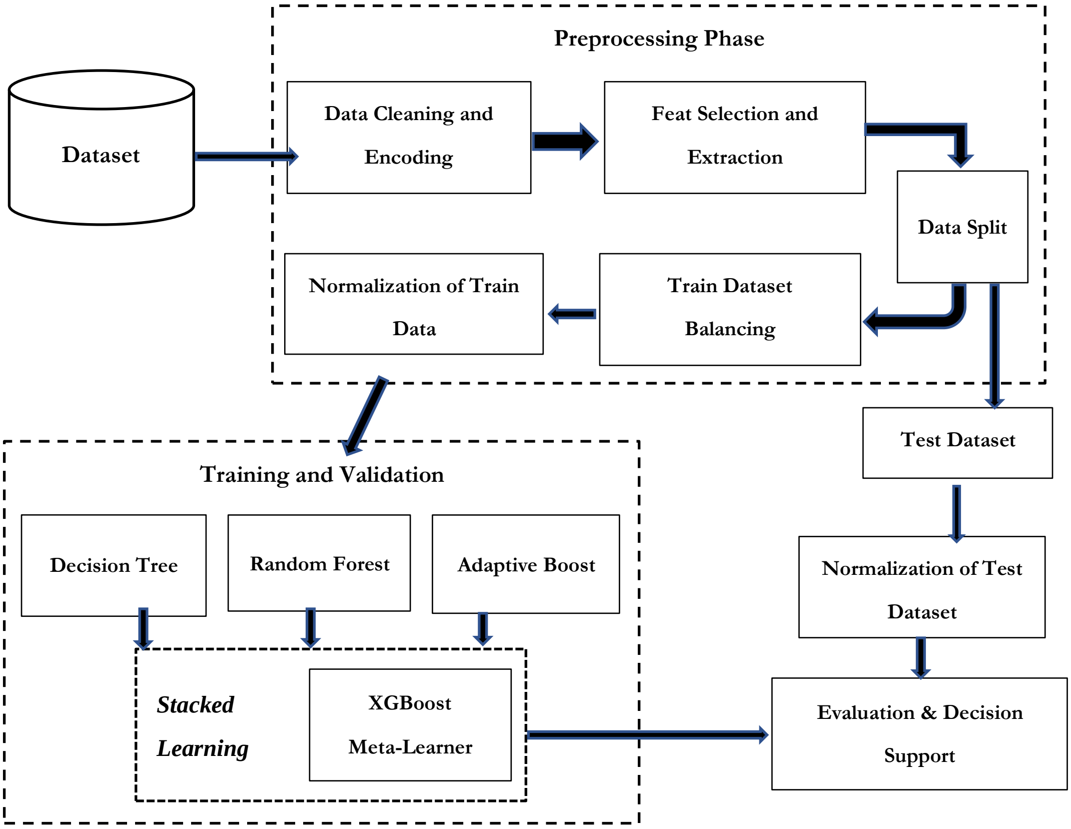


Figure 1. Proposed Stacking Ensemble with XGBoost meta-learner

3.1. Data Collection, Cleaning and Encoding:

This is achieved thus:

- **Collection:** Data is retrieved from Kaggle HBP and is available [web]: www.kaggle.com/datasets/jayaprakashpondy/blood-pressure. Figure 2 shows the detailed class distribution for Normal, Pre-HBP, Stage-1, and Stage-2 hypertension cases.

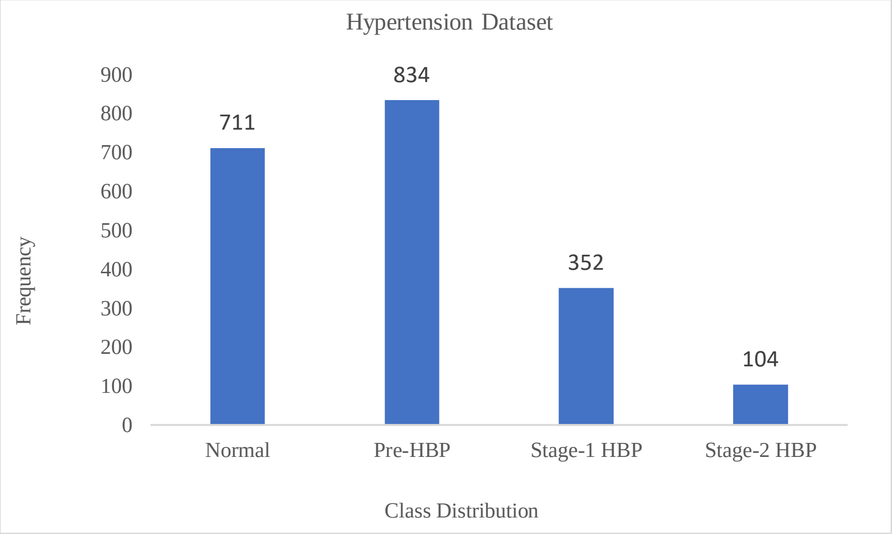


Figure 2. Dataset Distribution by Class

- **Data Encoding**: Rather than utilize the principal component analysis [83] to encode our dataset – we utilize the one-hot encoding technique, which converts categorical variables into a format that ML algorithms can understand [84]– because many ML schemes cannot handle category labels directly. One-hot encoding creates a binary numeric equivalence of the dataset by converting categorical variables into their binary forms.
- **Clean** removes duplicates to avoid redundancy and missing values to ensure data quality.

3.2. Feature Selection and Extraction

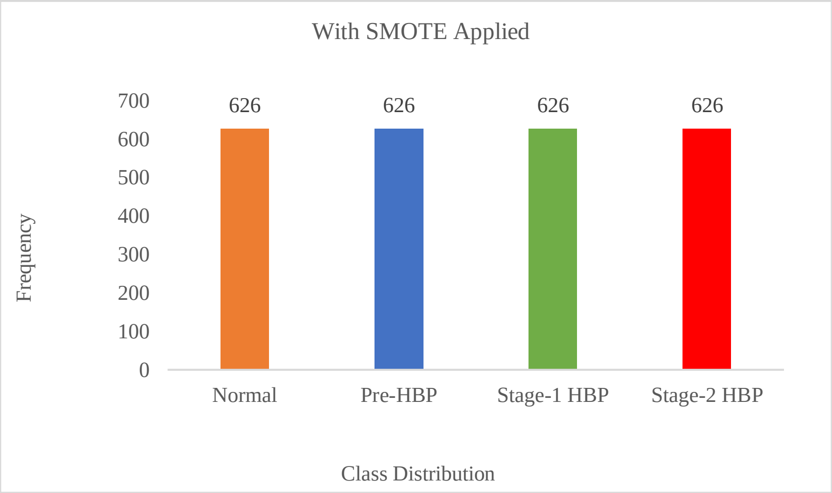
This stage selects and extracts the input data and determines what label will yield ensemble output (Y). It removes all docile cum irrelevant features with no relative importance to our quest for ground truth (i.e., target class). And in turn, reduces the dimensionality of the chosen dataset [85] vis-a-vis and fastens the ensemble's construction for enhanced/improved performance [86], [87], especially in cases where cost is a critical factor [88]. The efficiency of a selected feature is evaluated on how well the model fits [89] to ground truth (i.e., target class) [90]. Thus, we use the recursive feat elimination (RFE) [91] wrapper approach since our feats are engineered to unveil how relevant a selected feature supports our target class and test via frequency distribution to ascertain how its occurrence fits with the target class [92], [93]. With a computed RFE threshold value of 0.8991 – 11 features were selected from the original dataset instead of ground truth or target class 1, as in Table 1. These were examined instead of their contribution to the classification process [94].

3.3. Split and Balancing

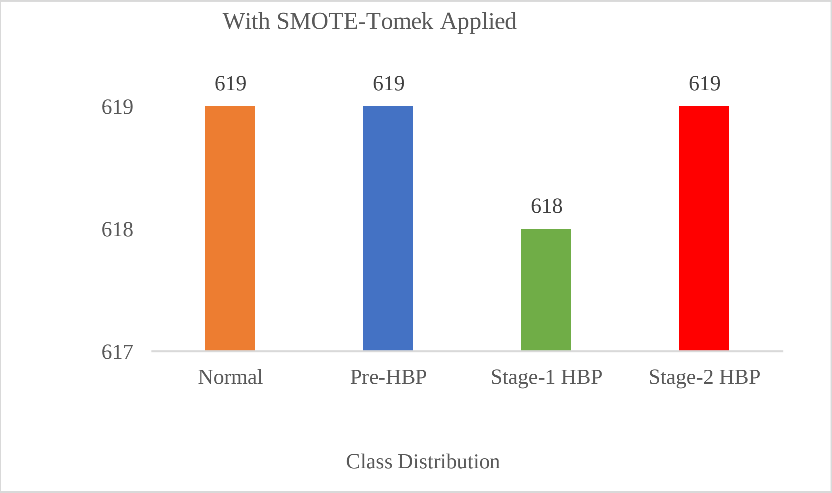
To track each feature towards our target class, the dataset is split into train (75%) and test (25%) subsets for this study. There is no rule on how they can be grouped – so we adopt this rule of thumb. Data balancing seeks to nearly (and evenly) redistribute data points in the training dataset to ensure an almost equal distribution between major and minor classes. While there are a variety of modes, We adopt SMOTE-Tomek thus [20]: (a) identifies major class, (b) interpolates to create synthetic data via the Tomek-link under-sample mode for the majority class, (c) adjusts data points to those of its closest neighbors so that new data-points overlaps, and (c) adds the generated synthetic data to the original dataset to yield a balanced dataset [95] as in Figure 3.

Table 1. Wrapper RFE Feature Selection for Kaggle Hypertension Dataset

Features	Data Type	Sample	RFE Score	Selected
patient_number	Int	Yes -1 /No – 0	0.2805	No
blood_pressure_abnormality	Binary	Yes -1 /No – 0	0.9318	Yes
hemoglobin_level	Float	11.28	0.9016	Yes
genetic_pedigree_coefficient	Float	0.9	0.8743	Yes
age	Int	34	0.9291	Yes
body_mass_indexi	Int	23	0.8659	Yes
sex	Binary	Male -1, Female – 0	0.5391	No
pregnancy	Binary	Yes -1 /No – 0	0.3528	No
smoking	Binary	Yes/No	0.5241	No
physical_activity	Int	9995	0.9805	Yes
salt_content_in_diet	Int	9607	0.9318	Yes
daily_alcohol_consumed	Int	205	0.9016	Yes
stress_level	Int	3	0.9732	Yes
chronic_kidney_disease	Binary	Yes -1 /No – 0	0.9291	Yes
adrenal_thyroid_disorder	Binary	Yes -1 /No – 0	0.9956	Yes



(a)



(b)

Figure 3. Applied Balancing on training dataset using (**a**) SMOTE only; (**b**) SMOTE-Tomek Links.

3.4. Normalization

Used for variable transformation to normalize skewed datasets. This seeks to ensure a nearness in the class distribution and may change our data distribution. Features are normalized via a standard scaler, which seeks to revert data features to yield a distribution with a mean value of 0 and a standard deviation of 1. We achieve this via Equation (2)– where x is the original value, μ is the mean, σ is the standard deviation, and z is our normalization process.

$$z = \frac{(x - \mu)}{\sigma} \tag{2}$$

3.5. Stacked Ensemble Construction

Stacked learning combines the predictive outcome of several base learners to acquire or achieve a more accurate prediction. It often involves 2-levels for which the first level consists of base learners (in this case, Adaptive Boosting, Decision trees, and Random Forest), and the second level aggregates the predictions of the first-level learners, usually called a meta-heuristic/learner (in this case, the XGBoost). Its major merits include (a) the diversification of models via the use of several algorithms [96], (b) enhanced generalization for the model, and (c) reduced risk in the overfitting of the ensemble [97]. Selecting a meta-learner is critical and crucial as they must optimize aggregated outputs and efficiently minimize prediction errors. The right meta-learner (especially for one) trained using the out-of-fold prediction from the base classifiers can significantly improve the ensemble accuracy, flexibility, and robustness – effectively harnessing the processing prowess of multiple good-fit base classifiers [98]. Thus, the utilized tree-based algorithms/classifiers include:

- XGBoost Meta-Learner is a tree-based leaner that scales the gradient boosting to classify data points. It yields a more robust classifier by aggregating its weaker (base) learner tree via majority voting schemes over a series of iterations on data points to yield an optimal fit solution. It expands its goal function by minimizing its loss function as Equation (3) to yield an improved model [99] to manage tree complexity effectively. For optimality, the XGBoost leverages the predictive power of weak base learners to yield a better decision tree with each iteration and account for weak performance, contributing to its knowledge of the task. Thus, with each tree trained on the candidate data, it expands the objective function via a regularization term $\Omega(f_t)$ and loss function $l\left(Y_i^t, \widehat{Y}_i^{t-1}\right)$to ensure an appropriate fit of the ensemble to yield improved generalization. This, ensures that both training dataset fits as re-calibrated solutions to remain within their solution's set boundaries and tunes its loss function for higher accuracy [100] configuration design as in Table 2.

$$L^t = \sum_{i=1}^{n} l\left(Y_i^t, \widehat{Y}_i^{t-1} + f_k(x_i)\right) + \Omega(f_t) \tag{3}$$

Table 2. The Extreme Gradient Boosting Ensemble Design Configuration

Configuration	Values	Description
n_estimators	250	Number of trees constructed
learning_rate	0.25	Step size learning to update the ensemble
max_depth	5	Max depth of each tree
random_state	25	The seeds for reproduction
eval_metric	["error', 'logloss']	Performance evaluation metrics
eval_set	(x,val, y_val)	Train dataset to evaluate performance
verbose	True	Determines if ensemble evaluation metric is printed at training

- A Decision Tree is a single-classifier that explores intricate sampling, tailored to mitigate the decision-making issues. To predict a target class, it starts from its root node to compare the root values with the records attribute. With this compared, it branches off to the next node as (a) begins at a tree with root node S that consists of a complete dataset, (b) finds the best attribute in the dataset via attribute selection measure, (c) divides S into train/test sub-datasets that contains possible values for the best attributes, (d) generate decision tree node, which contains best attributes, and (e) recursively make new decision trees using the subset of the dataset created [101]. Then, continue this process until the criterion for optimal solution is reached so that the tree can no longer classify the nodes. The leaf node reaches such through error pruning and/or cost-complex pruning. Its demerits are: (i) it is complex due to its many layers, (ii) it may result in overfit, resolved via a Random Forest ensemble, and (c) computational complexity increases for large datasets. Furthermore, its merits are numerous, and the feats used in our DT construction are shown in Table 3.

Table 3. The Decision Tree Classifier Design Configuration

Configuration	Values	Description
info_gain	120	Number of trees constructed
learning_rate	0.25	Step size learning to update the ensemble
min_sample_split	10	Minimal number of samples needed
min_sample_leaf	auto	Number of features to be considered in place of ground-truth
eval_set	(x,val, y_val)	Dataset used for evaluating ensemble performance at training
min_weight_fraction_leaf	0.1	Determines tree's structure based on the weight assigned to samples
max_depth	5	Max depth of each tree
random_state	25	The seeds for reproduction

- Adaptive Boosting combines multiple weak classifiers to build a strong one. Weak learners are called decision stumps, as they are DTs with a single split. The ensemble places more weight on hard-to-classify instances and less on data operating well. Stumps are produced for each feature iteratively and stored in a list until a lower error is received. The weight (s) assigned to each example determines its significance in the training dataset. Weights are updated with each iteration to yield stumps' performance. Ensemble sequentially trains its predictors so that each predictor tries to correct its predecessor [102]. Thus, they are more robust against overfitting and yield a more stable and improved performance. Table 4 shows the configuration therein.

Table 4. Adaptive Boosting Ensemble Design Configuration

Configuration	Values	Description
n_estimators	140	Number of trees constructed
learning_rate	0.25	Step size learning to update the ensemble
max_depth	5	Max depth of each tree
random_state	25	The seeds for reproduction
eval_metric	["error', 'logloss']	Evaluation metrics for ensemble performance
eval_set	(x,val, y_val)	Dataset used for evaluating ensemble performance at training
verbose	True	Determines if ensemble evaluation metric is printed at training

- Random Forest ensemble utilizes the bagging mode to grow successive trees independently. It uses bootstrap aggregation to construct each tree and to sample its train data using a majority vote at its prediction [103]. RF extends randomness via an extra layer that changes how it constructs trees. Each node is split using a binary-tree predictor, as RF splits its nodes and randomly selects the best predictor node from its learner subset. Its recursive structure helps it to capture interactions between various predictors. Its drawback is in their flexibility [104] with data diversity and complexity [105], as its outcome can yield lesser performance [106] for ground truth. To curb this, we adopt hyper-parameter tuning to significantly reduce model overfit, address imbalanced datasets, and enhance accuracy in its quest for ground truth [107]. Table 5 shows the RF configuration.

Table 5. Random Forest Ensemble Design Configuration

Configuration	Values	Description
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	5	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 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 the tree does not restart

3.6 Training and Validation

Our stack ensemble learns from scratch, and trees are iteratively constructed in each model for bootstrap training to yield the required enhancement using prediction probabilities from our normalized dataset. This enhances each tree's collective knowledge and helps the ensemble quickly identify intricate patterns present in each dataset since training blends normalized and original samples in its dataset, guaranteeing base-learner comprehensive learning. Hyperparameter tuning controls how a tree's complexity and weights are adjusted to gradient loss. The lower the value, the slower we travel on a downward slope to ensure how quickly a tree abandons old beliefs for new ones at training. As the tree learns – it identifies crucial, intrinsic feats. Our meta-learner yields a higher learning rate because the ensemble changes quickly as it learns newer feats. This flexibility yields ease of adaptability. The XGBoost meta-learner utilizes the regularization terms to change during learning quickly and ensures it adequately adjusts its learning to be devoid of poor generalization. Then, we carefully tuned these parameters: max_depth, n_estimator, learning_rate, and booster to ensure optimal performance [108].

Cross-validation is applied with 10 percent of the training dataset to estimate how well-learned skills by the ensemble perform on unseen data. It also evaluates the performance of the ensemble's accuracy and how well it has learned the feats of interest via resampled and balanced dataset technique. We use the stratified k-fold to rearrange the data so that each fold is a good representation of the dataset [109] and ensure our proposed stacking ensemble is devoid of overfit with improved generalization. We tested our resultant ensemble as an embedded system deployed via Flask API and Streamlit to help port the application onto various platforms as an embedded system.

4. Results and Discussion

4.1. Result Findings and Discussion

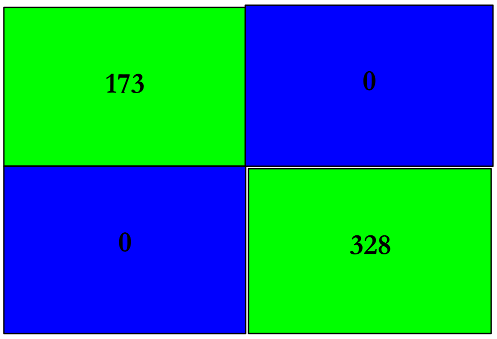
Table 6 shows the performance evaluation metrics for all base learners (Decision Tree, Random Forest, and AdaBoost) with the meta-learner (i.e., XGBoost Regressor). Note that tree-based ensemble learning aims to reduce the outcome relations conflict caused by the diversity and computational complexities of the dataset used. And in turn, ensure the ensemble is devoid of overfit considering the 3-base-learners. However, since the stacking ensemble can combine the performance of all 3-predictor classifiers – we decided to ensure simple and non-complex constructs for the tree-based base-learners used.

Table 6. Performance Evaluation of Stacked Ensemble 'Prior/After' RFE Feature Selection Mode and SMOTE-Tomek Links Data Balancing

Models	Prior RFE and SMOTE-Tomek				After RFE and SMOTE-Tomek			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
DT	0.6913	0.6846	0.6909	0.7100	0.9815	0.9805	0.9745	0.9805
AdaBoost	0.7134	0.7189	0.7208	0.7290	0.9968	0.9318	0.9848	0.9881
Random Forest	0.7323	0.7362	0.7409	0.7501	0.9981	0.9541	0.9881	0.9925
Proposed	0.8799	0.8223	0.8192	0.8874	1.0000	1.0000	0.9999	1.0000

Table 6 shows that prior to applying the RFE feature selection mode and SMOTE-Tomek links data balance approach. The base-leaners (i.e., DT, RF, and Adaboost) yield accuracy of 0.6913, 0.7134, and 0.7323, respectively - with recall of 0.6909, 0.7208 and 0.7409 respectively, And precision of 0.6846, 0.7189, and 0.7362 respectively – with F1 of 0.71, 0.7290, and 0.7501 respectively. Conversely, the meta-learner yields an Accuracy of 0.8799, recall of 0.8192, precision of 8223, and F1-score of 0.8874, respectively, which leverages the transfer learning flexibility via the stacking mode to yield improved metrics. The poor generalization before recursive feature elimination and data balancing agrees with [110]. It has also been attributed to the fact that sophisticated models also yield improved performance generalization.

Conversely, applying the RFE feature selection approach and SMOTE-Tomek data balancing positively impacted the stacking ensemble results. Adaboost and DT were observed to underperform when compared to Random Forest, which agrees with [41]. However, all 3-base leaners (i.e., DT, RF, and Adaboost) yield training accuracy of 0.9815, 0.9968, and 0.9981, respectively. With recall score(s) of 0.9745, 0.9848, and 0.9881, respectively; and precision of 0.9805, 0.9318, and 0.9541, respectively; and F1 of 0.9805, 0.9881, and 0.9925, respectively. Conversely, the meta-learner yields perfect accuracy, recall, precision, and F1 scores. Thus, our ensemble classifies hypertension data accurately as detected [104] dataset and has proven to efficiently reduce bias and variance as in the confusion matrix of Figure 5 – yielding a more stable and robust heuristic for new data and/or hidden underlying parameters within the training dataset.



173	0
0	328

Figure 5. Confusion Matrix for the Stacking Ensemble

The study supports that the recursive feature elimination approach greatly influences the selection of parameters good enough in the quest for ground truth. In turn, impacts the overall performance by identifying features of importance that influence model prediction [111]. It also enhanced efficiency in differentiating between true-positive and true-negative scores and between false-positive and false-negative scores.

4.2. Comparison

As we sought to benchmark this proposed model with existing studies – we achieved this by exploring the high performance of our stacked ensemble across the domain dataset to demonstrate its flexibility, adaptability, robustness, and prediction capability – on previous studies that utilized the same dataset. Our limited reviews found no study with the same dataset to enable such a comparison. Thus, we benchmarked the proposed ensemble against studies that explored similar (stacked learning) design constructs utilized in classification and regression tasks with domain datasets ranging from medical to IoT security, as seen in Table 8.

While some domain task datasets have proven much easier to be classified, Others have also conversely proven to be more painstaking [89]. Some domain task(s) such as medical and image records – require its 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.

Table 8. Performance Evaluation with Feature selection and Data Balancing

Method	F1	Accuracy	Precision	Recall
Ref [20]	1.0000	1.0000	0.9999	1.0000
Ref [112]	0.9981	0.9800	0.9800	0.9800
Ref [113]	0,9968	0.9318	0.9848	0.9881
Ref [114]	0.9999	0.9997	0.9991	0.9997
Our Method	1.0000	1.0000	1.0000	0.9999

5. Conclusions

Finding the balance between recall and specificity is also a crucial feat, as too much emphasis on one can ripple across the dataset – to yield a significant tradeoff for the other. In addition, accuracy can yield the idea of a model's reliability, which may also be less insightful for imbalanced datasets that sometimes render distorted perceived model performance [49]. However, in truth and practice – F1 has been utilized in assessing a heuristic's performance on criteria such as data imbalance – as it has been found to provide an altruist insight into a technique's effectiveness in classifying positive cases without the overprediction of false positives. In tree-based ensembles – bagging mode at its simplest form explores majority voting from several independent decision trees to aid its prediction. The boosting approach learns from the errors of its base learner such that each successor tree is sequentially based and/or linked to account for its predecessor's error. We argue that when making a decision, it is better to do it based on experiences from previous mistakes rather than deciding for the first time. This study proves that using a stacking ensemble with XGB as a meta-classifier (with its hyper-parameter tuning) can help perfect scores for F1, and other score criteria as required for many data mining tasks.

Author Contributions: Conceptualization: AAO, DRIMS, and VOG; Methodology: FOA, MDO, and APB; Software: CCO and AOE; Validation: AAO and EVU; Formal Analysis: VOG and TCA; Investigation: FOA, DRIMS, MDO, AAO, and APB; Data Curation: CCO, and REA; Writing—original draft preparation: DRIMS, AAO, CCO and AOE; Writing—review and editing: AEI; Visualization: AEI, VOG, CCO and APB; Supervision: MDO; Project administration: DRIMS and AAO; funding acquisition: All.

Funding: This research received no external funding.

Data Availability Statement: Data can be found at [web]: www.kaggle.com/datasets/jayaprakashpondy/blood-pressure.

Acknowledgments: We appreciate the Department of Data Science at the Dennis Osadebay University Asaba Nigeria, the Department of Informatic Engineering at Dian Nuswaranto University, Indonesia and the Department of Data Science, Federal University of Petroleum Resources Effurun Nigeria for provision of both administrative and technical support.

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

References

[1] A. S. Ali, E. H. Ali, S. W. Shneen, and L. H. Abood, "Adaptive Fuzzy Filter Technique for Mixed Noise Removing from Sonar Images Underwater," *J. Fuzzy Syst. Control*, vol. 2, no. 2, pp. 45–49, 2024, doi: 10.59247/jfsc.v2i2.176.

[2] 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.

[3] 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.

[4] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," *Mater. Today Proc.*, vol. 80, pp. 3200–3203, 2023, doi: 10.1016/j.matpr.2021.07.196.

[5] O. Emebo, B. Fori, G. Victor, and T. Zannu, "Development of Tomato Septoria Leaf Spot and Tomato Mosaic Diseases Detection Device Using Raspberry Pi and Deep Convolutional Neural Networks," *J. Phys. Conf. Ser.*, vol. 1299, no. 1, p. 012118, Aug. 2019, doi: 10.1088/1742-6596/1299/1/012118.

[6] M. M. Uba, R. Jiadong, M. N. Sohail, M. Irshad, and K. Yu, "Data mining process for predicting diabetes mellitus based model about other chronic diseases: a case study of the northwestern part of Nigeria," *Healthc. Technol. Lett.*, vol. 6, no. 4, pp. 98–102, Aug. 2019, doi: 10.1049/htl.2018.5111.

[7] S. Alobalorun Bamidele, A. Asinobi, N. Chidozie Egejuru, and P. Adebayo Idowu, "Survival Model for Diabetes Mellitus Patients' Using Support Vector Machine," *Comput. Biol. Bioinforma.*, vol. 8, no. 2, p. 52, 2020, doi: 10.11648/j.cbb.20200802.14.

[8] 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.

[9] F. O. Aghware, R. E. Yoro, P. O. Ejeh, C. C. Odiakaose, F. U. Emordi, and A. A. Ojugo, "Sentiment analysis in detecting sophistication and degradation cues in malicious web contents," *Kongzhi yu Juece/Control Decis.*, vol. 38, no. 01, p. 653, 2023.

[10] V. Geteloma, C. K. Ayo, and R. N. Goddy-Wurlu, "A Proposed Unified Digital Id Framework for Access to Electronic Government Services," *J. Phys. Conf. Ser.*, vol. 1378, no. 4, p. 042039, Dec. 2019, doi: 10.1088/1742-6596/1378/4/042039.

[11] 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.

[12] David Opeoluwa Oyewola, E. G. Dada, J. N. Ndunagu, T. Abubakar Umar, and A. S.A, "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, Nov. 2021, doi: 10.46481/jnsps.2021.173.

[13] V. O. Geteloma *et al.*, "Enhanced data augmentation for predicting consumer churn rate with monetization and retention strategies: a pilot study," *Appl. Eng. Technol.*, vol. 3, no. 1, pp. 35–51, Apr. 2024, doi: 10.31763/aet.v3i1.1408.

[14] J. K. Oladele *et al.*, "BEHeDaS: A Blockchain Electronic Health Data System for Secure Medical Records Exchange," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 231–242, Jan. 2024, doi: 10.62411/jcta.9509.

[15] C. Ma, H. Wang, and S. C. H. Hoi, "Multi-label Thoracic Disease Image Classification with Cross-Attention Networks," in *Singaporean Journal of Radiology*, vol. 21, 2019, pp. 730–738. doi: 10.1007/978-3-030-32226-7_81.

[16] J. Chung and J. Teo, "Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, pp. 1–19, Jan. 2022, doi: 10.1155/2022/9970363.

[17] M. Di Cesare, "Global trends of chronic non-communicable diseases risk factors," *Eur. J. Public Health*, vol. 29, no. Supplement_4, Nov. 2019, doi: 10.1093/eurpub/ckz185.196.

[18] J. E. Hall, J. M. do Carmo, A. A. da Silva, Z. Wang, and M. E. Hall, "Obesity, kidney dysfunction and hypertension: mechanistic links," *Nat. Rev. Nephrol.*, vol. 15, no. 6, pp. 367–385, Jun. 2019, doi: 10.1038/s41581-019-0145-4.

[19] R. Antia and M. E. Halloran, "Transition to endemicity: Understanding COVID-19," *Immunity*, vol. 54, no. 10, pp. 2172–2176, Oct. 2021, doi: 10.1016/j.immuni.2021.09.019.

[20] 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.

[21] 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.

[22] O. Jaiyeoba, E. Ogbuju, O. T. Yomi, and F. Oladipo, "Development of a Model to Classify Skin Diseases using Stacking Ensemble Machine Learning Techniques," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 22–38, May 2024, doi: 10.62411/jcta.10488.

[23] C. S. Htwe, Z. T. T. Myint, and Y. M. Thant, "IoT Security Using Machine Learning Methods with Features Correlation," *J. Comput. Theor. Appl.*, vol. 2, no. 2, pp. 151–163, Aug. 2024, doi: 10.62411/jcta.11179.

[24] I. Sahnoun and E. A. Elhadjamor, "Enhanced Freelance Matching: Integrated Data Analysis and Machine Learning Techniques," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 507–517, May 2024, doi: 10.62411/jcta.10152.

[25] A. A. Ojugo, M. I. Akazue, P. O. Ejeh, C. 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. 1, pp. 667–678, 2023.

[26] S. Basterrech and M. Wozniak, "Tracking changes using Kullback-Leibler divergence for the continual learning," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2022, pp. 3279–3285. doi: 10.1109/SMC53654.2022.9945547.

[27] 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.

[28] A. A. Ojugo and C. O. Obruche, "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.

[29] E. Ileberi, Y. Sun, and Z. Wang, "A machine learning based credit card fraud detection using the GA algorithm for feature selection," *J. Big Data*, vol. 9, no. 1, p. 24, Dec. 2022, doi: 10.1186/s40537-022-00573-8.

[30] 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.

[31] 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.

[32] A. P. Binitie and O. J. Babatunde, "Evaluating the privacy issues, potential risks, and security measures associated with using social media platforms," *Int. J. African Res. Sustain. Stud.*, vol. 3, no. 2, pp. 167–179, 2024.

[33] J. Herdiansyah, F. Ariefka, S. Putra, and D. Septiyanto, "Implementation of Zhang ' s Camera Calibration Algorithm on a Single Camera for Accurate Pose Estimation Using ArUco Markers," *J. Fuzzy Syst. Control*, vol. 2, no. 3, pp. 176–188, 2024, doi: 10.59247/jfsc.v2i3.256.

[34] 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.

[35] A. A. Ojugo and O. Nwankwo, "Tree-classification Algorithm to Ease User Detection of Predatory Hijacked Journals: Empirical Analysis of Journal Metrics Rankings," *Int. J. Eng. Manuf.*, vol. 11, no. 4, pp. 1–9, Aug. 2021, doi: 10.5815/ijem.2021.04.01.

[36] M. Ifeanyi 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.

[37] 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.

[38] 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. 1, no. 3, pp. 299–310, Feb. 2024, doi: 10.62411/jcta.10057.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] M. D. Okpor *et al.*, "Pilot Study on Enhanced Detection of Cues over Malicious Sites Using Data Balancing on the Random Forest Ensemble," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 109–123, Sep. 2024, doi: 10.62411/faith.2024-14.

[44] 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, pp. 108–115, Apr. 2024, doi: 10.11591/ijict.v13i1.pp108-115.

[45] 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.

[46] 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.

[47] 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.

[48] F. Jáñez-Martino, E. Fidalgo, S. González-Martínez, and J. Velasco-Mata, "Classification of Spam Emails through Hierarchical Clustering and Supervised Learning," *arXiv*. May 18, 2020. [Online]. Available: http://arxiv.org/abs/2005.08773

[49] C. Odiakaose *et al.*, "Hybrid Genetic Algorithm Trained Bayesian Ensemble for Short Messages Spam Detection," *Adv. Multidiscip. Sci. Res. J. Publ.*, vol. 12, no. 1, pp. 37–52, Mar. 2024, doi: 10.22624/AIMS/MATHS/V12N1P4.

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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. 1, no. 3, pp. 346–357, Feb. 2024, doi: 10.62411/jcta.9539.

[56] T. Edirisooriya and E. Jayatunga, "Comparative Study of Face Detection Methods for Robust Face Recognition Systems," in *2021 5th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI)*, Dec. 2021, no. December, pp. 1–6. doi: 10.1109/SLAAI-ICAI54477.2021.9664689.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] V. O. Geteloma *et al.*, "AQuamoAS: unmasking a wireless sensor-based ensemble for air quality monitor and alert system," *Appl. Eng. Technol.*, vol. 3, no. 2, pp. 70–85, Aug. 2024, doi: 10.31763/aet.v3i2.1409.

[62] 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.

[63] 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.

[64] S. B. N 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, Jan. 2021, doi: 10.12785/ijcds/100119.

[65] D. R. I. M. Setiadi, H. M. M. Islam, G. A. Trisnapradika, and W. Herowati, "Analyzing Preprocessing Impact on Machine Learning Classifiers for Cryotherapy and Immunotherapy Dataset," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 39–50, Jun. 2024, doi: 10.62411/faith.2024-2.

[66] M. Reza Rezvan, A. Ghanbari Sorkhi, J. Pirgazi, and M. Mehdi Pourhashem Kallehbasti, "AdvanceSplice: Integrating N-gram one-hot encoding and ensemble modeling for enhanced accuracy," *Biomed. Signal Process. Control*, vol. 92, no. August 2023, p. 106017, Jun. 2024, doi: 10.1016/j.bspc.2024.106017.

[67] 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.

[68] D. A. Oyemade and A. A. Ojugo, "A Property Oriented Pandemic Surviving Trading Model," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 5, pp. 7397–7404, Oct. 2020, doi: 10.30534/ijatcse/2020/71952020.

[69] A. Suruliandi, G. Mariammal, and S. P. Raja, "Crop prediction based on soil and environmental characteristics using feature selection techniques," *Math. Comput. Model. Dyn. Syst.*, vol. 27, no. 1, pp. 117–140, 2021, doi: 10.1080/13873954.2021.1882505.

[70] 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. 50–60, Oct. 2023, doi: 10.33633/jcta.v1i2.9259.

[71] 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, Oct. 2018, doi: 10.5815/ijitcs.2018.10.07.

[72] 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," *Computers*, vol. 13, no. 8, p. 191, Aug. 2024, doi: 10.3390/computers13080191.

[73] A. Ibor, M. Hooper, C. Maple, J. Crowcroft, and G. Epiphaniou, "Considerations for trustworthy cross-border interoperability of digital identity systems in developing countries," *AI Soc.*, no. August, Aug. 2024, doi: 10.1007/s00146-024-02008-9.

[74] 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.

[75] H. El Massari, S. Mhammedi, Z. Sabouri, and N. Gherabi, "Ontology-Based Machine Learning to Predict Diabetes Patients," in *Advances in Information, Communication and Cybersecurity*, 2022, pp. 437–445. doi: 10.1007/978-3-030-91738-8_40.

[76] N. Srividhya, K. Divya, N. Sanjana, K. Krishna Kumari, and M. Rambhupai, "Diabetes prediction using support vector machine," *EPRA Int. J. Multidiscip. Res.*, vol. 9, no. 10, pp. 421–426, 2023, doi: 10.36713/epra2013.

[77] A. A. Ojugo and A. O. Eboka, "Comparative Evaluation for High Intelligent Performance Adaptive Model for Spam Phishing Detection," vol. 3, no. 1, pp. 9–15, Nov. 2018, Accessed: Dec. 21, 2023. [Online]. Available: http://pubs.sciepub.com/dt/3/1/2/index.html

[78] 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.

[79] A. A. Ojugo, P. O.Ejeh, O. C. Christopher, 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.

[80] 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.

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

[82] S. Okperigho, B. Nwozor, and V Geteloma, "Deployment of an IoT Storage Tank Gauge and Monitor," *FUPRE J. Sci. Ind. Res.*, vol. 8, no. 1, 2024.

[83] I. Odun-Ayo, V. Geteloma, A. Falade, P. Oyom, and W. Toro-Abasi, "A Systematic Mapping Study of Utility-Driven Models and Mechanisms for Interclouds or Federations," *J. Phys. Conf. Ser.*, vol. 1378, p. 042008, Dec. 2019, doi: 10.1088/1742-6596/1378/4/042008.

[84] 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.

[85] H. Lu and C. Rakovski, "The Effect of Text Data Augmentation Methods and Strategies in Classification Tasks of Unstructured Medical Notes," *Res. Sq.*, vol. 1, no. 1, pp. 1–29, 2022.

[86] M. Bayer, M. A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter, "Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 1, pp. 135–150, 2023, doi: 10.1007/s13042-022-01553-3.

[87] 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.

[88] 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.

[89] 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.

[90] A. R. Muslikh, D. R. I. M. Setiadi, and A. A. Ojugo, "Rice Disease Recognition using Transfer Learning Xception Convolutional Neural Network," *J. Tek. Inform.*, vol. 4, no. 6, pp. 1535–1540, Dec. 2023, doi: 10.52436/1.jutif.2023.4.6.1529.

[91] 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.

[92] A. Taravat and F. Del Frate, "Weibull Multiplicative Model and Machine Learning Models for Full-Automatic Dark-Spot Detection from SAR Images," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XL-1/W3, no. September 2013, pp. 421–424, Sep. 2013, doi: 10.5194/isprsarchives-XL-1-W3-421-2013.

[93] P. M. 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.

[94] 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.

[95] 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.

[96] A. A. Ojugo and A. O. Eboka, "Empirical Evidence of Socially-Engineered Attack Menace Among Undergraduate Smartphone Users in Selected Universities in Nigeria," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 3, pp. 2103–2108, Jun. 2021, doi: 10.30534/ijatcse/2021/861032021.

[97] N. N. Wijaya, D. R. I. M. Setiadi, and A. R. Muslikh, "Music-Genre Classification using Bidirectional Long Short-Term Memory and Mel-Frequency Cepstral Coefficients," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 243–256, Jan. 2024, doi: 10.62411/jcta.9655.

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

[99] 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.

[100] 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. 201–211, Dec. 2023, doi: 10.33633/jcta.v1i2.9462.

[101] 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, Nov. 2021, doi: 10.46481/jnsps.2021.308.

[102] K. Muhamada, D. R. I. M. Setiadi, U. Sudibyo, B. Wijayanto, and A. A. Ojugo, "Exploring Machine Learning and Deep Learning Techniques for Occluded Face Recognition: A Comprehensive Survey and Comparative Analysis," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 160–173, Sep. 2024, doi: 10.62411/faith.2024-30.

[103] Y. Abakarim, M. Lahby, and A. Attioui, "An Efficient Real Time Model For Credit Card Fraud Detection Based On Deep Learning," in *Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications*, Oct. 2018, pp. 1–7. doi: 10.1145/3289402.3289530.

[104] 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.

[105] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, "Random forest for credit card fraud detection," in *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, Mar. 2018, pp. 1–6. doi: 10.1109/ICNSC.2018.8361343.

[106] 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.

[107] 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, May 2020, doi: 10.35877/454RI.asci2192.

[108] 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.

[109] J. Camargo and A. Young, "Feature Selection and Non-Linear Classifiers: Effects on Simultaneous Motion Recognition in Upper Limb," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 743–750, Apr. 2019, doi: 10.1109/TNSRE.2019.2903986.

[110] R. D. Joshi and C. K. Dhakal, "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches," *Int. J. Environ. Res. Public Health*, vol. 18, no. 14, p. 7346, Jul. 2021, doi: 10.3390/ijerph18147346.

[111] D. R. I. M. Setiadi, A. R. Muslikh, S. W. Iriananda, W. Warto, J. Gondohanindijo, and A. A. Ojugo, "Outlier Detection Using Gaussian Mixture Model Clustering to Optimize XGBoost for Credit Approval Prediction," *J. Comput. Theor. Appl.*, vol. 2, no. 2, pp. 244–255, Nov. 2024, doi: 10.62411/jcta.11638.

[112] M. A. Abbas *et al.*, "A novel meta learning based stacked approach for diagnosis of thyroid syndrome," *PLoS One*, vol. 19, no. 11, p. e0312313, Nov. 2024, doi: 10.1371/journal.pone.0312313.

[113] T. Ma, F. Wang, J. Cheng, Y. Yu, and X. Chen, "A Hybrid Spectral Clustering and Deep Neural Network Ensemble Algorithm for Intrusion Detection in Sensor Networks," *Sensors*, vol. 16, no. 10, p. 1701, Oct. 2016, doi: 10.3390/s16101701.

[114] N. Islam *et al.*, "Towards Machine Learning Based Intrusion Detection in IoT Networks," *Comput. Mater. Contin.*, vol. 69, no. 2, pp. 1801–1821, 2021, doi: 10.32604/cmc.2021.018466.