NIPES - Journal of Science and Technology Research
www.nipes.org

Effects of Data Balancing in Diabetes Mellitus Detection: A Comparative XGBoost and Random Forest Learning Approach

Fidelis Obukohwo Aghware[1], Maureen Ifeanyi Akazue[2], Margaret Dumebi Okpor[3], Bridget Ogheneovo Malasowe[4], Tabitha Chukwudi Aghaunor[5], Eferhire Valentine Ugbotu[6], Arnold Adimabua Ojugo[7], Rita Erhovwo Ako[8], Victor Ochuko Geteloma[9], Christopher Chukwufunaya Odiakaose[10], Andrew Okonji Eboka[11] and Sunny Innocent Onyemenem[12]

[1,4]Department of Computer Science, University of Delta, Agbor, Nigeria. fidelis.aghware@unidel.edu.ng, bridget.malasowe@unidel.edu.ng
[2]Department of Computer Science, Delta State University, Abraka, Nigeria. akazue@delsu.edu.ng
[2]Department of Cybersecurity, Delta State University of Science and Technology Ozoro, Nigeria. okpormd@dsust.edu.ng
[5]Department of Data Intelligence and Technology, Robert Morrison University, Pennsylvania, USA. tabitha.aghaunor@gmail.com
[6]Department of Data Science, University of Salford, United Kingdom. eferhire.ugbotu@gmail.com
[7,8,9]Department of Computer Science, Federal University of Petroleum Resources Effurun, Nigeria. ojugo.arnold@fupre.edu.ng, ako.rita@fupre.edu.ng, geteloma.victor@fupre.edu.ng
[10]Department of Computer Science, Dennis Osadebey University, Asaba, Nigeria. osegalaxy@gmail.com
[11,12]Department of Computer Science Education, Federal College of Education (Technical), Asaba, Nigeria. andrew.eboka@fcetasaba.edu.ng, innocentsunnyonyemenem@gmail.con

Article Info

***Keywords*:**
Data Balancing, Diabetes Mellitus Detection, XGBoost, Random Forest, Learning Approach

Received 11 January 2025
Revised 02 February 2025
Accepted 03 February 2025
Available online 5 March 2025

https://doi.org/10.37933/nipes/7.1.2025.1

eISSN-2682-5821, pISSN-2734-2352

Abstract

Diabetes is a prevalent chronic disorder, which has contributed to many underlying health challenges – as the World Health Organization has dubbed it the world's deadliest disease and a silent killer. As a non-communicable disease – it is difficult to diagnose at an early stage due to its types (that morph through many stages) that is broadly classified into type-I, type-II, gestational and pre-diabetes. Diabetes account for over 2-million deaths annually due to failed internal organs, high-blood pressure, etc. Thus, immediate action has become imperative for early detection and warning to (pre)carrier patients. There is also the problem inherent in real-world datasets due to imbalanced class(es) distributions rippling across poor generalization, high misclassification rates and low accuracy. Our study posits the utilization of data balancing techniques using the PIMA Indian Diabetes (PID) dataset to ascertain the impact of data balancing. We use 6-known schemes (RUS, UPS, SMOTE, ADASyn, SMOTE-Tomek and SMOTEEN) to resolve dataset imbalance in PID and evaluate how well these schemes fit with improved performance. The study explores tree-based XGBoost and Random Forest ensemble in identifying diabetes. The empirical (comparative) results from balancing approaches show that XGBoost performed best with SMOTE-Tomek; while the Random Forest model performed best with SMOTEEN.

1. Introduction

Diabetes Mellitus is a metabolic chronic disorder characterized by the presence of hyperglycemia or high blood sugar [1], which results in a body's inability to adequately break-down sugar [2] or secrete enough insulin as required to process glucose, and aid the normal functioning of the body. This, results in the critical failure of other body organs as is most prevalent in the elderly of the society. Diabetes also refer to a disease condition for which the body generates insufficient insulin or cannot properly utilize the insulin produced, and ripples across the organs, a range of underlying complications such as heart or kidney failure, risk of blindness, blood pressure, and nerve damage, etc [3]. World Health Organization (WHO) has tagged it a silent killer due to its difficulty in early detection. Even as a non-communicable disease, diabetes recorded over 7-million death in 2021 – making it the 7th-leader in cause-of-death globally [4], with

1-case every 5-seconds. Health expenditure on diabetes recorded USD966-billion dollars in 2021 with a projected rise of 316-percent over the next 15-years [5]. Diabetes is responsible for over 1.6million deaths directly [6], at 20.9 deaths per 100,000 population for which, about 47% of all deaths caused by diabetes occurred before the age of 70years [7]. While, its early detection is extremely tedious [8] – there is now the continued quest for effective schemes to accurately classify the disorder [9] and help manage teeming patients living with diabetes vis-à-vis reduce advances in carrier patients from serious cardiovascular, retinopathy, and nephropathy disorders/complications [10].

Diabetes is grouped into: (a) type-I is a chronic state in a patient where the patient's pancreas secretes little(no) insulin so that as sugar builds up in the bloodstream – it triggers life-threatening complications. With no available cure near in sight, its causes remain unknown [11], but have also been traced to life-style cum risk factors such as family genes [12], age, early ingestion of cow milk, exposure to Epstein-Barr virus [13], vitamin D deficiency, early introduction of infants to gluten diet, intake of nitrate-contaminated water [14], pregnant mothers with preeclampsia [15], infant jaundice [16]. While, it is insulin-dependent, its major symptoms are blurred vision, extreme hunger and thirst, fatigue, irritability, incessant urination, unintended weight loss, vaginal yeast infection, etc [17], and (b) type II is a chronic state that affect how a body metabolizes sugar [18]. Its slow development is triggered by a body's inability to produce enough insulin for metabolism to maintain normal glucose level, or for a body that resists its produced insulin's impact. Thus, while this stage is insulin non-dependent, it is commonly found in elderlies and obese persons; And, can be properly managed via proper eating habits, exercises, maintain a healthy weight and (extreme cases) administer insulin therapy. Its symptoms include weight loss, frequent urination, fatigue, blurred vision, acanthosis nigricans (darkened skin) [19] etc. Type-II has asymptomatic preclinical phase, which is not benign, and underscores the need for primary prevention and population screening in order to achieve early diagnosis and treatment.

Medical practitioners can manually diagnose diabetes – and previous studies have shown that patients with type II are more susceptible to a diverse range of both short/long-term consequences/complications that frequently lead to a rise in early mortality [20]. With diabetes as a cardiovascular ailment, it has become crucial and imperative to seek a more effective method to manage, diagnosis and early detect the diabetes disease – as previous studies have proven that drug-use can only be exercised as regulatory model [21], [22]. Related studies in health information management systems [23] – have proven a necessity to efficiently manage the veracity, value, and volume of healthcare patient data generated by healthcare facilities, especially with the provision of frontier healthcare to patients [24], [25]. Machine learning approaches have thus, become the much-needed and requisite tools via predictive models to glean insightful knowledge from such a volume of patient medical records [26]. Thus, the need to harness the predictive prowess of ML approaches to serve as decision support to clinicians and health experts in the early diagnosis and management of diabetes as the use of traditional approach are found to not yield cost-effective optimal solutions and alternatives [27].

1.1. Challenges with Imbalanced Dataset as Resolved with Data Balancing

ML schemes effectively improves predictions by reducing the overall variance and bias inherent in a dataset, whilst enhancing generalization [28]. Balancing allows prediction to benefits from the comprehensive knowledge of the explored model while focusing on error reduction to proffer a powerful model that exploits the learning depth of its base models. Also, performance of a scheme is often degraded by imbalanced nature of dataset [29]. By nature, datasets gathered using primary data collection techniques are unstructured in the naturalistic form, design and patterned labelling [30]. Thus, they yield unequal distribution in their major-and-minor classes, which pose challenges to the ML, if not adequately handled via data balancing [31]. With the obvious feat that many dataset(s) in their simplistic forms are rippled with imbalanced nature and schema [32]: (a) there exists the bias towards the major-class and the explored ML scheme or technique will end up ignoring the minor-class distribution as insignificant [33], (b) ML classifiers have been found and proven overtime to perform better for the major-class and worse for the minor-class, (c) the minor class if poorly classified yields high rate of misclassification even in a high-performing ML scheme [34], and (d) this imbalanced have proven to yield misleading accuracy with degraded model performance [35].

Furthermore, a critical factor that improves performance is the utilization of a rightly-formatted dataset so that the model can lean on: (a) its flexibility to appropriately encode the un(structured) dataset [36], (b) robust reuse of model for chosen or related task(s), and (c) yield cost-effective, optimal solution even with ambiguity, noise, and partial truth as contained in its (input) dataset [37]. Where a model does not learn the features of interest in an (un)structured, imbalanced dataset – this, leads to poor generalization and biased learning [38]. An imbalanced dataset overwhelms the chosen model – yielding poor generalization; while, a balanced dataset is a panacea for enhanced learning and the right recipe for improved generalization [39].

1.2. Learning Schemes: Literature Review

ML approaches have proven to be useful in the identification of intrusive anomalies. They achieve such a feat by learning the intrinsic patterns inherent in domain task (intrusion) predictor features as contained within the (un)structured dataset [40]. Identification tasks are grouped into [41]: machine learning (ML), deep learning (DL), and ensemble learning (EL). ML models as used in high-dimension tasks – are trained to identify hidden relations of interest in (un)structured dataset to support decision in the quest for truth (i.e. target class) [42]. Their robustness, reusability and flexibility lets them quickly learn such relations as changes occurs via feature engineering to eases outlier

identification in the functioning of a system [43], [44]. Thus, it determines crucial predictors selected for model construction as input; And in turn – recognize those to aggregate as output. Common traditional ML models include Random Forest [45], SVM [46], Naïve Bayes [47], etc. Conversely, DL are networks tailored [48] to capture underlying relations of interest in a dataset. Its vanishing gradient challenge impedes performance and hinders the widespread use of RNN. Its variant, the Long-Short-Term Memory (LSTM) resolves this challenge exploring input-gates that effectively manage how quickly and easily the model adapts to changes observed in the dataset [49]. A major hinderance to the LSTM is its requisition for longer training time and its inability to handle categorical large datasets [50].

To combat the challenges in both ML and DL – EL fuses both ML and DL using ML to overcome the issues in DL and vice-versa. Thus, the EL yields a single and stronger optimal fit classifier. This feat is achieved via: (a) vote, (b) bagging, (c) boost, and (d) stacked schemes [51]. In vote mode, classifier(s) are independently aggregated to yield a final output with enhanced performance. While, it does rely on their fused predictive relations – this unexplored fusion degrades performance if more diversity and outliers exists in the dataset [52]. Bagging (like vote) trains similar decision-trees with equal vote weight(s) [52]. It minimizes the variance and bias in a dataset by randomly training its tree with k-fold train-data so that model aggregates all trees predictions to yield greater accuracy with reduced errors [53]. With boost, it sequentially trains independent decision trees so that each iteration yields a learner/classifier that corrects the mistakes of its base (previous) learners in the output [54]. Thus, with each iteration – the ensemble learn and amends its predecessors incorrectly predicted data [55] to yield enhanced performance with ADAboost as an example [56]. Lastly, the stacked mode explores transfer-learning approach, which trains its (meta) classifier(s) to efficiently fuse the predictive outcome of its many base-classifiers to improve on the generalization performance of its (meta)classifier. This flexibility yields enhanced outcome with lesser convergence time and iterations [57], [58].

1.3. Challenges in Diabetes Detection

The study is motivated thus:

1. **Traditional Modes of Detection** via diabetes lipids screening and detection can be often cumbersome. Hence, the advent of wearable robotics to aid the efficient early detection and warning via sensor-based observations and alert units. Many patients do not experience its many symptoms early enough until the disease has long progressed with degenerated health conditions, and its limited awareness as a silent killer alongside the scarcity of clinical experiences to provision the necessary requisite skillset to diagnose the disease [59] compels its urgent attention. It has become imperative to deploy ML algorithms, embedded as wearable(s) sensor-based observation devices to aid early identification; And in turn, reduce its range of risks/complications; And ultimately enhance the survival rate and chances of prospective patients [60].
2. **Diabetes Detection:** The convention for identifying diabetes requires the plasma glucose criteria during a 75-g oral glucose tolerance test (OGTT) or A1C criteria as: (a) fasting plasma glucose (FPG) $\geq$ 126 mg/dL on 7.0 mmol/L (no calories intake for at least 8hrs), (b) 2-h plasma glucose (2-hPG) $\geq$ 200 mg/dL on 11.1 mmol/L during OGTT, (c) A1C $\geq$ 6.5% on a 48 mmol/mol using the NGSP certified/standardized to D method, and (d) in patients with classic symptoms of hyperglycemia crisis, a random glucose $\geq$ 200 mg/dL on 11.1 mmol/L during OGTT. These tested (at FGP, 2-hPG with 75-g OGTT, and A1C) are appropriate; And while these screening do not necessarily detect the disorder amongst same individuals, their efficacy for the primary prevention of the diabetes type-II have been demonstrated with patients who have impaired glucose tolerance with(out) elevated fasting glucose [61], [62].
3. **Limited Availability of Dataset:** The formal construction of machine learning models for their utilization in the identification and early prognosis of diseases/disorders requires access to right-quality datasets, which in turn will aid and fasten model construction as required for training and evaluation [63],
4. **Imbalanced Nature of Diabetes Dataset:** This is often experienced in many datasets for which the minor class records lag behind normal (major-class) records [64]. Previous studies posits: (a) detection models perform better if target class is in the major-class distribution, and perform worse for the minor class [65], [66], (b) complexity in the dataset with bias and variance inherent towards its major-class ensures that the models in some sense, is found to ignore the minor-class [67], (c) this yields poor performance with the minor-class poorly classified [68] – leading to high misclassification rates, and (d) such misleading results will yield sub-optimal performance [69], [70].

Thus, we seek to capture dynamic parameters to yield optimal fit solution that satisfies target class with improved generalization via proposing a variety of data balancing schemes to ascertain their impact on the chosen models. The study contributes thus: (a) demonstrates the utilization of data balancing technique on the chosen dataset to yield improved data quality and distribution for both class(es), (b) develop a variety of ML schemes to ascertain the flexibility and robustness with the chosen heuristics, and (c) evaluate the impact of the data balancing schemes on the chosen datasets.

2.0. Materials and Methods

Our proposed method as shown in Figure 1 – adopts the stacked learning mode with the following steps:

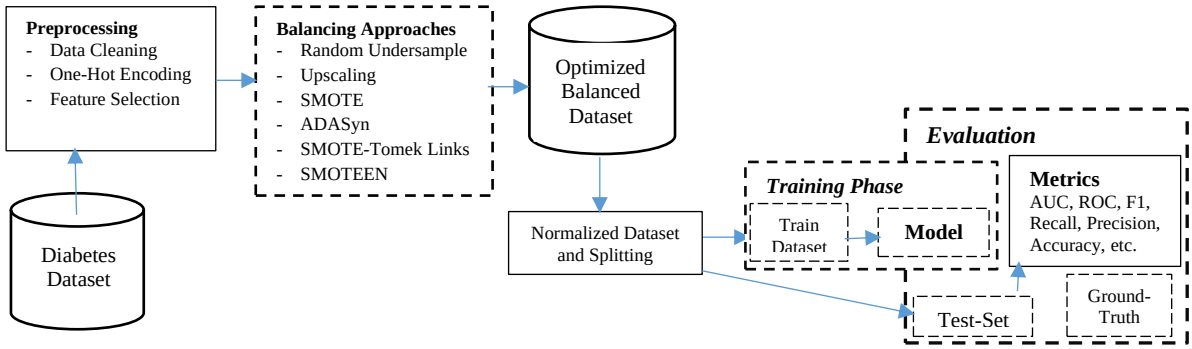

Figure 1. Proposed Methodology with Data Balancing/Resampling Approach(es)

1. **Step 1 – Dataset** was retrieved from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. The PIMA Indian Diabetes (PID) dataset consists 768-records from high prevalence of type-II diabetes with 9-attributes and all female patients of 21years and above. Dataset grouped its risk factors by association and yields a distribution of 500-records (major) no_diabetes, and 268-records (minor) diabetes class. No missing data – PID has duplicate records with plot in Figure 2a 0 read in input on to the Python DataFrame.
2. **Step 2 – Preprocessing** rids the dataset of discrepancies such as: (a) missing values, resolved by ignoring tuple or inserting missing values, (b) duplicates are resolved by removing redundant records, and (c) noisy (meaningless to interpret) are removed [71]. These, ensure data quality, improve data integrity, and yield an optimized, restructured dataset whilst retaining the labeled-classes. We use one-hot technique to encode and transform the categorical data into their binary equivalence for effective use by the model [72].
3. **Step 3 – Feature Selection** seeks to extracts and utilize for model construction – those data points that will form and be encoded as input (x), while – determining also the data-points that the model will predict as target class output (y) in its quest for ground-truth. The model evaluates how efficient the explored selection technique is – by how well and easily the model fits to the target-class. But, the PID is quite difficult to predict due to its homogeneity feature and limited predictor variables (i.e. 9-attributes). As thus, we utilize all the predictors therein for the model.
4. **Step 4 – Dataset Balance** resamples data-labels in a domain dataset to ensure an almost equal distribution of both the major-and-minor class(es) [73]. Our study seeks to evaluate the impact of a variety of balancing schemes [74] as explored in classification tasks to address the imbalanced nature of datasets as in Figure 2. These include:

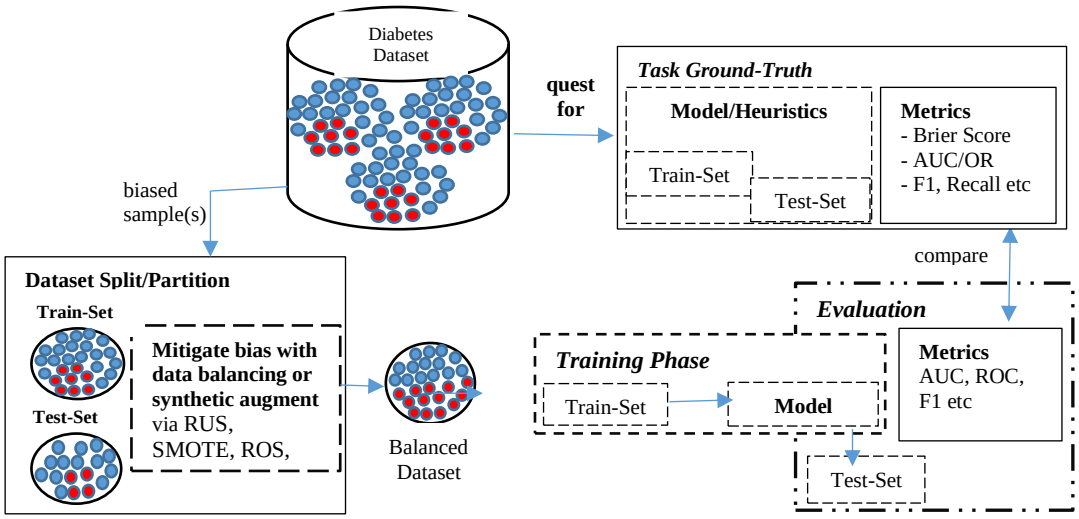

Figure 2. Data Balancing / Resampling techniques as explored in detection-based tasks

a. Random Under-sampling (RUS) randomly selects data points of interest, eliminates such instance(s) that are not closest neighbors to identified samples from the major-class [75], [76] in its bid to address inherent oversampling issue with the major-class distribution [77]. Figure 2b show its distribution plot.
b. Upscaling (UPs) approach is an over-sampling mode that randomly selects data-points of interest from the minor-class(es), interpolates to create synthetic points as closest neighbor(s) to samples from the minor-class(es), and

redistributes the synthetic point(s) onto the feature space – provisioning new, additional knowledge to the dataset to yield a more balanced distribution of the inherent classes. UPS mode yields the algorithm listing 3.1 with Figure 2c showing the distribution plot.

Algorithm Listing 3.1

Input: The original train data
1. choose an instance (xi) from the minor-class of the original dataset
2. interpolate to create a new instance at random in the minor-class
3. add newly created instance(s) into pool to yield new knowledge
4. repeat this procedure till the required threshold is attained: **end**

Output: Balanced version of the dataset

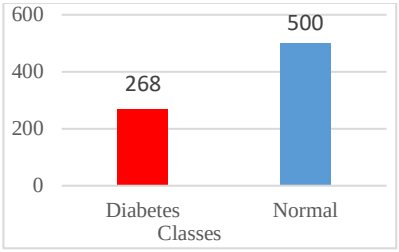

Figure 2a. Original Dataset

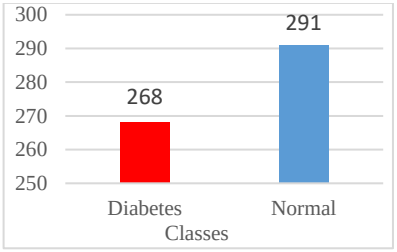

Figure 2b. RUS Balanced plot

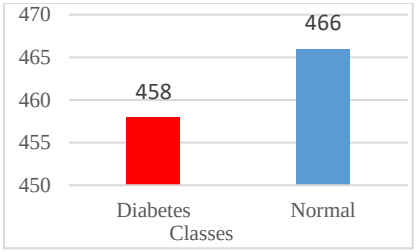

Figure 2c. UPs balanced dataset plot

c. The Synthetic Oversampling Technique (SMOTE) is an over-sampling technique, which seeks to balance the dataset class distribution via: (a) identifying minority class, (b) adjusting points to those of its closest neighbors, (c) interpolating points between the minor class instances and to its closest neighbors to create synthetic data, and (d) add the synthetic instances to original dataset to yield an oversampled, balanced dataset of both classes [78]. It yields the algorithm listing 3.2 with Figure 2d showing the distribution plot.

Algorithm Listing 3.2

Input: M(minor_class_sample); N(synthetic_sample); number_k_nearest_neighbor for ***i*** in range(N);
1. x = random_gen_sample(M) //generate random samples of the minor class
2. neighbors = k_nearest_neighbor(x)
3. y = random_gen_sample(neighbors)
4. sample = x + (y – x) * random_uniform(0,1) to create new_sample
5. T.add(sample): end

Output: Minor-class (newly created) samples added to yield balanced version of the dataset

d. Adaptive Synthetic Sampling (ADASyn) – extends the SMOTE approach using the nearest k-neighbour mode to generate more samples of the minor-class (which is often ignored and yields poor generalization); Thus, using the linear interpolation – it generates more synthetic instances between existing minor-class samples to yield that almost equals those of the majority-class. Figure 2e showing the distribution plot.

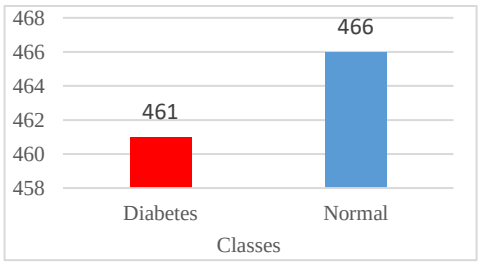

Figure 2d. SMOTE balanced distribution plot

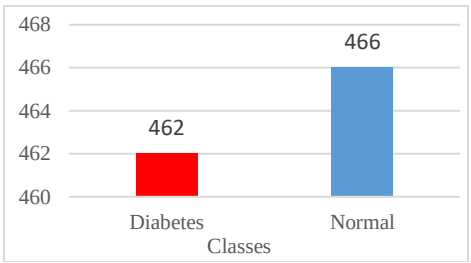

Figure 2e. ADASyn balanced distribution plot

e. SMOTE-Tomek links is a hybrid of SMOTE (over-sampler) and Tomek-links (under-sampler), which seeks to create synthetic samples for the minor-class with removal (under-sampling) of labels from major-class (closest to minor class) to yield a balanced distribution. With algorithm listing 3.3, Figure 2f showing the distribution plot.

Algorithm Listing 3.3

Input: M(minor_class_sample); N(synthetic_sample); number_k_nearest_neighbor for ***i*** in range(N);
1. from minor_class, choose random data-point //start SMOTE_mode
2. compute: relative_distance from randomly_selected_data and k_nearest_neighbor
3. choose rnd_val = random_value(0,1): rnd_val * relative_distance;
4. **if** simulated_samples = obtained **then** minor_class_new = minor_class + simulated_samples
5. repeat steps 2-to-4 until threshold_minor_class_new = reached;
6. select rnd_minor_class(data) //start Tomek_Links (under-sampler) approach

7. find k_nearest_neighbor(randomized_data)
8. **if** k_nearest_neighbor.selected = minor_class_new **then** TomekLink created
9. stop TomekLink procedure: end

Output: Balanced version of the dataset created

f. SMOTEEN fuses SMOTE (oversampler) and Edited nearest neighbour (under-sampler) by identifying and linking data points to its closest neighbor(s) to address both issues of over/under-sampling via data clean [79]. It resamples/creates synthetic instances for a minor-class, and randomly removes from a major-class to resolve the dataset imbalance via the closest neighbor approach. It generates new instances via the sampling ranges to its closest neighbor, balancing class distributions. It yields the algorithm listing 3.4 with Figure 2g showing distribution plot.

Algorithm Listing 3.4

Input: M(minor_class_sample); N(synthetic_sample); number_k_nearest_neighbor for ***i*** in range(N);

1. From minor_class, choose random data-point //start SMOTE_mode
2. compute: relative_distance from randomly_selected_data and k_nearest_neighbor
3. choose rnd_val = random_value(0,1): rnd_val * relative_distance;
4. **if** simulated_samples = obtained **then** minor_class_new = minor_class + simulated_samples
5. repeat steps 2-to-4 until threshold_minor_class_new = reached;
6. default number_clostest_neighbor = 3 //start Edited nearest neighbor (ENN) under-sampler approach
7. find k_nearest_neighbor(randomized_data)
8. **if** k_nearest_neighbor.selected = manor_class_new **then** TomekLink created
9. stop TomekLink procedure: end

Output: Balanced version of the dataset created

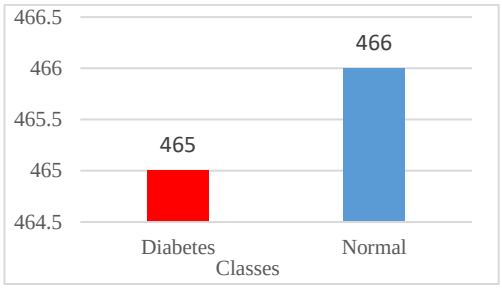

Figure 2f. SMOTE-Tomek Links balanced plot

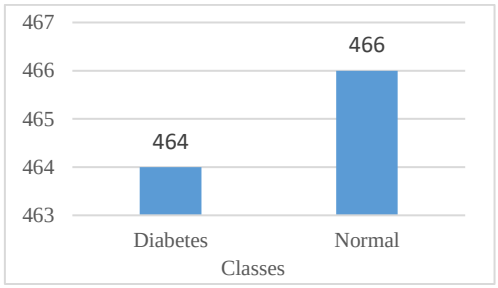

Figure 2g. SMOTEEN balanced plot

Merits of data balancing schemes includes that: (a) it prevents the skewed variance and bias inherent in real-world dataset, which often degrades performance, (b) it ensures that the model adequately learns hidden patterns to yield improved model generalization, (c) it helps the model to acknowledge the minor-class(es), which is often ignored and in turn, lessens the integrity of the entire dataset, and (d) it binds the majority-class to the impact of the minority-class, which in turn, helps the model to better understand the feature/predictor significance of each class to yield insightful evidence. Dataset is split using a 70% train-dataset and 30% test-dataset for all schemes.

5. **Step 5 – Model Construction and Train:** Quest for deploy of medical apps for early diabetes detection explores a variety of heuristics poised to improve its generalization using dataset such as PID [80], Iraqi Society Diabetes (ISD) [3], and Nigerian Diabetes Dataset [81], etc. While, the ISD is not as popular and as challenging as the PID – previous identification have results in Accuracy range between 0.69-to-0.89 [82]. While, [3] achieved a perfect score of 1.0000 – two (2) critical factor that hinder performance are: (a) imbalanced dataset that yield homogeneity complexity [83], and (b) explored must become more sensitive to identify the hidden patterns and become adaptive to capture predictor bias and variations. To study the impact of data balancing schemes – we will explore the performance of some traditional models (i.e. XGB and RF) evaluated with metrics such as Accuracy, Precision, Recall, F1, and Specificity [84]. Our traditional models are:

✓ **XGBoost** is a traditional tree-based leaner algorithm that explores the boost approach to yield a stronger learner that aggregates the output of its base (weaker) decision trees by correcting its predictor mistake to improve the learning outcome in the ensemble. It uses the majority voting over a set of iterations to yield optimal fit solution, and expands its goal function by minimizing its loss function [85] to manage its trees' complexities. It leans on the predictive power of weak learners, and each tree is trained to help it expands the objective function using regularization term $\Omega(f_t)$ and loss function $l(Y_i^t, \hat{Y}_i^t)$ to ensure improved generalization that fits the re-calibrated solution so that the data-points remain within set bounds to enable its loss function and hyper-predictors to be tuned for higher accuracy as in Table 1.

Table 1. XGBoost Parameter Design and Configuration

Features	Configuration
n_estimators	250
learning_rate	0.25
random_state	42
max_depth	Auto

- ✓ **Random Forest** is a tree-based, bagging scheme that aggregates its independent trees via the bootstrap construct to train its trees for prediction using majority vote. Its extra layer extends how random the trees are constructed with each node split via binary-tree predictor. The best split node(s) are then randomly selected using a recursive structure to help the trees capture interactions between the various predictors [86]. With increased diversity and complexity in the dataset [87] – it outputs poor performance. For this, the RF tunes its hyper-parameter to reduce overfit, address imbalanced datasets, and enhance accuracy as in Table 2, which shows its design and configuration.

Table 2. Random Forest Parameter Design and Configuration

Features	Configuration
n_estimators	150
learning_rate	0.25
random_state	42
max_depth	auto

3.0. Results and Discussion

3.1. Result Findings

Table 3 shows performance evaluation metrics for the explored XGBoost explored traditional model.

Table 3. XGBoost Performance Evaluation

Data Balancing	**F1**	**Accuracy**	**Precision**	**Recall**	**Specificity**	**Confusion Matrix**
Default	0.5602	0.5419	0.5401	0.5399	0.5325	TP (93) FN (19) FP (14) TN (104)
RUS	0.6305	0.6423	0.6320	0.6128	0.6370	TP (36) FN (17) FP (16) TN (98)
UPS	0.7025	0.7181	0.7241	0.7381	0.7026	TP (77) FN (09) FP (07) TN (184)
SMOTE	0.7881	0.7883	0.7665	0.7890	0.7790	TP (98) FN (12) FP (04) TN (164)
ADASyn	0.7875	0.7879	0.7705	0.7886	0.7784	TP (97) FN (12) FP (04) TN (165)
SMOTE-Tomek	**0.8189**	**0.8182**	**0.8028**	**0.8048**	**0.8200**	**TP (99) FN (09)** **FP (03) TN (168)**
SMOTEEN	0.8178	0.8179	0.8026	0.8049	0.8189	TP (98) FN (12) FP (04) TN (164)

Table 3 shows result of XGBoost on a variety of balancing schemes on PID dataset. Feature selection was undone due to the heterogeneity and limited predictor-set of the dataset. So, we used all predictors on the XGBoost with metric scores that show SMOTE-Tomek links data balancing approach yielded best result with harmonic mean (F1-score) of 0.8189 and an Accuracy of 0.8182 with Recall, Precision and Specificity values of 0.8028, 0.8048 and 0.8200 respectively to correctly classify 267-instances with 12-incorrectly classified instances [88]. Other balancing approaches as in Table 3 yields F1 range between 0.5602 to 0.8178, Accuracy range of 0.5419 to 0.8179, Recall range of 0.5399 to 0.8049, Precision range of 0.5401 to 0.8026, and Specificity range of 0.5325 to 0.8189 respectively, which indicates XGBoost can predict diabetes with high accuracy. With default values (i.e no data balancing) – we acknowledge that SMOTE-Tomek-based XGBoost influences ground-truth and impacted the overall performance prediction. Table 4 yields performance evaluation metrics for the Random Forest model.

Table 4. Random Forest Performance Evaluation

Data Balancing	**F1**	**Accuracy**	**Precision**	**Recall**	**Specificity**	**Confusion Matrix**
Default	0.5562	0.5413	0.5391	0.5394	0.5354	TP (91) FN (21) FP (16) TN (102)
RUS	0.6298	0.6253	0.6298	0.6282	0.6297	TP (32) FN (27) FP (14) TN (94)
UPS	0.7022	0.7180	0.7248	0.7240	0.7161	TP (72) FN (21) FP (19) TN (162)
SMOTE	0.7881	0.7883	0.7665	0.7890	0.7790	TP (98) FN (12) FP (04) TN (164)

ADASyn	0.7875	0.7879	0.7705	0.7886	0.7784	TP (97) FP (04)	FN (12) TN (165)
SMOTE-Tomek	0.8168	0.8183	0.8120	0.8146	0.8190	TP (99) FP (03)	FN (09) TN (168)
SMOTEEN	**0.8178**	**0.8199**	**0.8206**	**0.8149**	**0.8239**	**TP (99) FP (02)**	**FN (07) TN (170)**

Table 4 shows that for the RF model indicates that the SMOTEEN balancing approach yielded best result with F1-score of 0.8179 and Accuracy of 0.8199 with Precision, Recall, and Specificity values of 0.8206, 0.8149 and 0.8239 respectively to efficiently (and correctly) identify a total of 269-instances with an incorrect classification of only 9-instances. This does not suggest that Random Forest outperforms XGBoost – as we only seek to espouse the impact of explored balancing approach. Other balancing modes are as in Table 4 with F1 ranges between 0.5562 to 0.8178, Accuracy range of 0.5419 to 0.8199, Recall range of 0.5394 to 0.8149, Precision range of 0.5391 to 0.8206, and Specificity 0.5354 to 0.8239 respectively. These ranges also indicates that the Random Forest model is able to correctly predict diabetes with a high(er) accuracy if other components are to explored such as feature selection. The RF model supported by SMOTEEN proffered greater influence in the quest for ground-truth and impacted the overall performance prediction [89].

3.2. Validity Threat

All studies have expressed that there exist validity threats to a variety of degrees as thus:

a. Imbalanced Dataset: The imbalanced nature of the chosen dataset shows that only 268-instances of the recordset (i.e. 35percent) of the data records belongs to the minor (diabetes) class. This may/can be a threat against validity since the major-class instances often yields dominance effect on the model; whereas, the minor class (where not balanced) will yield poor generalization and in extreme cases, are ignored – as this will also negatively impact model performance [90]. However, our results reveals that balancing aids improved generalization as agreed.
b. Bias Introduced: Bias can also be introduced by the researcher in lieu of how the data is extracted from the original dataset. This can birth validity threat to the process therein. We can avert this by via a fully understanding of the dataset cum feature of interest. This is because, external threats to validity seek to evaluate whether or not we achieved improved generalization with the proposed model(s). This is not the case, as we sought to assess the impact of data balancing on imbalanced nature of the dataset as explored by traditional model(s) [91]. However, it is hoped that other researchers wishing to explore data-centric designs and model configuration with dataset for machine learning tailored approach can leverage and lean on the insights as provisioned therein this study.
c. **Specificity**: Some task datasets have proven easy to identify; while, others have proven more painstaking especially with medical dataset, where the chosen model/heuristics must infuse within its design [92] – the capability to measure specificity. This is because the specificity metric is strong correlates to the impact on diagnostic errors within the captured dataset as its relates to how sensitive the model is to minor shocks vis-à-vis its evaluation as directly related to the patient clinical outcomes [93].

4.Conclusion

The utilization of data balancing like SMOTE-Tomek [94] and SMOTEEN – alongside with feature selection technique, have improved performance generalization, which agrees with [95]. The use of tree-based ensemble have also shown that the heuristics can perfectly and correctly classified all test dataset with perfect accuracy. The use of feature selection technique helps the research to focus on critical feats for model construction to successfully detected spoofed sites with reduced errors that will secure user resources and provision enhanced user experience. Despite the enormous amount of data generated daily, diabetes class in the dataset – will always be found to lag behind in the quest for ground-truth. While, this study is a positive step in the right direction, we explored for our target delivery system and tested the ensemble as an embedded application program interface (API) in a standalone web app using flask API (a lightweight Python that enables us to easily integrate as embedded app for the targeted system), and Streamlit framework that provides the necessary platform to transform phishing detection ensemble into an accessible API [96]. Our Fast-API is deployed as a 3-phase system as thus: (a) the ***initialize*** function specifies communication routes required for the API, (b) its ***integrate*** function helps to connects API that allows for processing of incoming traffic requests [97], and (c) interoperability processes http requests from/to all the connected devices [98].

References

[1] S. Usman Gulumbe, S. Suleiman, S. Badamasi, A. Yusuf Tambuwal, and U. Usman, "Predicting Diabetes Mellitus Using Artificial Neural Network Through a Simulation Study," *Mach. Learn. Res.*, vol. 4, no. 2, p. 33, 2019, doi: 10.11648/j.mlr.20190402.12.

[2] A. A. Ojugo and E. O. Ekurume, "Predictive Intelligent Decision Support Model in Forecasting of the Diabetes Pandemic Using a Reinforcement Deep Learning Approach," *Int. J. Educ. Manag. Eng.*, vol. 11, no. 2, pp. 40–48, Apr. 2021, doi: 10.5815/ijeme.2021.02.05.

[3] 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.

[4] B. P. Tabaei and W. H. Herman, "A multivariate logistic regression equation to screen for diabetes: Development and validation," *Diabetes Care*, vol. 25, no. 11, pp. 1999–2003, 2002, doi: 10.2337/diacare.25.11.1999.

[5] P. Manickam *et al.*, "Artificial Intelligence (AI) and Internet of Medical Things (IoMT) Assisted Biomedical Systems for Intelligent Healthcare," *Biosensors*, vol. 12, no. 8, 2022, doi: 10.3390/bios12080562.

[6] A. A. Ojugo and O. D. Otakore, "Improved Early Detection of Gestational Diabetes via Intelligent Classification Models: A Case of the Niger Delta Region in Nigeria," *J. Comput. Sci. Appl.*, vol. 6, no. 2, pp. 82–90, 2018, doi: 10.12691/jcsa-6-2-5.

[7] P. Cihan and H. Coskun, "Performance Comparison of Machine Learning Models for Diabetes Prediction," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, IEEE, Jun. 2021, pp. 1–4. doi: 10.1109/SIU53274.2021.9477824.

[8] 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.

[9] 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.

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

[11] A. A. Ojugo, A. O. Eboka, R. E. Yoro, M. O. Yerokun, and F. N. Efozia, "Hybrid Model for Early Diabetes Diagnosis," in *2015 Second International Conference on Mathematics and Computers in Sciences and in Industry (MCSI)*, IEEE, Aug. 2015, pp. 55–65. doi: 10.1109/MCSI.2015.35.

[12] S. Habibi, M. Ahmadi, and S. Alizadeh, "Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining," *Glob. J. Health Sci.*, vol. 7, no. 5, Mar. 2015, doi: 10.5539/gjhs.v7n5p304.

[13] L. P. Joseph, E. A. Joseph, and R. Prasad, "Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture," *Comput. Biol. Med.*, vol. 151, p. 106178, Dec. 2022, doi: 10.1016/j.compbiomed.2022.106178.

[14] 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.

[15] C. P. American Diabetes Association, "Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes," *Diabetes Care*, vol. 45, no. Supplement_1, pp. S17–S38, Jan. 2022, doi: 10.2337/dc22-S002.

[16] A. Tuppad and S. Devi Patil, "An efficient classification framework for Type 2 Diabetes incorporating feature interactions," *Expert Syst. Appl.*, vol. 239, p. 122138, Apr. 2024, doi: 10.1016/j.eswa.2023.122138.

[17] A. A. Ojugo and I. P. Okobah, "Hybrid Fuzzy-Genetic Algorithm Trained Neural Network Stochastic Model for Diabetes Diagnosis and Classification," *J. Digit. Innov. Contemp Res. Sc., Eng Tech*, vol. 5, no. 4, pp. 69–90, 2017, doi: 10.22624.

[18] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *J. Diabetes Metab. Disord.*, vol. 19, no. 1, pp. 391–403, Jun. 2020, doi: 10.1007/s40200-020-00520-5.

[19] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Comput. Methods Programs Biomed.*, vol. 220, p. 106773, Jun. 2022, doi: 10.1016/j.cmpb.2022.106773.

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

[21] 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.

[22] 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.

[23] A. E. Ibor, B. E. Edim, and A. A. Ojugo, "Secure Health Information System with Blockchain Technology," *J. Niger. Soc. Phys. Sci.*, vol. 5, no. 992, p. 992, Apr. 2023, doi: 10.46481/jnsps.2023.992.

[24] A. C. Smith *et al.*, "Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19)," *J. Telemed. Telecare*, vol. 26, no. 5, pp. 309–313, Jun. 2020, doi: 10.1177/1357633X20916567.

[25] 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, 2024, doi: 10.62411/jcta.9509.

[26] 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.

[27] M. S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," *Heliyon*, vol. 10, no. 2, p. e24536, Jan. 2024, doi: 10.1016/j.heliyon.2024.e24536.

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

[29] Y. Wu *et al.*, "Large scale incremental learning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 374–382, 2019, doi: 10.1109/CVPR.2019.00046.

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

[31] A. Alqatawna, B. Abu-Salih, N. Obeid, and M. Almiani, "Incorporating Time-Series Forecasting Techniques to Predict Logistics Companies' Staffing Needs and Order Volume," *Computation*, vol. 11, no. 7, 2023, doi: 10.3390/computation11070141.

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

[33] 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.

[34] 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.

[35] R. E. Yoro, F. O. Aghware, M. . Akazue, A. E. Ibor, and A. A. Ojugo, "Evidence of personality traits on phishing attack menace among undergraduates in selected Nigerian universities," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 6, 2022.

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

[37] S. S. Olofintuyi, E. A. Olajubu, and D. Olanike, "An ensemble deep learning approach for predicting cocoa yield," *Heliyon*, vol. 9, no. 4,

p. e15245, Apr. 2023, doi: 10.1016/j.heliyon.2023.e15245.
[38] 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.
[39] 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.
[40] 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.
[41] A. P. Binitie *et al.*, "Stacked Learning Anomaly Detection Scheme with Data Augmentation for Spatiotemporal Traffic Flow," *J. Fuzzy Syst. Control*, vol. 2, no. 3, pp. 203–214, 2024, doi: 10.59247/jfsc.v2i3.267.
[42] 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)*, IEEE, Mar. 2018, pp. 1–6. doi: 10.1109/ICNSC.2018.8361343.
[43] M. Barlaud, A. Chambolle, and J.-B. Caillau, "Robust supervised classification and feature selection using a primal-dual method," Feb. 2019.
[44] D. R. I. M. Setiadi, A. R. Muslikh, S. W. Irananda, 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.
[45] 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.
[46] C. Li, N. Ding, H. Dong, and Y. Zhai, "Application of Credit Card Fraud Detection Based on CS-SVM," *Int. J. Mach. Learn. Comput.*, vol. 11, no. 1, pp. 34–39, 2021, doi: 10.18178/ijmlc.2021.11.1.1011.
[47] 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, 2024, doi: 10.31763/aet.v3i1.1408.
[48] 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.
[49] 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.
[50] 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.
[51] C. C. Odiakaose *et al.*, "Hypertension Detection via Tree-Based Stack Ensemble with SMOTE-Tomek Data Balance and XGBoost Meta-Learner," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 269–283, Dec. 2024, doi: 10.62411/faith.3048-3719-43.
[52] 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.
[53] 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.
[54] 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.
[55] 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.
[56] 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. 86–101, 2024, doi: 10.31763/aet.v3i2.1536.
[57] A. O. Eboka and A. A. Ojugo, "Mitigating technical challenges via redesigning campus network for greater efficiency, scalability and robustness: A logical view," *Int. J. Mod. Educ. Comput. Sci.*, vol. 12, no. 6, pp. 29–45, 2020, doi: 10.5815/ijmecs.2020.06.03.
[58] L. R. Zuama, D. R. I. M. Setiadi, A. Susanto, S. Santosa, and A. A. Ojugo, "High-Performance Face Spoofing Detection using Feature Fusion of FaceNet and Tuned DenseNet201," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 4, pp. 385–400, 2025, doi: 10.62411/faith.3048-3719-62.
[59] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.
[60] 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.
[61] M. A. Hambali and P. A. Agwu, "Adversarial Convolutional Neural Network for Predicting Blood Clot Ischemic Stroke," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 51–64, Jun. 2024, doi: 10.62411/jcta.10516.
[62] 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.
[63] A. A. Ojugo and D. A. Oyemade, "Boyer moore string-match framework for a hybrid short message service spam filtering technique," *IAES Int. J. Artif. Intell.*, vol. 10, no. 3, pp. 519–527, 2021, doi: 10.11591/ijai.v10.i3.pp519-527.
[64] 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.
[65] A. Panwar, V. Bhatnagar, M. Khari, A. W. Salehi, and G. Gupta, "A Blockchain Framework to Secure Personal Health Record (PHR) in IBM Cloud-Based Data Lake," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–19, Apr. 2022, doi: 10.1155/2022/3045107.
[66] 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.
[67] Y. Yang, Y. Wang, C. Wang, X. Xu, C. Liu, and X. Huang, "Identification of hub genes of Parkinson's disease through bioinformatics analysis," *Front. Neurosci.*, vol. 16, 2022, doi: 10.3389/fnins.2022.974838.
[68] N. R. Pratama, D. R. I. M. Setiadi, I. Harkespan, and A. A. Ojugo, "Feature Fusion with Albumentation for Enhancing Monkeypox Detection Using Deep Learning Models," *J. Comput. Theor. Appl.*, vol. 2, no. 3, pp. 427–440, 2025, doi: 10.62411/jcta.12255.
[69] A. Shoeibi *et al.*, "Detection of epileptic seizures on EEG signals using ANFIS classifier, autoencoders and fuzzy entropies," *Biomed. Signal Process. Control*, vol. 73, pp. 1–18, 2022, doi: 10.1016/j.bspc.2021.103417.
[70] A. A. Ojugo and R. E. Yoro, "Extending the three-tier constructivist learning model for alternative delivery: ahead the COVID-19 pandemic in Nigeria," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 21, no. 3, p. 1673, Mar. 2021, doi: 10.11591/ijeecs.v21.i3.pp1673-1682.
[71] J. Li *et al.*, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, 2017, doi: 10.1145/3136625.
[72] 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.
[73] A. O. Eboka *et al.*, "Pilot study on deploying a wireless sensor-based virtual-key access and lock system for home and industrial frontiers,"

Int. J. Informatics Commun. Technol., vol. 14, no. 1, p. 287, Apr. 2025, doi: 10.11591/ijict.v14i1.pp287-297.

[74] 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, 2024, doi: 10.62411/faith.2024-14.

[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] 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.

[77] A. R. Muslikh, D. R. I. M. Setiadi, and A. A. Ojugo, "Rice disease recognition using transfer xception convolution neural network," *J. Tek. Inform.*, vol. 4, no. 6, pp. 1541–1547, 2023, doi: 10.52436/1.jutif.2023.4.6.1529.

[78] 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.

[79] 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.

[80] 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, 2021, doi: 10.3390/ijerph18147346.

[81] T. A. Ojurongbe *et al.*, "Predicting Type 2 Diabetes Mellitus Using Patients' Clinical Symptoms, Demographic Features and Knowledge of Diabetes." pp. 0–32, Jul. 28, 2023. doi: 10.21203/rs.3.rs-3200975/v1.

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

[83] H. Qi, X. Song, S. Liu, Y. Zhang, and K. K. L. Wong, "KFPredict: An ensemble learning prediction framework for diabetes based on fusion of key features," *Comput. Methods Programs Biomed.*, vol. 231, p. 107378, Apr. 2023, doi: 10.1016/j.cmpb.2023.107378.

[84] 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.

[85] 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.

[86] 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.

[87] Y. Yu, M. Li, L. Liu, Y. Li, and J. Wang, "Clinical big data and deep learning: Applications, challenges, and future outlooks," *Big Data Min. Anal.*, vol. 2, no. 4, pp. 288–305, 2019, doi: 10.26599/BDMA.2019.9020007.

[88] 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.

[89] 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.

[90] B. O. Malasowe, F. O. Aghware, M. D. Okpor, B. E. Edim, R. E. Ako, and A. A. Ojugo, "Techniques and Best Practices for Handling Cybersecurity Risks in Educational Technology Environment (EdTech)," *J. Sci. Technol. Res.*, vol. 6, no. 2, pp. 293–311, 2024, doi: 10.5281/zenodo.12617068.

[91] 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.

[92] C. Ma, H. Wang, and S. C. H. Hoi, "Multi-label Thoracic Disease Image Classification with Cross-Attention Networks," *Singaporean J. Radiol.*, vol. 21, pp. 1–9, 2020.

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

[94] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.

[95] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3637–3647, Oct. 2018, doi: 10.1109/JIOT.2018.2816007.

[96] 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.

[97] B. O. Malasowe *et al.*, "Quest for Empirical Solution to Runoff Prediction in Nigeria via Random Forest Ensemble: Pilot Study," *Adv. Multidiscip. Sci. Res. J. Publ.*, vol. 10, no. 1, pp. 73–90, Mar. 2024, doi: 10.22624/AIMS/BHI/V10N1P8.

[98] A. Artikis *et al.*, "A Prototype for Credit Card Fraud Management," in *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems*, New York, NY, USA: ACM, Jun. 2017, pp. 249–260. doi: 10.1145/3093742.3093912.