Hybrid Multi-Step SMOTE-ENN Algorithm for Enhanced Breast Cancer Machine Learning Classification.

Abdullahi Muhammad[1], *Solomon Adelowo Adepoju*[2], *Sulaimon A. Bashir*[3] *and Opeyemi Aderiike Abisoye*[4]

[1]*Abdullahi Muhammad*, MTech, is a Student of Federal University of Technology, Minna, Nigeria. His teaching and research interests are in Artificial Intelligence, Machine Learning and Internet of Things. Mobile: +2347034245467, Email address: abdul8095@gmail.com

[2]*Solomon Adelowo Adepoju*, PhD, is a senior lecturer in the department of computer science, Federal University of Technology Minna, Nigeria. His teaching and research interests are in the area of Human Computer Interaction (HCI), web/data mining and ICT4D. Mobile +2348035829748, Email Address: solo.adepoju@futminna.edu.ng

[3]*Sulaimon A. Bashir,* PhD, is a senior lecturer in the department of computer science, Federal University of Technology Minna, Nigeria. His teaching and research interests are in the area of Human Computer Interaction (HCI), and Artificial Intelligence. Mobile +2349097622911, Email Address: bashirsulsimon@futminna.edu.ng

[4]*Opeyemi Aderiike Abisoye*, PhD, is a senior lecturer in the department of computer science, Federal University of Technology Minna, Nigeria. Her teaching and research interests are in Artificial intelligence. Mobile: +2348060546074, Email Address: o.abisoye@futminna.edu.ng

Abstract

Early detection of breast cancer is essential to prevent and reduce patient death roll, cost of operation on a patient, and provide early awareness for quick treatment. Researchers have proposed various methods for diagnosing and preventing breast cancer in women. It's identified that prediction accuracy highly depends on the size, quality, and distribution of the dataset class (balanced or unbalanced data class), considering the health sector most public data available including the breast cancer dataset is still imbalanced and those that are addressed using oversampling techniques could lead to the addition of noise in the source data. Hence this study aims to adopt a hybrid multi-step oversampling SMOTE-ENN (Synthetic Minority Oversample Technique and Edited Nearest Neighbor) to improve the quality of the breast cancer dataset and limit the possibility of noise in the data. As a result, this improves the dataset used to train the machine learning model, and 0.99% accuracy is achieved during model evaluation.

Key words: cancer, oversampling, machine learning, smote, ENN, accuracy, dataset

1.0 Introduction

Actually, there are more than a hundred distinct forms of cancer, in which every tissue in the body has a chance of developing cancerous cells, some even create several different types. It identifies that each type of cancer has unique characteristics. However, it appears that these diverse tumors are caused by largely comparable underlying mechanisms. In general, cancer is a disease that is developed as a result of a few body cells that multiply uncontrollably and spread to various parts of the body. Cancer can surge anywhere among the trillions of cells in the human body, and new cells emerge via cell multiplication and growth (Haldorsen et al., 2019).

Breast cancer is a type of tumor that emerges in the breast cell or tissue of women around the globe, which causes major fatalities among women. Breast cancer is known to be the second most cancer that results in death in Taiwan (Huang et al., 2008). The advancement of biomedical and computer technology has led to the recording of a number of clinical features associated with breast cancer. Many researchers have thought about using patient clinic data to forecast breast cancer for patients in order to combat the sharp rise in breast cancer cases.

Artificial intelligence is the domain of study that focuses on the development of an intelligent system to perform an intelligent task that is previously done by humans (Chen et al., 2020). However, machine learning is a subdomain in artificial intelligence which has the capacity to analyze data and draw relationships and essential features from a dataset. Machine learning can also be used to create computational models that more accurately explain data. Another critical need is for machine learning to be able to automatically detect cancer in order to assist doctors in reliably performing comparable tasks in huge numbers (Bakri et al., 2021). Considering the predictive capability and success rate that has been identified in the adoption of machine learning, large number of researchers and scientist are picking interest in the utilization of machine learning. However, medical domain is now adopting the predictive capability of machine learning due to the easy usage for diagnosing patient by medical professional.

An Imbalance Data set is said to be imbalance if the class proportions are skewed due to an uneven classification data set. Classes that account for the most sizeable portion of the data collection are referred to as the "majority classes." Unbalanced data may be used to illustrate a dataset's classes, such as a fraud and non-fraud data in which they are not equally distributed (Ali et al., 2019). In real word dataset majorly in the field of health data suffers from the issue of imbalance distribution of data sample from various data point categories. The most know traditional approach of addressing the issue of data imbalance is the use of Sampling Techniques. the oversampling of minority class and under sampling of the majority class is part of the traditional approach of tackling data imbalance. They various approach used for oversampling and under sampling data includes; cross fold validation, Synthetic Minority Oversampling Techniques (SMOTE), Cluster-Based Over Sampling (CBO), Edited Nearest Neighbor (ENN) Decision Tree, Naïve Bayes, and support Vector Machine (Domingues et al., 2018). The minority class is oversampled in the SMOTE algorithm by using each minority class sample and adding synthetic examples along the line segments connecting any or all of the k minority class nearest neighbor (Hasib et al., 2020).

This study identifies that early detection, classification and diagnosis of breast cancer patient will significantly minimize and prevent many deaths of cancerous patient. It’s essential to develop automated approach for efficient detection and classification of breast cancer in patient (Murtaza et al., 2020). In the domain of health, the issue of data imbalance has become problem face by many researchers, which hinder the performance of machine learning algorithm precision. (Susan & Kumar, 2019) identify that the oversampling of minority class introducing noises (irrelevant data) to the main data and there is high possibility of duplicate features in the dataset, as a result of this, overfitting of data occurs. The adoption of oversampling techniques such as ADASYN or SMOTE do not fully address the issues identified while generating synthetic data. Hence, this study introduces a more sophisticated approach thus, the multi-step SMOTE-ENN techniques.

Primarily, this study proposes a hybrid Multi-Step Minority Over Sample Techniques and Edited Nearest Neighbor for balancing breast cancer dataset class distribution, and to improve the diagnosing capability and prediction accuracy of machine learning algorithms. It’s essential to collect breast cancer dataset from an online repository and balance the data class distribution using SMOT-EENN multi-step approach, the resulted data is adopted for training the machine learning algorithm and finally an evaluation is carried out to test the performance of the developed model using standard performance metrics thus, accuracy, precision and recalls.

The significant application of this knowledge contribution is countless, most especially in medical field. Hospitals and health related center could utilize the improve model for accurate diagnosing and predicting the present of cancerous cell in women breast. This study will help various nation to address the high death rate that are cause due to lack of quick, early and accurate prediction of breast cancer. Adopting this model in real life will significantly decrease the expense of purchasing machinery or equipment used in the diagnosing breast cancer. However, this research work only considered the use of machine learning algorithm to develop the breast cancer prediction model. The use of deep learning algorithm is not considered in this sturdy.

2.0 Literature Review

Based on the research work of (Kothari et al., 2020) which identify that the tissue of adipose is a complex organ in the endocrine, which play a significant role in both cancer and obesity. It is commonly known that the female breast has a lot of adipose tissue, which is typically associated with excessive body fat. The adipose tissue organ encompasses adipocytes which is an extracellular matrix and immune cell, which take part in the dynamic change of woman breast ranging from puberty stage, pregnancy, lactation and involution. The researcher carried out a comprehensive review on adipose tissue in other to reveal the essence of adipose tissues in the development cycle of breast. The biology of breast adipose tissue is also compressively understood. Due to the multiple layers of adipose tissue in biological structure, and only few of the layer are known while many other parts are still not understood. It's concluded by the researchers that it's advised maintain a good healthy lifestyle and intake of energy is a very essential factor.

An hybrid principal component analysis (PCA) and related data mining models is proposed by (Wang & Yoon, 2015), which employs a principle component analysis method to reduce the feature space, to test the impact of feature space reduction. Two frequently used test data sets, Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995), are used to assess the performance of these models. This study compares various patient clinical records to find an accurate model that can predict the likelihood of developing breast cancer. Support vector machines (SVM), artificial neural networks (ANN), Naive Bayes classifiers, and AdaBoost trees are the four data mining models used in this paper. The outcomes of this analysis show a thorough trade-off between the adopted strategies and also give a comprehensive analysis of the models.

Its identify that the country such as American contain significant high record of breast cancer case and comes second in the death rate of women with breast cancer. A variety of methods have been used in data mining classification task to create an accurate prediction model and analyze the noteworthy risk factors. Many domains have utilized data mining techniques to gain hidden knowledge from that, this domain include the engineering, science, medicine and business. With the help of anthropometric information and parameters gathered during routine blood testing, this study aims to predict the trace of breast cancer. The researcher used different variety of classification methods, including K-NN, ANN, Decision Trees, and Naive Bayesian, and discovered that artificial neural networks (ANN) classify the attribute with the highest accuracy of (80.00%) (Ghani et al., 2019).

(Chaurasia et al., 2018) reveal that the second most occurring type of cancer in woman recorded in the world is the Breast cancer, in 2004 alone, there were about 1.1 million cases reported. This research paper's goal is to present a report on breast cancer in which the researchers adopted the most recent technological developments to create prediction models for the prognosis of the disease. The researchers adopted a large dataset (683 breast cancer cases) for developing the prediction models using three well-known data mining algorithms thus, Naive Bayes, RBF Network, and J48. The results showed that the Naive Bayes is the best predictor model with 97.36% accuracy on the holdout sample (this prediction accuracy is better than any reported in the literature), RBF Network came in second with 96.77% accuracy, and J48 came in third with 93.41% accuracy (based on the average accuracy Breast Cancer dataset).

One of the most prevalent illnesses affecting women worldwide is breast cancer. Numerous studies have been carried out to forecast the survival indicators, but the majority of these analyses were mostly carried out using fundamental statistical techniques. Prediction models were created using decision tree, random forest, neural networks, extreme boost, logistic regression, and support vector machines to identify the significant prognostic factors of breast cancer survival rate. All algorithms produced close results in terms of model accuracy and the calibration measure, with the lowest result coming from decision tree (accuracy = 79.8%) and the highest result coming from random forest (accuracy = 82.7%).

3.0 Methodology

This section introduces the use of Multi-Step Data Sampling Techniques using conceptual diagrams to illustration of the adopted approach, tools used, data collection and data analysis approach.

A. Multi-Step SMOTE-ENN

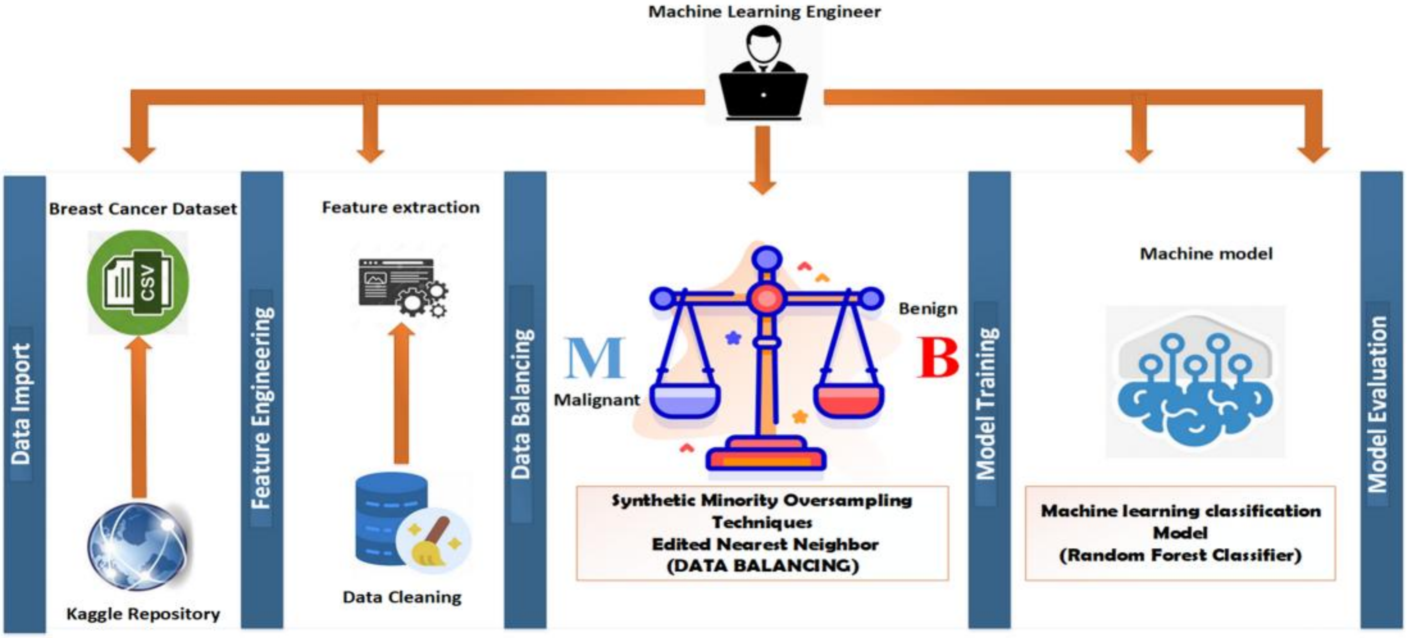


Fig 1. Multi-Step Breast Cancer Prediction Model Architecture

The figure 1 comprehensively illustrates the conceptual idea of the proposed Multi-Step SMOTE-ENN data balancing techniques for efficient and high accuracy prediction model. Based on figure 1 it's depicted in the first stage that the dataset is collected from an online data science repository call Kaggle, and the dataset is downloaded locally on the system in a csv format. The next stage involves data cleaning and feature extractions, which tends to remove columns and data point with empty, inconsistent and noisy data in the dataset. However, it's essential to extract and transform the dataset to format that is easily understood and processed by that system. Furthermore, the pre-processed data is balanced using the propose Multi-Step SMOTE-ENN data balancing techniques, this approach uses the hybridization of Synthetic Minority Oversampling Techniques and Edited Nearest Neighbor strategy to generate synthetic sample of the minority class distribution. The balanced breast cancer dataset can now be utilized in the training of machine learning algorithm, thus random forest classifier for prediction and diagnosing breast cancer patient. Finally, the machine learning developer is indicated with the avatar at the top of the figure which shows how machine learning engineering perform action and stages of the development of breast cancer prediction model.

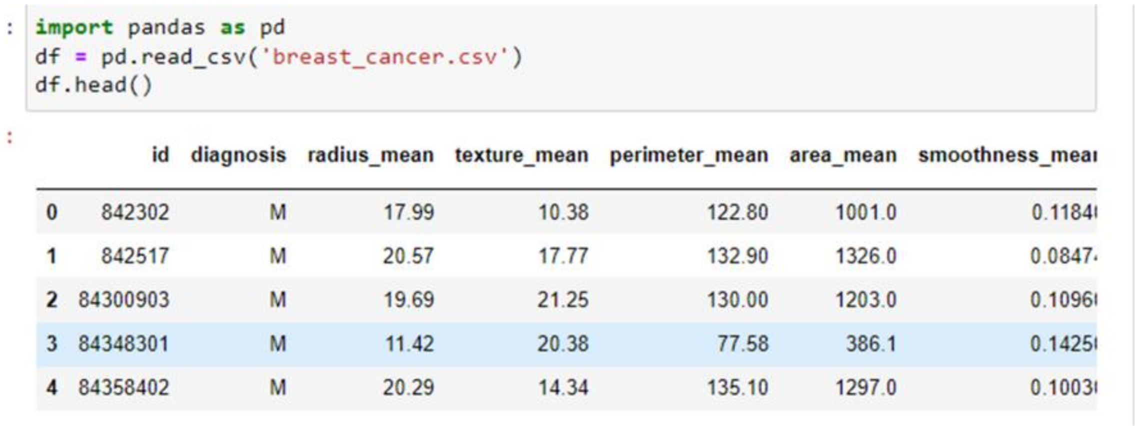
B. Tool Used

The developed breast cancer prediction model using the multi-step SMOTE-ENN data balancing approach and it's implemented within the Jupyter notebook integrated development environment. Python is considered as the choice of programming language in this study due to the large number of machine learning module available for performing machine learning operations. Module such as Numpy, Pandas, and Sklearn are popularly used for artificial intelligent operations. Other tools used in this study includes Microsoft Visio for designing the various conceptual diagram and Mendeley for academic reference and citation.

C. Data Collection

The breast cancer dataset is downloaded from the popular online open source data science repository call Kaggle (Casper et al., 2020). The dataset is downloaded in Comma Separate Value format and it's imported into the jupyter notebook environment using the panda module. The breast cancer dataset contains 567 data sample (rows) and 33 feature (columns). The diagnosis attribute indication the depended variable or the predicted attribute Malignant (for non-cancerous prediction) and Benign (for cancerous class prediction), both diagnosis class contain inequivalent number of data distribution based on data exploration.

The figure 2 below shows the dataset sample records within the jupyter notebook environment.



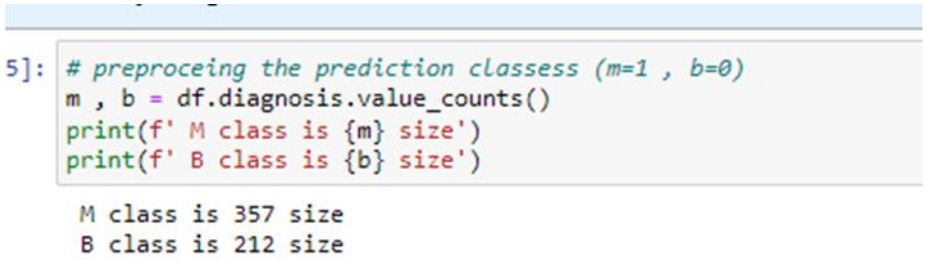
```
: import pandas as pd
  df = pd.read_csv('breast_cancer.csv')
  df.head()
```

:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mea
0	842302	M	17.99	10.38	122.80	1001.0	0.1184
1	842517	M	20.57	17.77	132.90	1326.0	0.0847
2	84300903	M	19.69	21.25	130.00	1203.0	0.1096
3	84348301	M	11.42	20.38	77.58	386.1	0.1425
4	84358402	M	20.29	14.34	135.10	1297.0	0.1003

Fig 2. Breast Cancer Sample Dataset.

The figure 2 show five sample record from the breast cancer dataset. It identifies that the dataset contains an id, diagnosis, radius means and lot more up to 33 feature sample. Its visually shown that most of the data entry are floating point in number, while the diagnosis attribute entry contain ‘M’ for Malignant and ‘B’ Benign as values. The class distribution is visually explored in the figure 2 below showing the class distribution indicating unbalance dataset.



```
5]: # preproceing the prediction classess (m=1 , b=0)
    m , b = df.diagnosis.value_counts()
    print(f' M class is {m} size')
    print(f' B class is {b} size')

     M class is 357 size
     B class is 212 size
```

D. *Methodology*

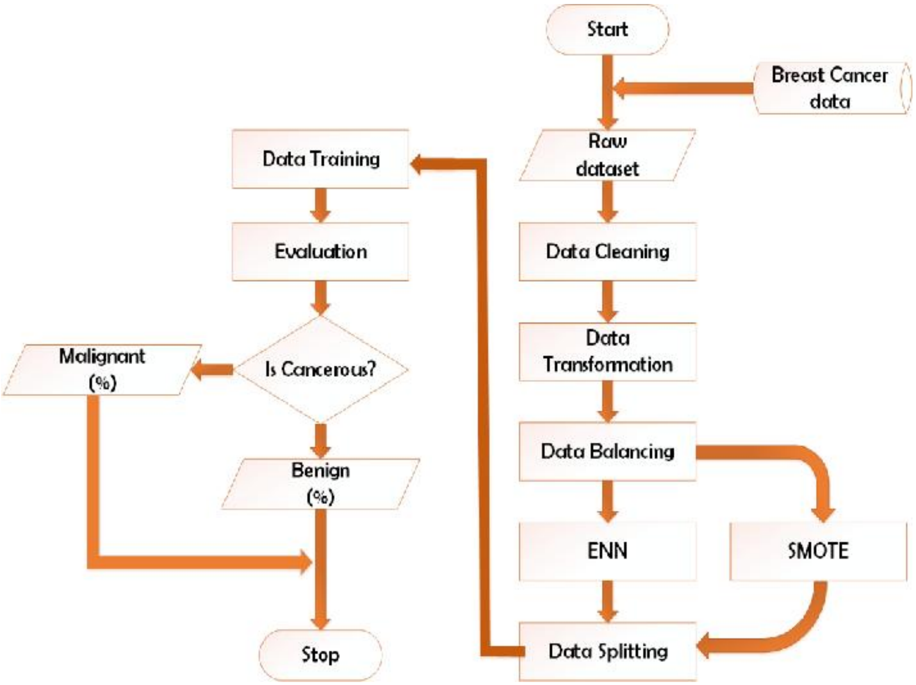
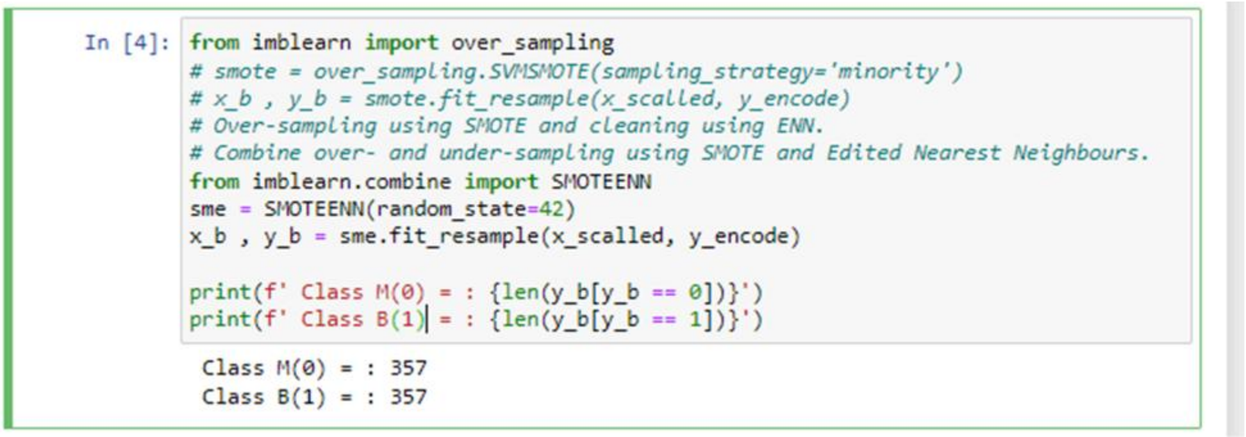


Fig 4 Conceptual Flow Data Diagram for Multi-Step Breast Cancer Prediction Model

The designed figure indicates the algorithm step carried out for the development of breast cancer dataset. Based on the figure the raw dataset is loaded from the database into the jupyter environment and data cleaning operation take place. The data cleaning helps in removing inconsistent, and noisy data from the raw dataset. The next step is the data transformation, thus converting the dataset into numerical and standard format for machine learning understanding and efficient classification. The diagnosis attribute is transformed into numerical (M=0 and B=1). Furthermore, it's essential to balance the breast cancer dataset and reduce the probability of introducing noise into the dataset. This research adopted the multi-step SMOTE-ENN approach for balancing the breast cancer dataset by generating synthetics samples using SMOTE and filtering synthetic data based on K-nearest neighbor algorithm. The data balancing stage is immediately followed by data splitting, model training using Random Forest classifier and finally model evaluation using standard performance metrics (Accuracy, precision and recall).

4.0 Results and Discussions

This section introduces the training process of the breast cancer prediction model using the balanced dataset generated with the Multi-Step SMOTE-ENN algorithm, and the result gotten will be discussed in this section, finally a comparative analysis will be done with existing study.

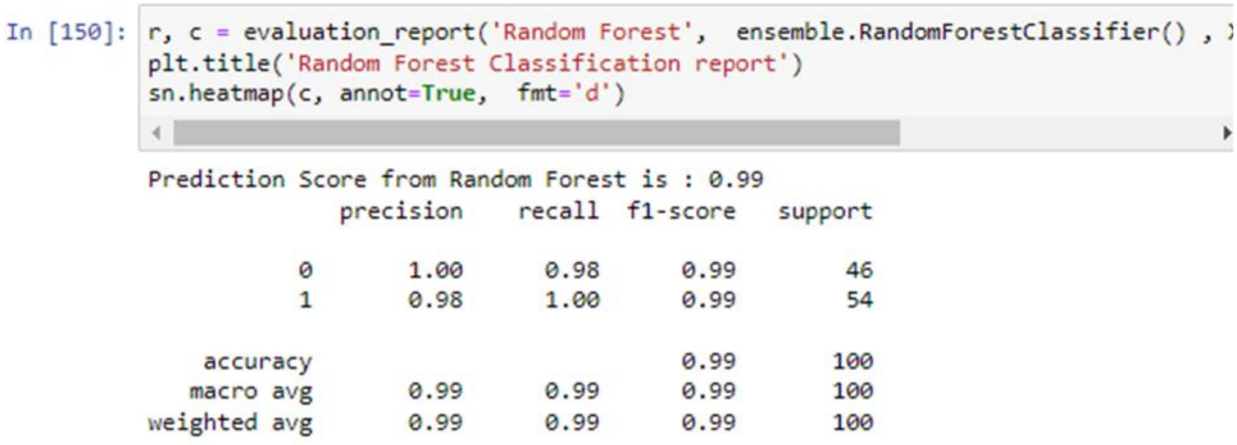


```
In [4]: from imblearn import over_sampling
        # smote = over_sampling.SVMSMOTE(sampling_strategy='minority')
        # x_b , y_b = smote.fit_resample(x_scalled, y_encode)
        # Over-sampling using SMOTE and cleaning using ENN.
        # Combine over- and under-sampling using SMOTE and Edited Nearest Neighbours.
        from imblearn.combine import SMOTEENN
        sme = SMOTEENN(random_state=42)
        x_b , y_b = sme.fit_resample(x_scalled, y_encode)

        print(f' Class M(0) = : {len(y_b[y_b == 0])}')
        print(f' Class B(1) = : {len(y_b[y_b == 1])}')

         Class M(0) = : 357
         Class B(1) = : 357
```

Fig 5. SMOTE-ENN Data Balancing Techniques

The figure 5 above shows the python implementation of the Multi-Step SMOTE-ENN Data Balancing Techniques using the Python Imblearn module. After balancing the breast cancer dataset, it can be seen in the figure above that the diagnosis prediction variable class distribution is balanced with a data sample of 357 in each class.

The balance dataset is used for training the machine learning algorithm, for evaluating the balance dataset the Random Forest Classifier is used to develop the breast cancer prediction model and the figure below show the performance measure of the adopted techniques.



```
In [150]: r, c = evaluation_report('Random Forest',  ensemble.RandomForestClassifier() ,
          plt.title('Random Forest Classification report')
          sn.heatmap(c, annot=True,  fmt='d')

          Prediction Score from Random Forest is : 0.99
                        precision    recall  f1-score   support

                     0       1.00      0.98      0.99        46
                     1       0.98      1.00      0.99        54

              accuracy                           0.99       100
             macro avg       0.99      0.99      0.99       100
          weighted avg       0.99      0.99      0.99       100
```

Fig 6. Breast Cancer Prediction Accuracy, Recall and F-1 Score

The figure 6 shows the performance measure of the random forest classifier model using the balance breast cancer dataset. The sturdy considers Accuracy, Precision, Recall and F1-score for measuring the performance of the developed machine learning model. Its identity that the developed model achieved an accuracy of 0.99%.

Table 1. Comparative Analysis of Result's

S/N	Machine learning techniques	Data balancing techniques	Accuracy
1	Random Forest (Bakri et al., 2021)	Cross-fold validation	0.99%
2	Support Vector Machine (Bakri et al., 2021)	Cross-fold validation	0.98%
3	Logistic Regression (Bakri et al., 2021)	Cross-fold validation	0.93%
4	Random Forest Classifier (Proposed)	None	0.97%
5	Random Forest Classifier (Proposed)	Proposed SMOTE-ENN	0.99%

The table 1. shows significant importance of data balancing in existing work and the proposed method.

5.0 Conclusion and Recommendation

Generally, data balancing techniques has great significant impact on improving the quality of dataset used for machine learning algorithm. The machine learning will be learning from dataset that has equal numbers of data point distribution from each prediction class. As a result, the prediction accuracy of the machine learning algorithm tends to increase. This study adopts the multi-step data balancing techniques that hybridize Synthetic Minority Oversampling Techniques SMOTE and Edited Nearest Neighbor (ENN). The study carried out an evaluation using Random Forest classifier with the raw dataset without balancing and an accuracy of 79% is achieved. However, the same Random Forest Classifier is trained using the balance dataset generated using the Multi-step SMOTE-ENN and an accuracy of 99% is achieved. Hence, the experiment proved that the proposed data balancing techniques efficiently improve the Machine Learning performance.

Furthermore, its highly recommended for medical domain to adopt AI in diagnosing breast cancer in women, this can significantly reduce the death rate statistic of any economy. However, most proposed and existing method for predicting, and diagnosing breast cancer are based on machine learning algorithm and not deep learning algorithm. This occurs as a result of limited dataset available on breast cancer. It's recommended to hybridize multiple dataset with similar feature for training a more robust model using deep learning.

References

Ali, H., Salleh, M. N. M., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science*, *14*(3), 1552–1563. https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563

Bakri, M., Amin, & Ekawati, I. (2021). *Breast Cancer Prediction Model Using Machine Learning*. *002*(August), 1–8.

Casper, B., Bojer, S., & Meldgaard, J. P. (2020). *Learnings from Kaggle ' s Forecasting Competitions*.

Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms and Computational Technology*, *12*(2), 119–126. https://doi.org/10.1177/1748301818756225

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. *IEEE Access*, *8*, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Domingues, I., Amorim, J. P., Abreu, P. H., Duarte, H., & Santos, J. (2018). Evaluation of Oversampling Data Balancing Techniques in the Context of Ordinal Classification. *Proceedings of the International Joint Conference on Neural Networks*, *2018-July*(October). https://doi.org/10.1109/IJCNN.2018.8489599

Ghani, M. U., Alam, T. M., & Jaskani, F. H. (2019). Comparison of Classification Models for Early Prediction of Breast Cancer. *3rd International Conference on Innovative Computing, ICIC 2019*, *January 2020*. https://doi.org/10.1109/ICIC48496.2019.8966691

Haldorsen, I. S., Lura, N., Blaakær, J., Fischerova, D., & Werner, H. M. J. (2019). What Is the Role of Imaging at Primary Diagnostic Work-Up in Uterine Cervical Cancer? *Current Oncology Reports*, *21*(9). https://doi.org/10.1007/s11912-019-0824-0

Hasib, K. M., Iqbal, M. S., Shah, F. M., Mahmud, J. Al, Popel, M. H., Showrov, M. I. H., Ahmed, S., & Rahman, O. (2020). A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem. *Journal of Computer Science*, *16*(11), 1546–1557. https://doi.org/10.3844/JCSSP.2020.1546.1557

Huang, C. L., Liao, H. C., & Chen, M. C. (2008). Prediction model building and feature selection with support vector machines in breast cancer diagnosis. *Expert Systems with Applications*, *34*(1), 578–587. https://doi.org/10.1016/j.eswa.2006.09.041

Kothari, C., Diorio, C., & Durocher, F. (2020). The importance of breast adipose tissue in breast cancer. *International Journal of Molecular Sciences*, *21*(16), 1–33. https://doi.org/10.3390/ijms21165760

Murtaza, G., Shuib, L., Abdul Wahab, A. W., Mujtaba, G., Mujtaba, G., Nweke, H. F., Al-garadi, M. A., Zulfiqar, F., Raza, G., & Azmi, N. A. (2020). Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*, *53*(3), 1655–1720. https://doi.org/10.1007/s10462-019-09716-5

Susan, S., & Kumar, A. (2019). SSO Maj -SMOTE-SSO Min : Three-step intelligent pruning of majority and minority samples for learning from imbalanced datasets. *Applied Soft Computing Journal*, *78*(February), 141–149. https://doi.org/10.1016/j.asoc.2019.02.028

Wang, H., & Yoon, S. W. (2015). Breast cancer prediction using data mining method. *IIE Annual Conference and Expo 2015, October*, 818–828.