Panamerican Mathematical Journal
ISSN: 1064-9735
Vol 34 No. 4 (2024)

BES-Optimized SMOTE Variant to Improve Dataset Scaling for Enhanced Privacy-Preserving Classification

Vijayendra S.Gaikwad[1], Dr. Kishor H. Walse[2], Dr. Mohammad Atique Mohammad Junaid[3]

[1,3]P.G. Department of Computer Science & Engineering, Sant Gadge Baba Amravati University, Maharashtra, India
vij711@gmail.com

[2]Sant Bhagwanbaba Kala Mahavidyalaya, Sindkhed Raja, Buldana, Maharashtra, India

Article History:

Received: *18-06-2024*

Revised: *04-08-2024*

Accepted: *23-08-2024*

Abstract:
The research paper undertakes a comprehensive comparative assessment of various Synthetic Minority Over-sampling Technique (SMOTE) variants in the context of medical datasets. The study is dedicated to mitigating the challenges posed by imbalanced data in medical applications. The performance of SMOTE, SMOTE-ENN, Borderline SMOTE, SMOTE-Tomek Links, Safe Level SMOTE, Density-based SMOTE, SMOTE-Rough Set Theory, and a developed BES-Optimized SMOTE are evaluated across five distinct medical datasets such as heart disease, arrhythmia, sick euthyroid, hypothyroid and breast cancer. Additionally, the K-Nearest Neighbors (KNN), Decision Forest (DF), Random forest (RF), and Support Vector Machine (SVM) algorithms are utilized to analyze the performance of SMOTE variants on each dataset, where the developed Bald Eagle search (BES) optimized SMOTE shows superior performance than other SMOTE variants, indicating its effectiveness in generating balanced scaled datasets. Subsequently, the BES-optimized SMOTE is chosen as the most effective SMOTE variant to scale and balance the respective datasets. Following a thorough examination of results, datasets that are appropriately scaled with the selected SMOTE variant are employed for ensuing research endeavors. These refined and balanced datasets will serve as the foundation for performance analysis of our future work on efficient privacy preserving KNN classification. To facilitate this, the scaled datasets in this work will be encrypted using the Paillier Encryption Scheme.

Keywords: scaled datasets, SMOTE variants, BES optimization.

1. Introduction

Electronic health records (EHRs) have made it possible for patients to receive first-class care based on data analysis. As a result, medicinal practice, which had formerly mainly depended on the expertise of doctors or other healthcare professionals, has transitioned to a more data-driven strategy that extracts value from data [1]. Data analysis is challenging because there are still a lot of issues in the clinical field. A few of the difficulties that make data analysis challenging include concerns over data privacy, such as the unauthorized release of data due to cyber threats [2], and the spread of data across various hospitals for different purposes, like the movement of patients within the facility. These problems are often faced in the analysis of clinical data, and several research projects have been dedicated to addressing each of these issues [3-6]. A noteworthy amount of the data produced and collected in medical facilities frequently includes details about patients. As a result, the process of removing personal identifiers is crucial for data utilization. However, when this anonymous data is

exchanged among various entities, there remains a potential for privacy breaches [7-8]. EHRs help healthcare professionals make clinical decisions, but they can also lead to privacy violations for patients, which problem arises when the proportion of cases suggesting one class to those showing the other class is significantly lower. This is called a class imbalance problem, which also arises from a lack of events. Class imbalanced data primarily affects the degree of susceptibility resulting from the dataset with fewer observations of the minority class, in addition to compromising machine learning (ML) outcomes [9].

Class imbalances are a common problem in clinical data analysis that hinders pattern detection. If there is variability in the distributions between the classes, the dataset is deemed Imbalanced. Unbalanced learning [10] is the learning data mining algorithm with an unbalanced dataset. The problematic scenario arises from a variety of contexts [7]. Due to a number of factors, the analysis of such imbalanced datasets produces less trustworthy results. However, these issues can be resolved by conducting appropriate exploratory data analysis (EDA), which includes feature selection, algorithmic, and data pre-processing techniques [11-12]. Four issues arise from imbalanced datasets such as dataset size, overlap, feature vector size, and bias [13]. Numerous strategies, including cost-sensitive methods, kernel-based methods, sampling techniques, and combinations of these have been proposed as solutions to the problem [7]. Different oversampling strategies have been developed to increase discrimination, reduce the danger of overfitting, and strengthen class boundaries. The SMOTE method was introduced in [14]. By interpolating current minority samples with their nearest minority neighbors, it generates synthetic minority samples. Many variations of SMOTE have been developed as a result of its success in producing synthetic samples, which include SMOTE-Tomek [15], SMOTE-ENN [16], Borderline-SMOTE [17], SVM-SMOTE, and K-Means-SMOTE [18], [19].

Conventional ML algorithms are built on the supposition of an equal number of instances for each class. As such, these algorithms have a tendency to become partial in support of the majority class, which is usually the population that is in good health. The ensuing bias has the potential to seriously impair these algorithms' forecast accuracy. Taking up the problem of imbalanced classification becomes even more crucial in the field of healthcare, where promptness and accuracy of diagnosis are vital [19]. Recent studies on ML for imbalanced data difficulties demonstrate significant advancements; Li *et al.* [20] noted that while Convolutional Neural Network (CNN) provides extremely high performances, the issue of class imbalance persists. The authors introduced a loss function that incorporates an extra class-imbalance aware regularization, increasing CNN's sensitivity to the minority class samples. Additionally, Zhang *et al*. [21] asserted that a major misclassification of deep belief networks (DBNs) is caused by the imbalanced class issue. The authors divide up the misclassification costs among classes to resolve the difficulty, and then they apply the costs to DBN in order to obtain the correct classification. While the aforementioned research has been able to contribute to the solution of the class imbalance problem, the sampling methodology was found to be a more effective way to rebuild the samples for the minority class by Han et al. [17], [22].

The research objective is to investigate the effect of SMOTE-based augmentation techniques on imbalanced medical datasets. Through a comprehensive evaluation, the research aim to quantify the

improvements in classification performance achieved by employing various SMOTE variants in conjunction with state-of-the-art ML algorithms. Additionally, a BES-optimized SMOTE is developed in this research and compared with other SMOTE variants for the evaluation of performance using KNN, DT, RF and SVM classifiers on heart disease, arrhythmia, sick-euthyroid, hypothyroid, and breast cancer datasets. By leveraging the balanced datasets generated through SMOTE-based techniques, a Paillier Homomorphic Encryption [23] is used in this research to safeguard patient privacy while maintaining the integrity of diagnostic processes.

The Bald Eagle Search (BES) optimization [24] is combined with SMOTE (BES-Optimized SMOTE) to effectively handle the imbalanced data in the datasets. The BES aids in the improvement of SMOTE by determining the best k-neighbor, which plays a key role in handling imbalance data.

The remainder of the manuscript is organized as follows, the existing researches on SMOTE for imbalance issues is reviewed in section 2. The SMOTE with its types and importance as well as the Paillier Homomorphic Encryption are detailed in Section 3. Section 4 delves into the methodology employed for dataset preparation and augmentation; and discusses the selection of ML algorithms. Additionally, section 5 presents a comprehensive evaluation of the results obtained and section 6 concludes the paper.

2. Literature Survey

Huang and Chiu [25] combined feature selection and SMOTE technique for breast cancer prediction. The outcomes of experiments conducted on two breast cancer datasets reveal that the combined function of feature selection and over-sampling is more effectual than using either method individually for datasets that are highly imbalanced. However, for datasets that are less imbalanced and have fewer features, only using over-sampling is a more suitable approach.

Abid Ishaq *et al.*, [26] applied the SMOTE to address imbalanced data in a dataset pertaining to the survival of cardiovascular patients. Their primary objective was to predict patient survival and to achieve this, they evaluated the performance of nine distinct classification models. Their experimental outcomes indicate that when SMOTE is employed, the Extra Tree Classifier (ETC) surpasses the other models, achieving an impressive accuracy value of 0.9262 in the prediction of heart patient survival, as reported in [4].

Azhar *et al*. [27] studied various SMOTE-based methods on 24 binary class imbalanced datasets for classification problems. They analyzed which SMOTE variant to use for various data complexities like small disjuncts, class overlap, and noise examples.

Hui Han *et al.*, [17], introduced two novel minority over-sampling techniques: borderline-SMOTE1 and borderline-SMOTE2. These methods specifically target the over-sampling of minority examples located in proximity to the class border. Through experimentation, they demonstrated that these approaches, when applied to the minority class, yielded superior true positive (TP) rates and F-values in comparison to both SMOTE and random over-sampling methods.

Rok Blagus and Lara Lusa [28] explored the characteristics of SMOTE from both theoretical and practical perspectives, utilizing both simulated and actual high-dimensional data. In high-dimensional data scenarios, they found that SMOTE may not effectively address class imbalance and

can even exacerbate majority class bias for most classifiers. However, when dealing with high-dimensional data, k-nearest neighbor (k-NN) classifiers utilizing Euclidean distance benefit from SMOTE, especially when combined with variable selection. Using more neighbors further enhances its effectiveness. Additionally, they found that SMOTE in high-dimensional data doesn't significantly alter class-specific means but reduces data variability and introduces sample correlation, impacting class predictions.

Waqar *et al.*, [29] introduced a cost-effective, feature engineering-free solution using ML algorithms on a UCI dataset to predict heart attacks. The approach includes SMOTE to address class imbalance. Results demonstrate that a SMOTE-based artificial neural network, when tuned properly, outperforms other models, offering a reliable method for heart attack prediction.

Ahmed Arafa *et al.*, [30] presented a reduced noise SMOTE (RN-SMOTE) to preprocess the imbalanced data. At first, the model oversampled the data using SMOTE, and the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) reduced the noise. RN-SMOTE offered improved performance on different classifiers like RF, Adaboost, SGD-LR, and SVM with critical imbalanced ratios and different dimensionalities on real datasets. However, the RN-SMOTE had a high time complexity, which affected the performance of the model.

Hatice Nizam-Ozogur and Zeynep Orman [19] GASMOTEPSO-ENN approach to address the imbalanced data issue in health datasets. The GASMOTEPSO-ENN was a combination of SMOTE, particle swarm optimization (PSO), genetic algorithm (GA), and edited nearest neighbor (ENN). Despite the fact that the GASMOTEPSO-ENN approach had managed class imbalance with remarkable outcomes and works effectively on a range of health datasets, it had several drawbacks the model was limited to binary classification and non-adaptive to classification complexities.

Geometric SMOTE for Nominal and Continuous features (G-SMOTENC) was introduced by Joao Fonseca and Fernando Bacao [31] and utilized the data selection and creation methods of G-SMOTE to generate a dataset with mixed data types. The G-SMOTENC produced data quality was evaluated across twenty datasets with varying imbalance ratios, metric/non-metric feature ratios, and class counts. The findings demonstrated that G-SMOTENC outperformed the more well-known SMOTENC, Random Oversampling, Random Undersampling, and a recently suggested oversampling method for mixed data types (SMOTE-ENC). However, determining the optimal parameters for tuning is difficult.

R. Sujithaa and B. Paramasivan [32] developed a SMOTE-SVM to solve the imbalance issues in data. k-NN was deployed to balance the samples and compute the difference between them. Grey Wolf Optimizer (GWO) was employed to attain improved performance. The developed method balanced the samples for each class and the use of GWO reduced the over-generalization issues with enhanced performance. However, the processing and execution time for the individual dataset classification was increased hindering the model's performance in balancing the data.

A. Mary Sowjanya and Owk Mrudula [13] introduced Bi-phasic SMOTE (BP-SMOTE) and Distance-based SMOTE (D-SMOTE) to reduce the class imbalance in datasets. A stacked approach was utilized based on stacked RNN and stacked CNN, which improved the accuracy and reduced the overfitting issues. However, the stacked approaches were difficult to train as it required more

computational resource.

3. Primitives

A. SMOTE

The SMOTE is a crucial method used in the field of ML to address imbalanced class distribution issues. Chawla and her team introduced this innovative approach in 2002 [14] to tackle the problems associated with imbalanced datasets. Under these circumstances, the learning process is dominated by the majority class, which produces biased models that ignore the minority class and favor the majority class. To address this, SMOTE aims to improve the model's performance by producing synthetic examples for the minority class, so balancing the class distribution.

The idea behind SMOTE is to create synthetic samples by interpolating between examples of the minority class that already exist. Choosing a random instance from the minority class and its k-neighbors is required for this. Next, along the line segments that link the selected instance to its neighbors, the algorithm generates new synthetic cases. This results in a more comprehensive representation of the minority class space, which in turn enhances the classifier's ability to make generalizations.

B. Types of SMOTE

i) SMOTE: SMOTE [14] is an oversampling method for addressing class imbalance in ML. In this way, class representation is essentially balanced while creating synthetic samples through interpolating between minority class examples that already exist.

ii) SMOTE Edited Nearest Neighbors (SMOTE-ENN): SMOTE-ENN [16] combines the SMOTE and ENN techniques. It first applies SMOTE to create synthetic samples for the minority class and then uses ENN to remove noisy or irrelevant examples, improving the quality of the oversampled dataset.

iii) Borderline SMOTE (BDS): Borderline SMOTE [17] is an extension of SMOTE that selectively generates sy nthetic samples for borderline instances, which are close to the decision boundary. This approach aims to address class imbalance while focusing on the samples that are more challenging to classify.

iv) SMOTE with Tomek Links (SMOTE-TL): SMOTE- TL combines SMOTE with Tomek Links [15], a method for identifying and removing instances that are Tomek links, or close neighbors of different classes. This combined approach aims to enhance the quality of synthetic samples created by SMOTE.

v) Safe-Level SMOTE (SL-SMOTE): SL SMOTE [33] is an advanced variant of SMOTE that generates synthetic samples by considering the "safe level" of a data point, which is based on its proximity to other minority class samples. This method aims to improve the robustness and quality of the oversampling process by focusing on the samples with the highest potential for improving model performance.

vi) Density-Based SMOTE DB-SMOTE (DB-SMOTE): This method [34] uses a way of finding groups of data points based on their density, and it's designed to create more examples from clusters

of data points that have an unusual shape. It does this by making new examples close to the center of these clusters. These new examples are packed close together near the center but spread out further away from it.

vii) SMOTE Rough Set Theory (SMOTE-RSB): This [35] approach combines the SMOTE with a Rough Set Theory-based editing technique to create new samples for preprocessing imbalanced datasets. The editing technique involves using the lower approximation of a subset to refine the generated samples.

viii) Optimized SMOTE: Optimized SMOTE is the combination of SMOTE with meta-heuristic optimization algorithms, where these algorithms improve the SMOTE's performance in balancing the data.

C. Importance of using SMOTE

i) Data Diversity Preservation: SMOTE preserves the diversity of the minority class by creating synthetic samples through interpolating between instances of the minority class that already exist. In contrast, simpler methods like random oversampling merely duplicate existing examples, potentially leading to overfitting.

ii) Effective Handling of Class Imbalance: SMOTE tackles class imbalance and lowers the possibility of model bias in favor of the majority class by creating synthetic data points. This is especially helpful in cases where the minority class is underrepresented in imbalanced datasets.

iii) Reducing Overfitting: SMOTE introduces new instances that are in the vicinity of the original minority class data points. This aids in reducing the likelihood of overfitting, as the generated samples are plausible representations of the minority class, thereby enhancing model generalization.

iv) Compatibility with Various Algorithms: SMOTE is algorithm-agnostic and can be employed in conjunction with a various ML algorithms. It is not limited to specific models, making it a versatile choice for different tasks and applications.

v) Numerous Variants: SMOTE has several variations, such as Borderline SMOTE and SMOTE-ENN, which can be tailored to suit specific dataset characteristics. This adaptability allows for fine-tuning the oversampling strategy to match the data's specific challenges.

D. Paillier Homomorphic Encryption:

Paillier Homomorphic Encryption [23] is a fundamental cryptographic method that enables computations to be carried out on encrypted data without requiring decryption. It was introduced by Pascal Paillier in 1999 [14] and is widely used in privacy-preserving applications, particularly in scenarios where sensitive information needs to be processed while preserving confidentiality.

At its core, Paillier encryption relies on the mathematical properties of large prime numbers and their modular operations. The scheme is based on the computational difficulty of factoring the product of two large prime numbers, which forms the foundation of its security. Paillier Encryption scheme [23] possesses two primary homomorphic properties:

i) **Additive Homomorphism:** The scheme allows for the addition of encrypted values, yielding

an encryption of the sum of the corresponding plaintexts. Mathematically, if $E(m_1)$ and $E(m_2)$ are the encryptions of plain texts m_1 and m_2 then $E(m_1) * E(m_2) \bmod n^2$ results in an encryption of $m_1 + m_2$.

ii) **Scalar Multiplication Homomorphism:** Paillier encryption also supports scalar multiplication, enabling the encryption of a plain text raised to a constant exponent. An encryption $E(m)$, raising it to a power k results in an encryption of $m * k$. Mathematically, $E(m)^k \bmod n^2$ is an encryption of $m * k$.

4. Methodology

The Methodology section is where the step-by-step processes that are followed to investigate the research question is detailed to obtain reliable results.

A. ***Problem Statement***

The initial step involved preprocessing the data, followed by the application of various SMOTE variants to address the inherent data imbalance issue. Subsequently, the preprocessed datasets underwent evaluation using two ML algorithms, namely KNN and SVM. The resulting oversampled datasets were encrypted using Paillier encryption scheme [23] and archived for future utilization.

The primary objective is to ascertain the most efficient variant for mitigating class imbalance within the dataset. The subsequent section provides an extensive exposition of the proposed framework and its constituent elements.

B. ***Dataset Description***

The datasets that have been used are taken from the UCI ML Repository. Missing data was imputed by appropriate imputation techniques, categorical variables were handled by label encoding and one-hot encoding in appropriate cases. The data was standardized. Below is the description of each dataset:

TABLE 1. DATASET DESCRIPTION

Name	Samples	Features	Ratio
heart[36]	1025	14	1.05
arrhythmia[37]	452	177	1.18
Sick euthyroid[38]	2700	14	10.20
hypothyroid[38]	2000	14	15.39
Breast cancer[39]	569	33	1.68

C. ***SMOTE Variants***

Various SMOTE variants were implemented using the imbalanced-learn (imblearn) library. The specific imports for each variant are SMOTE, Borderline-SMOTE, SVM-SMOTE, SMOTEN, SMOTEENN, SMOTE-Tomek, and SVM-SMOTE. These libraries and variants collectively allowed us to comprehensively address class imbalance in our medical datasets and conduct a comparative

analysis of their effectiveness.

i) Reasons for Using Various SMOTE Variants:

The choice to employ multiple SMOTE variants in our study was motivated by the need to comprehensively evaluate and compare their performance in addressing class imbalance in medical datasets. Each SMOTE variant has certain characteristics that make it appropriate for specific scenarios. Our reasons for using these variants are as follows,

a) ***SMOTE:*** The standard SMOTE is utilized as a baseline to compare the performance of other variants. It provides a fundamental understanding of oversampling techniques.

b) ***SMOTE-ENN:*** SMOTE-ENN mergeS SMOTE with ENN, offering a way to simultaneously oversample and reduce noisy samples. This is beneficial when dealing with datasets contaminated with noise.

c) ***Borderline SMOTE:*** Borderline SMOTE focuses on producing synthetic samples close to decision boundary, which can be particularly effective in situations where the minority class is clustered around the majority class.

d) ***SMOTE TL:*** SMOTE TL integrates Tomek links to remove problematic samples before oversampling. This can help mitigate the impact of noisy data and overlapping classes.

e) ***SL SMOTE:*** SL SMOTE, which uses safe levels, provides a more structured and controlled approach to oversampling, preserving the dataset's underlying distribution.

f) ***DBSMOTE:*** DBSMOTE incorporates local data density, making it effective for handling regions of varying data concentration in imbalanced datasets.

g) ***SMOTE-RSB:*** SMOTE-RSB combines SMOTE with Rough Set Theory to refine synthetic samples, enhancing their relevance and effectiveness in preprocessing imbalanced data.

h) ***Optimized SMOTE:*** BES optimized-SMOTE is employed in this research as one of the SMOTE variants. Here the SMOTE is optimized using Bald eagle search (BES) optimization [24]. The BES algorithm exploits the hunting behavior of bald eagles to find the best value of k-neighbors in SMOTE that handles the class imbalance issues effectively.

ii) Bald Eagle Search Optimization:

The BES optimization [24] is motivated by the characteristics of the Bald eagle such as searching, swooping, and hunting the prey. Enabling certain characteristics into the model helps to find the best solution without falling into local optima and improves the SMOTE performance in balancing the data. The bald eagle's hunting approach serves as the basis for the BES optimization algorithm, a meta-heuristic optimization method. Large, open spaces with plenty of prey and mature trees for nesting are part of the bald eagle's habitat. Thanks to their exceptional vision and dual-vision abilities, they can identify their prey at a considerable distance. They split their hunting activity into three stages as selecting, searching, and swooping. The predator first selects the search location, where there are lots of prey. In the second stage, it searches the assigned area to find possible prey. Finally, it takes the information from the second step to determine the best place to attack and concentrates on assaulting from that position. The BES approach uses Bald's dynamic features to

address the convergence issue, which increases the algorithm's ability to reach the optimal solution faster than with conventional optimization algorithms, which usually produce the local best solution. The phases involved in the BES optimization are detailed as follows,

a) Initialization: Initially, the solution is initialized as Q which are the K-neighbors of SMOTE.

***b) Fitness evaluation*:** The fitness determines which solution is the best. In this case, the best solution is determined by selecting the solution with the highest accuracy, which is taken to be the best fitness and is represented as,

$$F_t(Q) = accuracy(Q) \qquad (1)$$

c) Solution Update: After determining the fitness function, the solutions are updated based on various stages, which are described below,

(i) Selection phase: In this phase, the solution selects the best area in the search space to hunt prey and explores the best position. This selection of space is done based on information from previous iterations and this area is located near the previous area. The selecting behavior is represented as,

$$Q_k^{t+1} = Q_{best} + \beta * r\left(Q_{mean} - Q_k^t\right) \qquad (2)$$

where, Q_k^{t+1} is the position of the k^{th} solution at $(t+1)^{th}$ iteration, and Q_k^t is the k^{th} solution's position at t^{th} iteration. Q_{best} is the best position of the solution in the search space, Q_{mean} indicates that the solutions have used all of the knowledge from the earlier points, β is the parameter to control the position change in range $[1.5, 2]$, and r is the random number in range $[0,1]$.

(ii) Searching phase: In order to expedite the search and discover a better dive site, the solutions in this stage look for prey and travel in different directions within the spiral space. The updated position based on this phase is expressed as,

$$Q_k^{t+1} = Q_k^t + x(k) * \left(Q_k^t - Q_{mean}\right) + y(k) * \left(Q_k^t - Q_k^{t-1}\right) \qquad (3)$$

where, $x(k)$ and $y(k)$ denotes the polar coordination of the solutions, which is estimated as,

$$x(k) = \frac{xr(k)}{\max(|xr|)} \ and \ y(k) = \frac{yr(k)}{\max(|yr|)} \qquad (4)$$

$$xr(k) = r(k) * \sin(\theta(k)) \ and \ yr(k) = r(k) * \cos(\theta(k)) \qquad (5)$$

where, $\theta(k)$ and $r(k)$ denotes the polar angle and polar diameter of the spiral equation respectively, and are expressed as,

$$\theta(k) = g * \pi * rand \qquad (6)$$

$$r(k) = \theta(k) + V * rand \qquad (7)$$

where, *rand* is the random number between $[0,1]$. V lies in the range of $[0.5,2]$ that finalizes the search cycles.

(iii) Swooping phase: The solution dives towards the target prey from the best location and the other solutions follow the best location and attack the prey. The updated position based on this phase is expressed as,

$$Q_k^{t+1} = rand * Q_{best} + x_1(k) * \left(Q_k^t - h_1 * Q_{mean}\right) + y_1(k) * \left(Q_k^t - h_2 * Q_{best}\right) \quad (8)$$

where, h_1 *and* h_2 *are in the range* $[1,2]$ and are the intensities of solutions towards best and center points.

$$x_1(k) = \frac{xr(k)}{\max\left(|xr|\right)} \text{ and } y_1(k) = \frac{yr(k)}{\max\left(|yr|\right)} \quad (9)$$

$$xr(k) = r(k) * \sin\left(\theta(k)\right) \text{ and } yr(k) = r(k) * \cos\left(\theta(k)\right) \quad (10)$$

$$\theta(k) = g * \pi * rand \quad (11)$$

$$r(k) = \theta(k) \quad (12)$$

d) Declaration of best solution: After the solutions are updated, their fitness value needs to be re-evaluated to declare the best solution.

e) Termination: The optimization continues to iterate by reevaluating fitness and updating solutions to attain the best solution by checking $t < t_{\max}$. If the condition is met, the process ends or else it continues to reevaluate the fitness.

The best solution obtained from the BES algorithm is determined as the best *k*-neighbors in SMOTE, which generates the new synthetic sample using interpolation between the best K neighbor and minority class sample. This effectively balances the datasets. The flowchart of the BES algorithm is depicted in Figure 1 and algorithm 1 shows the pseudo code of BES algorithm.

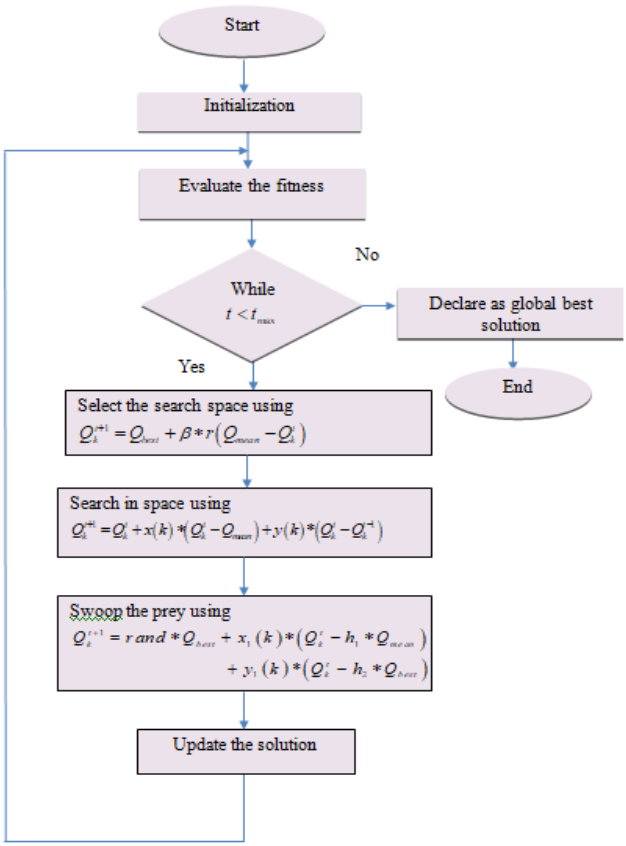

Figure 1. Flowchart of BES optimization algorithm

Algorithm 1: Pseudo code for BES optimization algorithm

Pseudo code for BES optimization
Initialization Q
Evaluate the fitness
While $\left(t < t_{\max}\right)$
For each population
Selection phase
Update the solution using $Q_k^{t+1} = Q_{best} + \beta * r\left(Q_{mean} - Q_k^t\right)$
Searching phase
Update the solution using $Q_k^{t+1} = Q_k^t + x(k) * \left(Q_k^t - Q_{mean}\right) + y(k) * \left(Q_k^t - Q_k^{t-1}\right)$
Swooping Phase
Update the solution using $Q_k^{t+1} = rand * Q_{best} + x_1(k) * \left(Q_k^t - h_1 * Q_{mean}\right) + y_1(k) * \left(Q_k^t - h_2 * Q_{best}\right)$
Return to fitness evaluation
Declare the best solution
End while
Terminate the process

By applying a range of SMOTE variants, the research aimed to assess their individual strengths and weaknesses in handling class imbalance within medical datasets. This comprehensive analysis allows

us to make informed recommendations and draw meaningful comparisons based on the unique characteristics of each variant.

D. Machine Learning Models

i) K-Nearest Neighbour (KNN):

K-nearest neighbors [40], a supervised ML algorithm, is employed for labeling datasets. This method involves finding neighboring data points for a given data point and then utilizing their predictions to make predictions for an unknown data point's label. In this research project, the sci-kit-learn library is utilized to implement KNN.

ii) Decision Tree (DT):

A non-parametric supervised learning algorithm [41], DT is employed for regression and classification tasks. Each tree is made up of nodes and branches. Each node indicates the features that are need to classified in the category and the subset denotes the value that taken by the node. By learning the simple decision rules, DT predicts the target variable's value. In this research project, the sci-kit-learn library is utilized to implement DT.

iii) Random Forest (RF):

Random forests [42], also known as random decision forests, are ensemble learning techniques for classification, regression, and other problems. They work by building a large number of decision trees during the training phase. High dimensionality, huge datasets can be handled with RF. RF keeps the overfitting problem at bay and improves the model's accuracy. In this research project, the sci-kit-learn library is utilized to implement RF.

iv) Support Vector machine (SVM):

The SVM [43] is a supervised ML technique, similar to KNN, but it outperforms KNN in terms of both cost and accuracy. Unlike KNN, SVM calculates a support vector along the decision boundary instead of measuring the distance to every data point. This support-vector is employed to classify the provided dataset. In this research project, the sci-kit-learn library is utilized to implement SVM. The workflow for the experiments is depicted in Figure 2.

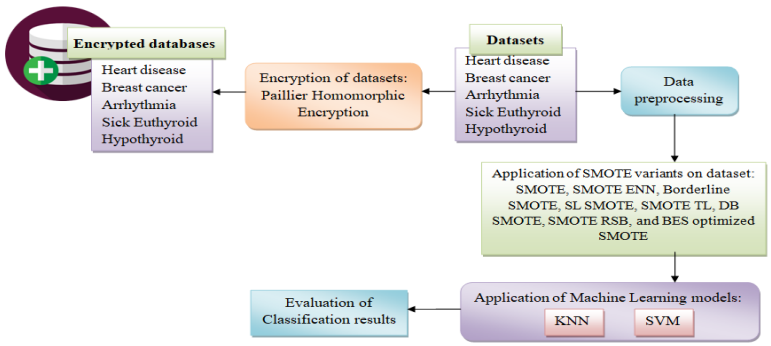

Figure 2. Workflow for experiments

*E. **Encryption Workflow***

Since further research involves working in the privacy preserving environment, datasets are encrypted and stored in a database. The encryption process involves using the Pallier homomorphic encryption technique. The Python- pallier library [44] is used for implementing Pallier encryption. Paillier scheme was selected because of its homomorphic nature. Pallier encryption is considered secure against known attacks, and its security relies on the hardness of the decisional composite residuosity assumption.

*F. **Experimental Setup***

The experiments were performed on a Google Colab environment with the following specifications: Processor: Intel(R) Xeon(R) CPU @ 2.20GHz (Dual-Core), 56320 KB Cache Size.

5. EMPIRICAL RESULTS

The tables present the performance metrics of the KNN, DT, RF and SVM applied to the dataset with SMOTE, SMOTE-ENN, Borderline SMOTE (BDS), SMOTE-Tomek Links (SMOTE-TL), Safe Level SMOTE (SL-SMOTE), Density-based SMOTE (DB-SMOTE), SMOTE-Rough Set Theory (SMOTE-RSB), and the developed BES optimized SMOTE. It includes accuracy, recall, F1 score, and precision, all of which indicate the algorithm's effectiveness in classifying the data.

Below are the performance metrics with the most prominent results among all variants.

A. Results with Heart Dataset

The accuracy scores for K-Nearest Neighbors (KNN) applied to the heart, scaled using various SMOTE variants are as follows: Without SMOTE: 0.9073, SMOTE: 0.9219, SMOTE-ENN: 0.8146, BDS: 0.9156, SMOTE-TL: 0.9245, SL-SMOTE:0.9256, DB-SMOTE:0.8293, SMOTE-RSB:0.9073, and BES optimized SMOTE: 0.93451. Additionally, the DT, RF, and SVM also offers efficient performance in the heart dataset on various SMOTE variants in which the BES-optimized SMOTE attains accuracies of 0.9652, 0.9586, and 0.9943 respectively, which show improvements of 9.38%, 10.2%, 1.18% over SL-SMOTE. These results indicate that the developed BES-optimized SMOTE attains more efficient performance than other existing SMOTE variants.

Tables 2,3, 4, and 5 are performance metrics for the Heart Dataset without SMOTE, with SMOTE, with SL-SMOTE, and with BES-optimized SMOTE respectively.

TABLE 2. HEART: WITHOUT SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.907317	0.912621	0.908213	0.903846
DT	0.835941	0.814769	0.324715	0.825372
RF	0.854167	0.823641	0.835419	0.846572
SVM	0.979321	0.963417	0.942389	0.954387

TABLE 3 HEART: SMOTE

Model	Accuracy	Recall	F1Score	F1Score
KNN	0.921951	0.912621	0.921569	0.921569
DT	0.862471	0.830047	0.839475	0.839475
RF	0.896477	0.854149	0.854789	0.854789
SVM	0.978695	0.956841	0.963524	0.963524

TABLE 4. HEART: SL-SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.925674	0.914576	0.927578	0.931435
DT	0.874601	0.843651	0.854237	0.861457
RF	0.860021	0.844721	0.846662	0.853021
SVM	0.982547	0.970874	0.985222	0.975861

TABLE 5. HEART: BES-OPTIMIZED SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.934514	0.927821	0.932569	0.941489
DT	0.965233	0.951463	0.945471	0.954326
RF	0.958635	.923611	0.943842	0.946857
SVM	0.994321	0.982054	0.988231	1.000000

B. Results with Arrhythmia Dataset

The accuracy scores for K-Nearest Neighbors (KNN) applied to the arrhythmia, scaled using various SMOTE variants are as follows: Without SMOTE: 0.6154, SMOTE: 0.6044, SMOTE-ENN:0.5934, BDS:0.5562, SMOTE-TL:0.5824, SL-SMOTE:0.6264, DB-SMOTE:0.6374, SMOTE-RSB:0.5521, and BES optimized SMOTE: 0.8017. Furthermore, the DT, RF, and SVM also provides effective performance in the arrhythmia dataset on various SMOTE variants in which the BES-optimized SMOTE attains accuracies of 0.8012, 0.7823, and 0.9118 respectively, which indicate improvements of 26.31%, 31.41%, and 18.04% over DB-SMOTE. These findings show that compared to previous SMOTE variants, the developed BES-optimized SMOTE achieves efficient performance.

Tables 6, 7, 8, and 9 are performance metrics for the Arrhythmia Dataset without SMOTE, with SMOTE, with DB- SMOTE, and with BES-optimized SMOTE respectively.

TABLE 6. ARRHYTHMIA: WITHOUT SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.615385	0.795455	0.666667	0.573770
DT	0.498525	0.465712	0.457471	0.486521
RF	0.594985	0.554863	0.547865	0.574611
SVM	0.747253	0.931818	0.780952	0.672131

TABLE 7. ARRHYTHMIA: SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.604396	0.659091	0.617021	0.585790
DT	0.532141	0.514789	0.547321	0.526894
RF	0.445387	0.425766	0.426213	0.432574
SVM	0.758242	0.818182	0.765957	0.723478

TABLE 8. ARRHYTHMIA: DB-SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.637363	0.704545	0.652632	0.607843
DT	0.585744	0.567576	0.565629	0.571228
RF	0.536582	0.521059	0.516103	0.527889
SVM	0.747253	0.840909	0.762887	0.698113

TABLE 9. ARRHYTHMIA: BES-OPTIMIZED SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.801718	0.629059	0.683127	0.750000
DT	0.801285	0.782546	0.797491	0.793468
RF	0.782374	0.761369	0.781347	0.771549
SVM	0.911828	0.735442	0.800774	0.881579

C. Results with Sick_euthyroid Dataset

The accuracy scores for K-Nearest Neighbors (KNN) applied to the sick_euthyroid, scaled using various SMOTE variants are as follows: Without SMOTE: 0.9277, SMOTE: 0.82407, SMOTE_ ENN:0.7463, BDS: 0.8344, SMOTE_TL:0.7926, SL_SMOTE:0.8148, DB-SMOTE:0.7093, SMOTE _RSB: 0.8444, and BES optimized SMOTE: 0.98614. Furthermore, in the hypothyroid dataset, the DT, RF, and SVM also provides effective performance on various SMOTE variants in which the BES-optimized SMOTE attains accuracies of 0.9876, 0.9861, and 0.9724, respectively, which indicate improvements of 17.08%, 21.03%, and 9.69% over RSB-SMOTE. Compared to other SMOTE variants, these results show that the developed BES-optimized SMOTE achieves efficient performance.

Tables 10, 11, 12, and 13 are performance metrics for the Sick_euthyroid Dataset without SMOTE, with SMOTE, with SMOTE RSB, and with BES-optimized SMOTE respectively.

TABLE 10. SICK-EUTHYROID: WITHOUT SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.927778	0.610638	0.740426	0.755319
DT	0.931129	0.921047	0.904809	0.901344
RF	0.929247	0.917592	0.902025	0.912239
SVM	0.912457	0.612766	0.732558	0.756414

TABLE 11. SICK-EUTHYROID: SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.824074	0.670213	0.780702	0.845681
DT	0.810871	0.7919857	0.802518	0.807602
RF	0.764650	0.7518502	0.741580	0.743658
SVM	0.876667	0.742553	0.681633	0.756124

TABLE 12. SICK-EUTHYROID: SMOTE-RSB

Model	Accuracy	Recall	F1Score	Precision
KNN	0.844444	0.762490	0.812754	0.830922
DT	0.818961	0.806803	0.795879	0.796924
RF	0.778702	0.742468	0.749723	0.751356
SVM	0.878148	0.728448	0.868925	0.823280

TABLE 13. SICK-EUTHYROID: BES-OPTIMIZED SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.986140	0.985915	0.965517	0.945946
DT	0.987691	0.949776	0.958724	0.965760
RF	0.986147	0.957069	0.948414	0.950642
SVM	0.972456	1.000000	0.986111	0.972603

D. Results with Hypothyroid Dataset

The accuracy scores for K-Nearest Neighbors (KNN) applied to the hypothyroid, scaled using various SMOTE variants are as follows: Without SMOTE: 0.9875, SMOTE: 0.9730, SMOTE_ ENN:0.9779, BDS:0.9625, SMOTE-TL:0.9775, SL-SMOTE: 0.9675, DB SMOTE: 0.9750, SMOTE-RSB: 0.97794 and BES optimized SMOTE: 0.993710. Furthermore, in the hypothyroid dataset, the DT, RF, and SVM also provides effective performance on various SMOTE variants in which the BES-optimized SMOTE attains accuracies of 0.996, 0.993, and 0.9927 respectively, which indicate improvements of 2.99%, 3.24%, and 2.28% over SMOTE-ENN. Compared to other SMOTE variants, these results show that the developed BES-optimized SMOTE achieves efficient performance.

Tables 14, 15, 16, and 17 are performance metrics for the Hypothyroid Dataset without SMOTE, with SMOTE, with SMOTE ENN and with BES optimized SMOTE respectively.

TABLE 14. HYPOTHYROID: WITHOUT SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.987545	0.997347	0.973395	0.989474
DT	0.988769	0.975065	0.965695	0.974359
RF	0.991921	0.964765	0.946062	0.988707
SVM	0.982543	0.975832	0.970802	0.981771

TABLE 15. HYPOTHYROID: SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.973492	0.981432	0.986667	0.991957
DT	0.975235	0.946978	0.951536	0.968076
RF	0.979528	0.966054	0.947845	0.958769
SVM	0.970381	0.976127	0.983957	0.991914

TABLE 16. HYPOTHYROID: SMOTE-ENN

Model	Accuracy	Recall	F1Score	Precision
KNN	0.977915	0.973475	0.982597	0.991892
DT	0.966753	0.950871	0.942234	0.963425
RF	0.961248	0.943569	0.943549	0.952465
SVM	0.970047	0.976127	0.983957	0.991914

TABLE 17. HYPOTHYROID: BES-OPTIMIZED SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.993710	0.984142	0.989377	0.994667
DT	0.996567	0.994129	0.993799	0.995674
RF	0.993457	0.991573	0.992235	0.992446
SVM	0.992710	0.976185	0.986624	0.997290

E. Results with Breast Cancer Dataset

The accuracy scores for K-Nearest Neighbors (KNN) applied to the breast_cancer, scaled using various SMOTE variants are as follows: Without SMOTE: 0.9540, SMOTE: 0.9261, SMOTE_ ENN:0.9698, BDS:0.9474, SMOTE_TL: 0.9146, SL_SMOTE:0.9298, DBSMOTE:0.9474, SMOTE_ RSB: 0.9193, and BES optimized SMOTE: 0.9846. Furthermore, the DT, RF, and SVM also provides effective performance in the breast_cancer dataset on various SMOTE variants in which the BES-optimized SMOTE attains accuracies of 0.975, 0.976, and 0.9894 respectively, which indicate improvements of 0.88%, 1.27%, and 1.78% over SMOTE-ENN. These findings show that compared to previous SMOTE variants, the developed BES-optimized SMOTE achieves efficient performance.

Tables 18, 19, 20, and 21 are performance metrics for the breast Cancer Dataset without SMOTE, with SMOTE, with SMOTE ENN and with BES optimized SMOTE respectively.

TABLE 18. BREAST CANCER: WITHOUT SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.954053	0.901408	0.934307	0.969697
DT	0.935657	0.904583	0.917861	0.924663
RF	0.929015	0.913560	0.900134	0.917463
SVM	0.981912	0.943662	0.971014	0.945861

TABLE 19. BREAST CANCER: SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.926140	0.945915	0.965517	0.945946
DT	0.881751	0.849631	0.86830	0.865891
RF	0.879563	0.860351	0.844603	0.8536213
SVM	0.981912	0.956842	0.978214	0.958746

TABLE 20. BREAST CANCER: SMOTE-ENN

Model	Accuracy	Recall	F1Score	Precision
KNN	0.969825	0.957746	0.944444	0.931507
DT	0.967065	0.941365	0.935160	0.950324
RF	0.9640486	0.942562	0.933035	0.954045
SVM	0.971825	0.964217	0.975656	0.959624

TABLE 21. BREAST CANCER: BES-OPTIMIZED SMOTE

Model	Accuracy	Recall	F1Score	Precision
KNN	0.984684	0.985915	0.979021	0.972222
DT	0.975670	0.9505683	0.934853	0.964893
RF	0.976460	0.946234	0.951665	0.960532
SVM	0.989456	1.000000	0.986111	0.972603

Figures 3,4, 5, 6, and 7 depict the accuracy scores applied on considered five datasets with SMOTE, SMOTE-ENN, BDS, SMOTE-TL, SL-SMOTE, DB-SMOTE, SMOTE-RSB, and BES-optimized SMOTE for KNN, DT, RF and SVM classification algorithms. The accuracy scores at the Y axis are logarithmically scaled.

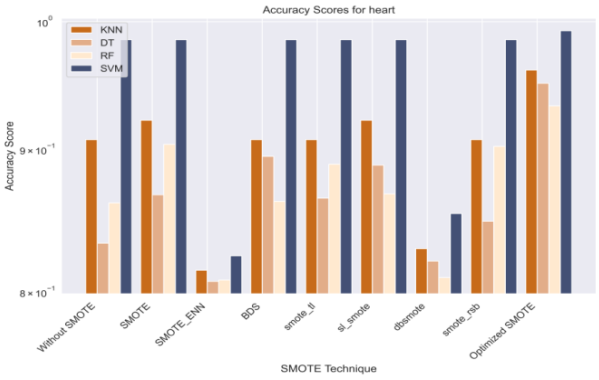

Figure 3: Accuracy Scores for Heart Disease

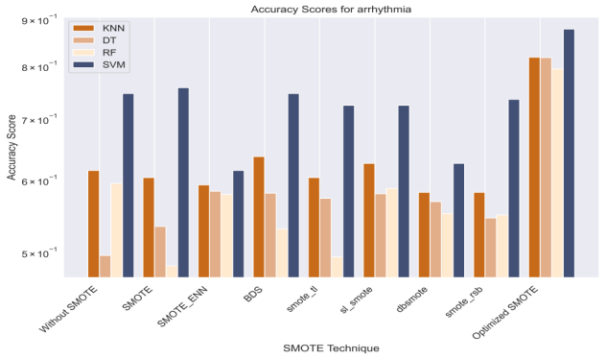

Figure 4: Accuracy Scores for Arrhythmia

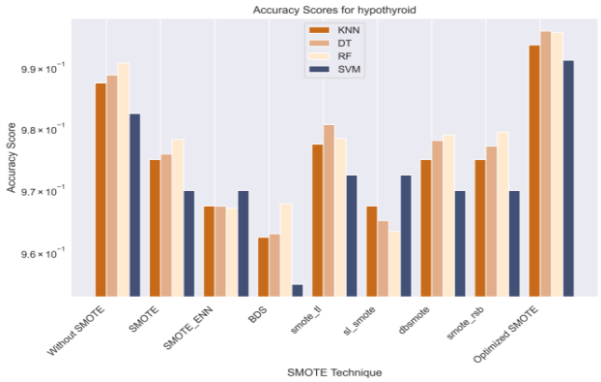

Figure 5: Accuracy Scores for hypothyroid

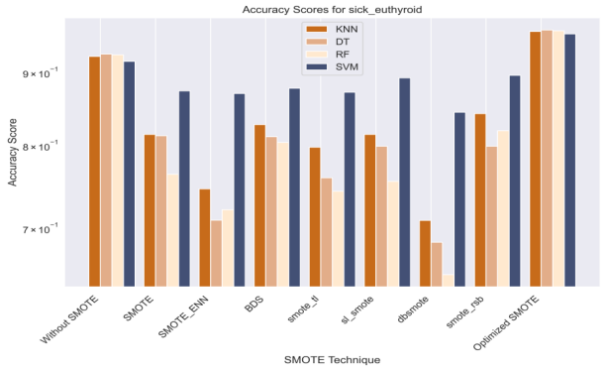

Figure 6: Accuracy Scores for Sick_Euthyroid

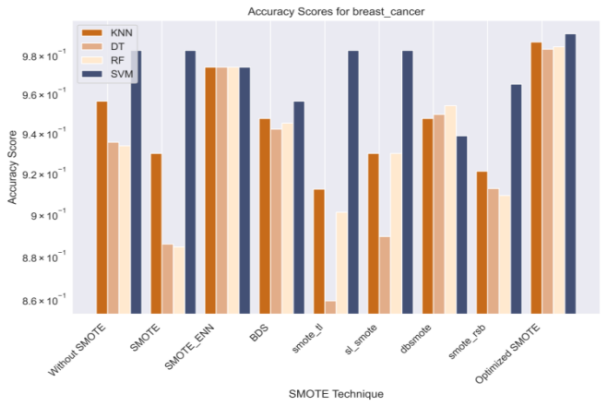

Figure 7: Accuracy Scores for breast cancer

The results indicate that the developed BES-optimized SMOTE worked well for all four datasets. So, to scale the data, BES-optimized SMOTE technique is decided to be used on these specific datasets. The scaled datasets were split into training and testing data, encrypted using Paillier homomorphic encryption [23] and stored separately for further research. The time required for encryption is mentioned in table 22 below.

Table 22. Dataset and their Encryption Times

Dataset Name	Dimensions	Time Required (in minutes)
Breast cancer (train)	(490, 31)	80
Breast cancer (test)	(114, 31)	15
Arrhythmia (train)	(402, 177)	332
Arrhythmia (test)	(91, 177)	81

The presented graph in Figure 8 illustrates the relationship between the number of values in the selected datasets and the corresponding time required for their encryption. The graph reveals a linear increase in encryption time with respect to the number of values, highlighting a consistent trend across the analyzed datasets and hence, we can infer that homomorphic encryption is computationally costly option for privacy preserved classification. So, developing an efficient approach would be our endeavor.

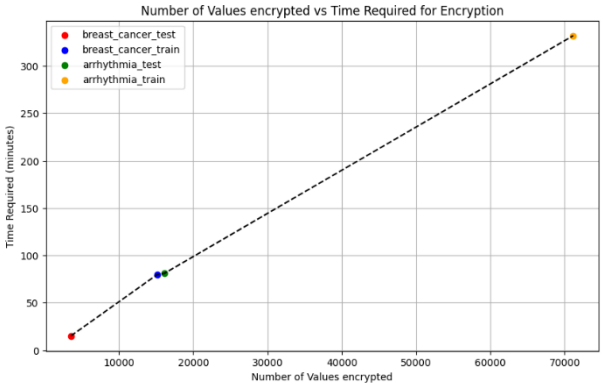

Figure 8. Encryption Time Analysis for selected datasets

6. **Conclusion and Future Scope**

This study undertook a comprehensive analysis of nine diverse SMOTE variants, assessing their impact on five medical datasets using KNN, DT, RF and SVM models. The outcomes of these evaluations have provided crucial insights into the appropriateness of SMOTE techniques for addressing class imbalance in medical data.

Upon careful analysis of the results, it is evident that the developed 'BES-optimized SMOTE' variant exhibited superior performance on specific datasets. For the 'heart' dataset, along with 'BES optimized SMOTE', 'SL-SMOTE' also demonstrated remarkable improvements in accuracy and precision. Conversely, BD-SMOTE also emerged as the optimal choice for the 'Arrhythmia' dataset next to the 'BES-optimized SMOTE'.

In the case of the 'breast cancer' dataset, 'SMOTE-ENN' also emerged as the most effective oversampling technique, yielding significant advancements in both accuracy and precision following BES-optimized SMOTE. For the hypothyroid and 'sick euthyroid' datasets, the 'BES-optimized SMOTE' variant showed better results.

Looking ahead, the insights gathered from this study will help guide future investigations. The focus will be on leveraging the identified best-performing SMOTE variants in tandem with privacy-preserving KNN classification. This strategic fusion aims to bolster the security and confidentiality of patient data, a critical aspect in the realm of healthcare analytics.

As per the obtained results, breast cancer and arrhythmia datasets gave good results with BES-optimized SMOTE. Hence, BES-optimized SMOTE technique is used for scaling above mentioned datasets and use encrypted version of these scaled datasets in our future work on efficient privacy-preserving KNN classification.

References

[1] Agrawal, R., Delen, D., & Benjamin, D. B. (2019). Clinical Intervention Research with EHR: A Big Data Analytics Approach.

[2] Abouelmehdi, K., Beni-Hessane, A., & Khaloufi, H. (2018). Big healthcare data: preserving security and privacy. *Journal of big data*, *5*(1), 1-18.

[3] Liu, X., Lu, R., Ma, J., Chen, L., & Qin, B. (2015). Privacy-preserving patient-centric clinical decision support system on naive Bayesian classification. IEEE journal of biomedical and health informatics, 20(2), 655-668

[4] Gutierrez, O., Saavedra, J. J., Zurbaran, M., Salazar, A., & Wightman, P. M. (2018, October). UserCentered Differential Privacy Mechanisms for Electronic Medical Records. In 2018 International Carnahan Conference on Security Technology (ICCST) (pp. 1-5). IEEE.

[5] Raisaro, J. L., Troncoso-Pastoriza, J., Misbach, M., Sousa, J. S., Pradervand, S., Missiaglia, E., ... & Hubaux, J. P. (2018). Medco: Enabling secure and privacy-preserving exploration of distributed clinical and genomic data. IEEE/ACM transactions on computational biology and bioinformatics

[6] Huang, L., Shea, A. L., Qian, H., Masurkar, A., Deng, H., & Liu, D. (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. Journal of Biomedical Informatics, 103291.

[7] Hahn, Seok-Ju, and Junghye Lee. "Privacy-preserving federated bayesian learning of a generative model for imbalanced classification of clinical data." arXiv preprint arXiv:1910.08489 (2019).

[8] Culnane, C., Rubinstein, B. I., & Teague, V. (2017). Health data in an open world. arXiv preprint arXiv:1712.05627

[9] Ilavarasi, A. K. (2020, October). Class imbalance learning for Identity Management in Healthcare. In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 995-1000).

[10] He, H., & Garcia, E. A. (2008). Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, (9), 1263-1284.

[11] Napierala, K., & Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. Journal of Intelligent Information Systems, 46, 563-597.

[12] Mrudula, O., & Mary Sowjanya, A. (2020). A prediction model for imbalanced datasets using machine learning. J Crit Rev, 7(08), 2132-2140.

[13] Sowjanya, A. M., & Mrudula, O. (2023). Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. Applied Nanoscience, 13(3), 1829-1840.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, June 2002.

[15] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset," *Sensors*, vol. 22, p. 3246, Apr. 2022.

[16] M. Lamari, N. Azizi, N. E. Hammami, A. Boukhamla, S. Cheriguene, N. Dendani, and N. E. Benzebouchi, "SMOTE–ENN-based data sam- pling and improved dynamic ensemble selection for imbalanced medical data classification," in *Advances on Smart and Soft Computing*, pp. 37– 49, Springer Singapore, Oct. 2020.

[17] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over- sampling method in imbalanced data sets learning," in *Lecture Notes in Computer Science*, pp. 878–887, Springer Berlin Heidelberg, 2005.

[18] Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. Information Sciences, 465, 1–20.

[19] Nizam-Ozogur, H., & Orman, Z. (2024). A heuristic-based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data. Expert Systems, e13596.

[20] F. Li, S. Li, C. Zhu, X. Lan, and H. Chang, "Class-imbalance aware CNN extension for high resolution aerial image based vehicle localization and categorization," in Proceedings of the 2017 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, China, June 2017.

[21] C. Zhang, K. Tan, and R. Ren, "Training cost-sensitive deep belief networks on imbalance data problems," in Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, Canada, July 2016.

[22] Liu, Y., Li, X., Chen, X., Wang, X., & Li, H. (2020). High-performance machine learning for large-scale data classification considering class imbalance. Scientific Programming, 2020.

[23] P. Paillier, "Public-key cryptosystems based on composite degree residu- osity classes," in *Advances in Cryptology — EUROCRYPT '99*, pp. 223– 238, Springer Berlin Heidelberg.

[24] Chhabra, A., Hussien, A. G., & Hashim, F. A. (2023). Improved bald eagle search algorithm for global optimization and feature selection. *Alexandria Engineering Journal*, *68*, 141-180.

[25] M.-W. Huang, C.-H. Chiu, C.-F. Tsai, and W.-C. Lin, "On combining feature selection and over-sampling techniques for breast cancer predic- tion," *Applied Sciences*, vol. 11, p. 6574, July 2021.

[26] A.Ishaq,S.Sadiq,M.Umer,S.Ullah,S.Mirjalili,V.Rupapara,and M. Nappi, "Improving the prediction of heart failure patients' survivalusing SMOTE and effective data mining techniques," *IEEE Access*,vol. 9, pp. 39707–39716, 2021.

[27] N. A. Azhar, M. S. M. Pozi, A. M. Din, and A. Jatowt, "An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2022

[28] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, Mar. 2013.

[29] M.Waqar,H.Dawood,H.Dawood,N.Majeed,A.Banjar,andR.Alhar-bey, "An efficient SMOTE-based deep learning model for heart attackprediction," *Scientific Programming*, vol. 2021, pp. 1–12, Mar. 2021.

[30] Arafa, A., El-Fishawy, N., Badawy, M., & Radad, M. (2022). RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification. Journal of King Saud University-Computer and Information Sciences, 34(8), 5059-5074.

[31] Fonseca, J., & Bacao, F. (2023). Geometric SMOTE for imbalanced datasets with nominal and continuous features. Expert Systems with Applications, 234, 121053.

[32] Sujitha, R., and B. Paramasivan. "Optimal progressive classification study using SMOTE-SVM for stages of lung disease." Automatika 64, no. 4 (2023): 807-814.

[33] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level- SMOTE: Safe-level-synthetic minority over-sampling TEchnique for handling the class imbalanced problem," in *Advances in Knowledge Discovery and Data Mining*, pp. 475–482, Springer Berlin Heidelberg, 2009.

[34] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "DB- SMOTE: Density-based synthetic minority over-sampling TEchnique," *Applied Intelligence*, vol. 36, pp. 664–684, Apr. 2011.

[35] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, "SMOTE-RSB: a hybrid preprocessing approach based on oversampling and under- sampling for high imbalanced data-sets using SMOTE and rough sets theory," *Knowledge and Information Systems*, vol. 33, pp. 245–265, Dec. 2011.

[36] W. S. Andras Janosi, "Heart-disease," UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X.

[37] B. A. H. Guvenir, "Arrhythmia," UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C5BS32.

[38] R. Quinlan, "Thyroid disease," UCI Machine Learning Repository, 1986. DOI: https://doi.org/10.24432/C5D010.

[39] M. O. S. N. Wolberg, William and W. Street, "Breast Cancer Wis- consin (Diagnostic)," UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C5DW2B

[40] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, pp. 986–996, Springer Berlin Heidelberg, 2003.

[41] Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for machine learning." Journal of Applied Science and Technology Trends 2, no. 01 (2021): 20-28.

[42] Cutler, Adele & Cutler, David & Stevens, John. (2011). Random Forests. 10.1007/978-1-4419-9326-7-5.

[43] N. Cristianini and E. Ricci, "Support vector machines," in Encyclopedia of Algorithms, pp. 928–932, Springer US, 2008.

[44] C. Data61, "Python paillier library." https://github.com/data61/ python-paillier, 2013