ISSN 2278-3091
Volume 14, No.3, May - June 2025
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse041432025.pdf
https://doi.org/10.30534/ijatcse/2025/041432025

A Novel Resampling Model for Classifying an Imbalanced Breast Cancer Dataset

Hana Babiker Nassar[1], Abdelrahman Elsharif Karrar[2], Waleed Ibrahim Osman[3]
[1] College of Computer studies
The National Ribat University, Sudan
hananassar2011@gmail.com
[2] College of Computer Science and Engineering
Taibah University, Saudi Arabia
akarrar@taibahu.edu.sa
[3] College of Computer Studies
The National Ribat University, Sudan
walidibrahimosman@gmail.com

Received Date: April 20, 2025 Accepted Date: May 23, 2025 Published Date: June 06, 2025

ABSTRACT

Breast cancer is a significant health concern within medical care systems, necessitating accurate classification. The patient data are recorded and statistically analyzed, revealing an increasing number of files. And then transferred to the statistics department with increasing numbers. This study investigates breast cancer data imbalance utilizing Khartoum State Hospital. An imbalanced data problem occurs when one class has a significantly larger number of samples than another. To address this, resampling, attribute selection, handling missing values, classifier algorithms (ANN, REP TREE, SVM, J48), and ensemble learning models were employed. The base classifier yielded the first result, the meta-learning algorithms (Bagging, Boosting, and Random Subspace) the second, and an ensemble model the third. The boosting with the J48 ensemble model achieved the highest accuracy, **95.2797 %,** outperforming bagging with j48 (**90.559%**) and random subspace with j48 (**84.2657%**).

Key words: Breast Cancer, Imbalance dataset, attribute selection, Resampling Methods, Ensemble Model.

1. INTRODUCTION

When breast cells start to proliferate uncontrollably, the first stage of breast cancer starts. These cells can be felt as a bump or seen on X-rays to create a tumor. New swelling or buildup is a common sign of breast cancer. Breast malignancies can be spherical, soft, or painful, but a firm mass with uneven borders that is painless is more likely to be cancer. They are even capable of causing pain st[1], [2]. Breast cancer can be either benign or malignant, and it develops when cells grow and divide uncontrollably. Since a few risk factors raise a woman's likelihood of acquiring breast cancer, researchers have attempted to pinpoint the precise cause of the disease. One of the most important considerations when choosing a treatment choice is the cancer stage, which is determined by the Tumor, Nodes, and Metastasis (TNM) system. From stage 0 (the least advanced) to the most advanced stage, this type of tumor is identified [3]. In many real-world applications, imbalanced data classification frequently occurs. The underlying training set is assumed to be uniformly distributed in many classification approaches. However, when the training set is extremely unbalanced, those methods suffer from a serious bias issue. Severe learning issues for unbalanced classes are a common concern in the actual world.

In the real world, the majority of the data is unbalanced. This circumstance arises when the target class's distribution.Across the various class levels, it is irregular. One of the most difficult issues in machine learning is the classification of this kind of data, which has attracted a lot of attention lately. A class with more instances is referred to as a substantial mass in an unbalanced data set, whereas a class with a relative and several instances is remembered as a minor class [4]. To address this imbalanced data issue, several approaches were created, including sampling techniques, ensemble learning, cost-sensitive learning, feature selection, and algorithmic modification[5].

2. LITREATURE REVIEW

One of the two groups of breast cancer data in this study has more samples than the other, indicating that the data is imbalanced. To categorize this imbalanced data, a variety of pre-processing methods are used, such as resampling, attribute selection, and handling missing values. Afterward, several classifier models are constructed. In the real world, the majority of data is imbalanced. This circumstance arises when the target class's distribution across the various class levels is irregular.

One of the most challenging issues in machine learning is the classification of this type of data, which has garnered considerable attention recently. According to G.I. Salama et al.., the Quinlan C4.5 algorithm was implemented using Decision Tree J48, which produced a pruned and unpruned C4.5 tree. The decision trees produced by J48 can be utilized for categorization. J48 uses the idea of information entropy to construct decision trees from a set of labeled training data.

In addition to using cost-sensitive learning with a base classifier and analyzing breast cancer using data mining techniques, it also makes use of machine learning techniques such as Decision Tree (C4.5), Artificial Neural Networks, RK, and support vector machines to predict breast cancer. This is by J. Joshi et al.'s construction of the work to determine the effectiveness of pre-processing algorithms on datasets that are used to achieve more accurate results.

Dataset description, tool selection, pre-processing, resampling, attribute selection, classification algorithms, ensemble model, evaluation results, and transformation are the first steps in the methods for this work. In this step, we address an imbalanced dataset by using missing values and the attribute selection resamples method. The basic classifier, Meta classifier, ensemble model, evaluation findings, and model were the algorithms chosen for this study.

The data sets in this study are imbalanced; the major class is one with a greater number of instances, whereas a mini class or class has comparatively fewer examples. Sampling approaches employed in this investigation were used to solve imbalanced data problems with the distribution of a dataset. There are two kinds of sampling procedures: under-sampling and over-sampling. Under-sampling is a random under-sampling method that tries to balance the class distribution by randomly removing the majority class sample, and random oversampling also aids in this process. Balanced class distribution of classes through the replication of minatory class samples. Classification algorithms used in this study used four base classifiers: j48, SVM, rep tree, and neural network, as well as three Meta classifiers- bagging, boosting, and random subspace – were employed in this study classification procedure. To increase classification accuracy. I employed a model in this research for a classification composite model ensemble consisting of a mix of base classifiers and meta-classifiers. Ensemble methods can be used to increase overall accuracy by learning and combining a series base classifier model. Bagging, boosting, and random subspace are popular ensemble methods. The ensemble Model combined different types of classifiers to find the optimal classification performance from the combination sub-model. It contains two layers; the first layer consists of base classifiers, and the second layer is a Meta classifier, which receives the prediction of the base classifiers as an input and then generates the final prediction.

The experiment and result for this study consist of three experiments ensemble model combination Meta classifier with base classifier with resampling and attribute selection.

In the evaluation of classification results, Tagging, boosting, and random subspace algorithm is tried with many classification algorithms (J48, REP, Random Forest tree, SVM, and Neural Network), and the best performance of each ensemble classifier before and after resampling is obtained. Depending on the final result of the model that is constructed, the classification model efficiency is evaluated based on correct/incorrect instances, accuracy regarding correct and incorrect instances generated with a confusion matrix, Precision, Recall, f-measure, and time taken to build the model. The results of all base and Meta classifiers with attribute selection and resampling and all ensemble Classification experiments before using the resampling technique. This study aims to build a model to classify the imbalanced breast cancer data available in the IT departments in Sudanese hospitals. This will help us to predict cancer recurrence or non-recurrence events.

The research concluded that boosting the ensemble learning algorithm with a single misclassified J48 is the best model of classification that can be used in breast cancer data. In this research, the accuracy of classification techniques is evaluated based on a selected single classifier with a combination ensemble Meta algorithm with three popular Meta-learning algorithms (bagging, boosting, and random subspace). Also, the accuracy of classification techniques is evaluated based on the resampling method [5].

This paper studies different classification models that are used for both resampling techniques, machine learning classifier models used, SVM, Naïve baize R (logistic regression), Decision tree, Random Forest, Gradient boosting classifier, Bagging classifier with NB, Bagging classifier with DT, and Ad boosting.

As per G. Cohen,et el, the study suggested the resampling methods due to the difficulty of identifying the minority target. They applied a new resampling method by which equally oversampling of infrequent positives and under-sampling of the non-infected majority depending on synthetic circumstances created by class-specific sub-clustering, and they stated that their new resampling technique achieved better results than traditional random resampling.

In addition, the method introduced by G. Lemaitre, F. Nogueira, and C. K. Aridas uses k-means grouping to balance the imbalanced instances by decreasing the number of majority instances and also I. Mani and I. Zhang "applied an under-sampling method to remove information points from the majority of instances constructed on their spaces between each other. In the methodology for this study, two resampling techniques, which depend on changing the class distribution. Also, different classification models are used for both resampling techniques to compare between classifiers.

The main objective for sampling minatory class is to balance class spreading through the random repetition of minority target instances also this technique has two limitations First, it will raise the probability of over-fitting, as it creates the same reproductions of the minority class instances and second it, makes learning procedure more time overwhelming especially if the original dataset is now equally huge, but imbalanced; the same as our dataset.

This study used different classification models to predict with random oversampling techniques with different classifications and showed that random forest has the highest score among all evaluation metrics.

Under-sampling majority class second experiment used the under-sampling technique the best simple under-sampling

algorithm is random under-sampling. It is a non-heuristic algorithm that tries to balance target distributions over eliminating randomly from majority class instances. This operation may remove possibly valuable data that can be essential for classifier models, but it is useful when you have a lot of data.

In random under-sampling techniques, the score of the classifier models was poor when compared to the random over-sampling techniques, and Naive Bayes (NB) in the under-sampling method got a higher score compared to the other classifiers [9].

The imbalance is a problem that is very commonly found in disease-related datasets, such as the breast cancer dataset used in this study, where the class with a greater number of instances is known as the majority class, whereas the one with comparatively a smaller number of instances is known as the minority class.

To solve the problem of class imbalance, various sampling techniques have been introduced, which include under-sampling, oversampling, and a combination of both. Sampling strategies are introduced to overcome the class imbalance issue through the removal of some data from the majority class (undersampling) or the addition of some artificially synthesized or replicated data to the minority class (oversampling).

To build a good prediction model from the training set, the data must be well-balanced. However, the class labels of the target variable, cancer in the breast cancer dataset used in this study, are not balanced [6]

This study applied three different classification techniques on the areal breast cancer dataset, The first used specified classification techniques without any resampling techniques. Second, several resampling methods to get better performance. Classification techniques were used in this study: Decision tree, Random Forest, and Extreme Gradient Boosting.

Machine Learning (ML) or Data Mining (DM) algorithms are applied in the medical domain to assist with the decision-making process, for example, for the prediction of cancer risk. ML and DM algorithms can be classified into supervised or unsupervised learning, depending on the goal of the data mining task.

Classification is a supervised learning technique, and the goal of the classification model is to predict qualitative or categorical outputs that assume values in a finite set of classes (e.g., Yes/No or Benign-cancer/Malignant-cancer, etc.) without an explicit order. The primary objective of traditional classifiers is to get higher accuracy by reducing the overall classification error, however, the overall classification error is biased towards the majority class for imbalanced data problems.

Many approaches deal with this problem, such as cost function-based and sampling-based solutions. In this research, we focused on sampling-based approaches that can be classified into three major categories - random under-sampling, random over-sampling, and a hybrid of over-sampling and under-sampling.

In the methodology for this study, three different classifiers, namely Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost), were used to train the breast cancer data set of imbalanced data (original data as well as modified training data obtained by using different resampling methods.

Decision tree: A DT is a supervised learning approach that learns from class-labeled instances. It works very well with different types of data, and the results are easy to interpret.

Random Forest: RF is a powerful classification and regression tool that generates a forest of classification trees rather than a single classification tree. RF creates decision trees on randomly selected data samples, obtains the prediction from each tree, and selects the best solution using voting.

There are two stages in the RF algorithm. The first one is RF building, and the second stage is to predict the RF classifier created during the first stage.

Extreme Gradient Boosting (XGBoost): XGBoost is an implementation of gradient-boosted decision trees designed for speed and performance.

XGBoost is a scalable and accurate implementation of the gradient boosting machines, and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed.

To obtain a better classification performance, we used spec-iced classifiers to train the model using the original training data. We also used various types of resampling methods on the training data to train the model using specified classifiers with the modified training data.

There are steps in the model to handle imbalanced data; this step includes the classification model data and test for classification, and several resampling techniques were used on training data. The instances of the majority class were removed, or instances of the minority class were added. The classification model data was trained with the specified classifiers. First, we used the original training data without using any sampling methods and built models using the specified classifiers. Second, for the training, we used the modified training data obtained by applying the different sampling techniques. Each of these training data sets was used to train all three classifiers. All of the above models were saved for the prediction of the test data, and the last step generated the prediction of the test data.

In this study, three different classifier models on the original training data and different resampling methods on the training data were used to modify the training data accordingly. The modified training data sets were used for the training of the specified classifiers. Results were obtained from the models applied to the test data.

The overall performance of the DT classification models (built based on the modified training data) on test data for all the different training data sets [7].

In this study, the original dataset will be grouped according to the existing class so that it can produce majority and minority data. Data grouping is very necessary because this study will focus on minority data only, and then the data in the minority class is defined in the discrete attribute or attribute continuously, after defining the generation of synthetic data by the stages in each attribute and repetition of several majority data and minority data.

The RWO sampling method includes an oversampling method that works by forming or generating new data from minority classes. To form new data, the RWO sampling method is based on the average and standard deviation of the minority class data.

The classification algorithms used in this study, C4.5, Naive Bayes, and Neural Network, the accuracy, F-measure, and G-mean methods were used to measure the performance of the proposed method and the validation method using 10-fold cross-validation.

The classification algorithm used in this research is a Neural Network and Support Vector Machine (SVM) to measure the performance of the proposed method using AUC, f-measure, and G-mean, and a 5-fold cross-validation method is used for the validation method [8].

The purpose of this study is to explore the impact of various class balancing techniques on a model built on an imbalanced breast cancer dataset, the various methods in this study, sampling techniques – Random under-sampling (RUS), adaptive synthetic (ADAYSN), and oversampling techniques and hybrid techniques using RUS and ADAYSN also in this paper using various of ensemble model using voting and stacking are built for improvement in performance in predicting of breast cancer. The various steps required to be performed to correct diagnosis of breast cancer referred to pre-processing, resampling of the data applying random under-sampling techniques (RUS) and adaysn techniques, hybrid sampling techniques RUS+adaysn.

The classifier algorithms built on a dataset in this study are SVM using SGD, random forest (RF), logistic regression (LR), and XG Boost. Ensemble techniques in this study use a voting mechanism and stacking mechanism and, finally, evaluate the model.

In this study, different sampling techniques were applied to the training dataset:

Random undersampling (RUS) is the simplest and fastest undersampling technique. It randomly selects samples from the majority class and deletes them from the training set until a balanced distribution is achieved. All these undersampling techniques remove only a few samples from the majority class and are unable to achieve a balanced distribution.

Adaptive Synthetic Sampling (ADAYSN) generates minority class data synthetically, and it is a modification of the Synthetic minority oversampling technique (SMOTE). It generates more samples of those minority classes that are difficult to learn; that is, it generates more samples in the region where the density of the minority samples is low as compared to the region where the density is large.

Hybrid sampling RUS + ADAYSN combines both oversampling and under-sampling techniques using ADAYSN and RUS. These hybrid techniques improve the balance by removing instances of the majority class and adding instances of the minority class. Hybrid techniques combine the advantage of both under-sampling and oversampling techniques and are known to produce results. After applying the sampling techniques, the classifiers are built in this study using the LR, XGBoost, support vector machine with Stochastic Gradient Descent optimization (SVM-SGD), and random forest techniques. Cost-sensitive (weighted) classifiers are also built on unsampled data using these four algorithms. The weighted classifiers take care of class imbalance issues by penalizing the misclassification of the minority class.

SVM is a powerful classification technique that can be used to build a linear classifier and a non-linear classifier with the help of kernels. Stochastic Gradient Descent (SGD) is one of the optimization techniques that, when used with the SVM model, not only optimizes the accuracy of the model but also reduces execution time.

Random Forest is another supervised learning technique, and also, random forest reduces the bias in the final model as each tree is constructed on different samples and different features.

Extreme Gradient Boosting (XGBoost) XGBoost is a powerful ensemble of a decision tree used for classification in boosting. XGBoost is the most popular algorithm due to its scalability, speed, and capability to handle sparse data. A class-weighted XGBoost handles class imbalance issues by assigning more weight to the minority class. This study evaluates the model on the XGBoost with all the data sampling techniques, as well as the weighted approach. [9]

This paper presents a study of the different techniques that are used to handle the imbalanced dataset and finally proposes a novel oversampling technique to tackle the binary classification of the imbalanced dataset problem.

In this paper, we argue that oversampling techniques can yield better results in handling imbalanced dataset problems if the majority of data is considered during the oversampling process. The proposed technique tackles the imbalanced dataset problem by using a sample of the majority data to create a better new sample of minority data. The evaluation results show that such an oversampling technique outperforms the standard oversampling algorithm.

There are three main approaches to tackling imbalanced data problems. The data level method, also called the external method, works to adapt the number of data instances to balance the distribution. On the other hand, the algorithm-level method (called the internal method) adapts the traditional algorithms of learning to minimize the bias, increase accuracy, and get the benefit of mining data that have skewed distributions. Hybrid methods combine both data and algorithm-level methods.

With data-level methods, the goal is to modify the dataset to make it more suitable to apply a traditional learning algorithm. Three sub-approaches to modify the dataset are under-sampling, feature selection, and oversampling. Undersampling is removing samples from the majority class, whereas oversampling is generating new objects for the minority class. Feature selection means the algorithms that output a subgroup of the input feature set that is more relevant and helps a classifier enhance its performance. The oversampling adds synthetic samples to the minority class to balance the distribution of the classes. The simplest method of oversampling is the replication of instances of the minority class.

The algorithm-level method modifies traditional learning algorithms to lessen the bias found toward the majority class. To achieve this, a good understanding of the learning algorithm is needed, as well as a clear analysis of reasons for its failure in learning from imbalanced datasets.

In this cost-sensitive approach, the traditional learning algorithm is adapted to include varying penalties for each class of samples. There are a lot of methods for ensemble algorithms used in algorithm-level methods, such as bagging, boosting, random forest, and rotation forest. Till now, several approaches have been developed, and improvements to traditional methods have been designed to solve the issue of imbalanced distributions. Bagging is a machine learning approach that is used to improve accuracy while reducing variance in classifying samples. Boosting means that poor classifications can be combined to create a more correct decision. That is to say, boosting means several algorithms that use weights to make weak learners more accurate.

In hybrid methods, preprocessing is done to the data samples with imbalanced distribution. This is done by using over or under-sampling at first, and then using a cost-sensitive approach. Classification algorithms used in hybrid methods: sampling-based- approach with cost-sensitive learning, in these methods, preprocessing is done to the data samples with imbalanced distribution. This is done by using over or under-sampling at first, and then using a cost-sensitive approach.

Random forest and random subspace methods are the second classifications used in hybrid methods. Random trees are still growing, mostly because of their flexibility and good performance. The random forest is treated as a simple-to-tune technique, unlike other techniques (e.g., GBM), which require careful tuning. Random forests utilize a large number of integrated decision trees.

Extremely randomized trees, the third classification used in the hybrid method, use randomness in the training stage to produce different sets. In addition to the random subgroup of attributes that chooses the most distinctive feature, defining attributes are randomized when extremely random trees are applied.

Finally, ensemble methods and deep neural networks, in many fields such as speech recognition and object detection and many other fields, DNN has improved dramatically in the last few years. As per Batista et al. (2004), research presented a respectable study of sampling methods. Several strategies of over and under-sampling and dynamic/hybrid processes have been tested and examined carefully on thirteen datasets. While most of them had improved performance, in all experimental datasets, there has been no method overwhelming others.

The paper explored the nature of the imbalanced data and its current real-life application. At the end of this paper, the proposed techniques for handling the imbalanced data problem are presented. In the future, the aim is to further apply it to categorical data sets, and the direction is to apply the approach to multiclass datasets [10].

In this paper, an oversampling method, the aim is to augment the original dataset with synthetically created observations of the minority classes. Introduced a new oversampling technique based on variation autoencoders. The experiments show that the new method is superior in augmenting datasets for downstream classification tasks when compared to traditional oversampling methods.

This study used SMOTE techniques, which forces the decision region of the minority class to be more general. Altering the class distributions of a dataset does have downsides, however; under-sampling the majority class may lead to discarding useful data, and oversampling the minority class can lead to over fitting. Also used is ADAYSN; the algorithm uses a density distribution to determine the number of additional synthetic examples needed to be generated for each minority sample. This is in contrast to SMOTE, where an equal amount of synthetic data is generated for each minority data sample. In contrast to SMOTE and ADAYSN, cost-sensitive learning techniques do not modify the imbalanced data distribution directly. These techniques can also consider learning when error costs are unequal. When misclassification costs are known, they can be incorporated directly into the cost function. In this study, generative approaches have shown promise by outperforming traditional sampling or cost-sensitive techniques. The author generated synthetic data points from the minority class by first learning the probability distribution of the minority class and subsequently adding to a resampled

The process was set until the desired proportion between minority and majority classes was reached.

In this paper, we similarly focus on generative methods for oversampling and introduce a new generative modeling approach using Variation Autoencoders (VAE) to oversample the minority class in an imbalanced dataset, with a focus on binary target variables. However, the approach can be used easily in multi-class situations.

To conclude, this paper introduced a new generative approach for oversampling based on variation inference. In particular, used a two-stage latent structure VAE to learn a sampling distribution of the original dataset. To learn the minority class distribution, the target responses augment z encodings to learn the second encodings z our experimental results illustrated the superior performance of the new oversampling method versus SMOTE as well as ADASYN, and indeed demonstrate the promise of this new method for dealing with imbalanced datasets [11].

This paper proposes a novel two-stage resampling methodology in which we initially use the oversampling techniques in the image space to leverage a large amount of data for the training of a convolutional neural network.

In this study a novel approach for handling data imbalance in the image recognition task, in which we apply data resampling in two stages: first of all, we oversample the data directly in the image space and use it for the initial training of the model, and afterward we under-sample the data in the high-level feature space produced, based on the input images, by the previously trained network, to fine-tune its last layers.

The first work related to the subject of the impact of data on neural networks can be traced to Masko and Hensman who used random oversampling (ROS) who used random oversampling (ROS) before training the convolutional neural network, and Lee et al. who used random under-sampling (RUS) in combination with transfer learning, Pouyanfar et al. who introduced a dynamic sampling approach, and Buda et al.who consider the impact of RUS, ROS, and two-phase training, an approach in which the network is first trained on the balanced dataset, and afterward finetuned on the original data. The method advocates for the use of resampling in both the image space and the feature space. Koziarski et al. considered the impact of data imbalance on the performance

of a convolutional neural network in the breast cancer recognition task, as well as the possibility of applying different data resampling techniques directly in the image space. Additionally, successfully applied the Radial-Based sampling algorithm in the high-level feature space, achieving an improvement in the performance. Both the over- and the under-sampling techniques have their limitations that have to be addressed to achieve a satisfactory performance, in particular in the image recognition setting. Traditional oversampling algorithms producing synthetic observations, such as SMOTE and its derivatives, were not designed to be used on the image data. This study proposes a conceptually simple strategy of two-stage data resampling intertwined with a traditional convolutional neural network training procedure. The motivation behind the approach is to, first of all, leverage a high amount of data, further enhanced by applying to oversample in the initial training of the convolutional network, and afterward fine-tune the fully connected head of the network on a smaller amount of undersampled data, uncontaminated by the synthetic observations. It is important to note that even though we propose using oversampling in the first stage of the algorithm and undersampling in the second stage, the exact choice of both resampling strategies can be treated as a parameter of the method. Another theoretically viable strategy is applying one of the variants of SMOTE in the second resampling stage directly on the high-level features extracted from a previously trained convolutional network. In this paper, we considered the case of an originally balanced benchmark dataset with artificially introduced data imbalance. Specifically, we used a colorectal cancer histology dataset published by Kather et al. It consists of a total of 5,000 histological images of human colorectal cancer divided.

Into eight different types of tissue. The dataset included textures extracted at different scales, from individual cells to larger structures. Each image had a dimensionality of 150_150 pixels. In addition to the baseline setting in which no data resampling was applied, we considered three popular data-level approaches for handling data imbalance: random under-sampling (RUS), random oversampling (ROS), and SMOTE. The methods were applied in one of three ways: directly in the image space (IS), in which case the data was vectorized before resampling and reshaped to the original format afterward; in the feature space (FS). In conclusion, consider the imbalanced image recognition problem with an application to the multi-class texture analysis in colorectal cancer histology. Discussed the shortcomings of the existing data-level strategies of dealing with data imbalance in the context of image recognition and proposed a novel approach, two-stage data resampling, to mitigate the described deficiencies of over- and under-sampling. Finally, in the conducted experimental analysis, we empirically confirmed the usefulness of the proposed approach. We evaluated a combination of oversampling the data in the image space and later undersampling it in the high-level feature space, and we were able to achieve additional improvement in performance [12].

This research presents a framework for software defect prediction by using feature selection and ensemble learning techniques. Software Defect Prediction (SDP) is an effective way to resolve this issue, which ensures the high quality of software with a limited number of resources. Machine learning techniques have been widely used for software defect prediction for the last two decades. These techniques are categorized as supervised, unsupervised, and hybrid. In supervised learning, the classes are known in advance. These learning techniques need the pre-classified data (training data) for training, during which classification rules are made, and then these rules are used to classify the unseen data (test data). In unsupervised learning, classes are not known,

These techniques use particular algorithms to explore and identify the structure of data. The hybrid learning or semi-supervised learning approach integrates both supervised and unsupervised techniques. The objective of this research is to contribute to improving the prediction of defect-prone software modules. For this purpose, a framework is presented for software defect prediction by using feature selection and ensemble learning techniques. The preprocessing stage of the proposed framework consists of three activities: Normalization, Feature Selection, and Class Balancing. All of these activities aim to improve the structure of data so that higher results can be achieved from the classification process. The classification stage uses the 'Stacking' technique to implement the ensemble learning. The classification used in this research includes: "Naïve Bayes (NB), Multi-Layer Perceptron (MLP), Radial Basis Function (RBF), Support Vector Machine (SVM), K Nearest Neighbor (KNN), k Star (K*), One Rule (One R), PART, Decision Tree (DT), and Random Forest (RF)". It is observed that the proposed framework performed well compared to all of the base classifiers.

Many researchers have used machine learning techniques to solve binary classification problems such as Sentiment Analysis, Rainfall Prediction, Network Intrusion Detection, and Software Defect Prediction. The classification techniques include: "Naïve Bayes (NB), and Multi-Layer Perceptron (MLP). Radial Basis Function (RBF), Support Vector Machine (SVM), K nearest Neighbor (KNN), k Star (K*), One Rule (One R), PART, Decision Tree (DT), and Random Forest (RF)". The performance is analyzed by using Precision, Recall, f-measure, Accuracy, MCC, and ROC Area. Researchers proposed a feature selection-based ensemble classification framework. The framework is implemented in two dimensions, one with feature selection and the second without feature selection. The performance is analyzed by using Precision, Recall, F-measure, Accuracy, MCC, and ROC. The researchers used six classification techniques to predict the software defects. The classification techniques include Discriminant Analysis, Principal Component Analysis (PCA), Logistic Regression (LR), Logical Classification, Holographic Networks, and Layered Neural Networks. The back-propagation technique was used in ANN for training. Performance was evaluated by various measures, including Verification Cost, Predictive Validity, Achieved Quality, and Misclassification Rate. The researchers used SVM to predict the software bugs in datasets. The researchers discussed the significance of metric selection for software bug prediction. They discussed that some metrics are more important than others when predicting software defects. They used the ANN model to identify the significant metrics. The selected metrics were then used by the

researchers to predict the software defects through another ANN model. The performance of the proposed method was compared with Gaussian kernel SVM, and the dataset was used for the experiment. The results reflected that SVM performed better than ANN in binary defect classification. The researchers presented an integrated method that consists of a Hybrid Genetic algorithm and a Deep Neural Network. The Hybrid Genetic algorithm selects the optimum features, and the Deep Neural Network performs the prediction by classifying the modules as defective and non-defective. The experiments were performed on various datasets. The results reflected that the proposed approach showed higher performance as compared to other techniques.

This paper presents a feature selection-based ensemble classification framework. The framework consists of datasets and preprocessing, Preprocessing proposes three activities: normalization, feature selection, and class balancing. The process of normalization aims to bring the values of the complete dataset into the range of 0 to 1 for effective classification results. For feature selection, the wrapper approach with Artificial Neural Network (MLP) is used as a feature subset evaluator, and full datasets are used for training. Six search methods are used, including Best First (BF), Greedy Stepwise (GS), Genetic Algorithm/Search (GA), Particle Swarm Optimization (PSO), Rank Search (RS), and Linear Forward Selection (LFS). For each of the used datasets. For classification ensemble learning techniques, stacking along with Meta classifier. Moreover, the base classifiers include Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K Nearest Neighbor (kNN), and Bayes Net (BN).

The proposed framework's results are evaluated using various measures such as F-measure, Accuracy, MCC, and ROC. These measures are calculated by using the parameters of the confusion matrix. The proposed framework is implemented on Datasets by using six widely used search methods. The search methods include Best First (BF), Greedy Stepwise (GS), Genetic Algorithm/Search (GA), Particle Swarm Optimization (PSO), Rank Search (RS), and Linear Forward Selection (LFS). In this study, the results of the proposed framework (including all search methods) in each dataset are compared with the results. The used machine learning techniques include Naïve Bayes (NB) and Multi-Layer Perceptron (MLP). Radial Basis Function (RBF), Support Vector Machine (SVM), K Nearest Neighbor (KNN), kStar (K*), One Rule (OneR), PART, Decision Tree (DT), and Random Forest (RF). Presents the highest scores of base classifiers on the same datasets that are used in this research.

In conclusion for this research, this study presented a feature selection-based ensemble classification framework for effective software defect prediction. This paper presented a feature selection-based ensemble classification framework for effective software defect prediction. The proposed framework is implemented on datasets, and the performance is analyzed in terms of F-measure, Accuracy, MCC, and ROC. Performance evaluation is performed in two dimensions: first, the scores of all search methods within the framework are compared with each other, and second, the results of the proposed framework with all search methods are compared with the results of 10 well-known machine learning classifiers. The results reflected the higher performance of the proposed framework as compared to all classifiers. [13].

Data imbalance in Machine Learning refers to an unequal distribution of classes within a dataset.

In this study, there are experiments with two resampling adopted techniques: oversampling and undersampling. Several researchers consider it a challenging issue that needs more attention to resolve the imbalance problem. One of the common approaches was to use resampling techniques to make the dataset balanced. Resampling techniques can be applied either by under-sampling or oversampling the dataset. Undersampling is the process of decreasing the amount of majority target instances or samples. Some common under-sampling methods contain tomeks' links, cluster centroids, and other methods. Oversampling can be performed by increasing the amount of minority class instances or samples by producing new instances or repeating some instances. An example of the oversampling method is Borderline-SMOTE. Many related works shown in this study suggested the resampling methods due to the difficulty of identifying the minority target. They applied a new resampling method by which equal oversampling of infrequent positives and under-sampling of the non-infected majority, depending on synthetic circumstances created by class-specific sub-clustering. They stated that their new resampling technique achieved better results than traditional random resampling. According to E. Duman, Y. Ekinci, and A. Tanriverdi, the study applied three dissimilar methods to an advertising dataset. Logistic regression, Chi-squared automatic interaction detection, and neural network.

The performance of the three methods was created by means of accuracy, AUC, and precision. They compared several different imbalanced datasets produced from the real dataset. They stated that precision is a good measure for an imbalanced dataset. Also, I. Mani and I. Zhang's study applied an under-sampling method to remove information points from the majority of instances constructed on their spaces between each other. The data set is used to tackle and review the imbalanced data problem. In general, the whole dataset in this study contains 202 features and 200,000 entries. This study used two resampling techniques, which depend on changing the class distribution. Also, we studied different classification models. The experiment for this study used the oversampling techniques. A non-heuristic algorithm is known as random oversampling. Its main objective is to balance class spreading through the random repetition of minority target instances and also use different classifier models to predict with random oversampling techniques. Classifier model used with oversampling techniques: SVM, logistic regression, decision tree, random forest, gradient boosting, bagging with naïve Bayes, bagging with decision tree, Ada boosting. After showing all the evaluation metrics for all the classifiers mentioned in this study, they can see that random forest has the highest score among all of the evaluation metrics and performs better than the other classifiers.

The second experiment of this study used the under-sampling techniques. The best simple under-sampling algorithm is random under-sampling. It is a non-heuristic algorithm that tries to balance target distributions over eliminating randomly from majority class instances. This operation may remove

possibly valuable data that can be essential for classifier models, but it is useful when you have a lot of data. Sampling techniques used the same classifier model of oversampling to predict sampling. When the random under-sampling techniques were used, the score of classifier models was poor when compared to random oversampling techniques, and notice that some classifiers had the same scores as in the oversampling method or got worse in some other classifiers. Naive Bayes (NB) in the under-sampling method got a higher score compared to the other classifiers.

In conclusion, this study presented two techniques to handle the problem of class imbalance and applied them to different machine learning classification models. They have used the dataset provided by the competition from the Kaggle website," Santander Customer Transaction Prediction". They have used this data to tackle and review the imbalanced data problem. We tried the oversampling technique for the dataset and measured the classifiers with different evaluation metrics, as well as the other technique, under-sampling. We noticed how oversampling performs better than under-sampling for different classifiers and gets higher scores in different evaluation metrics.

In my opinion, for future work, I plan to apply different deep learning techniques with both resampling techniques to compare them. [14].

The typical solution for an imbalance dataset includes data level (under-sampling or over-sampling) or algorithmic level (cost-sensitive learning approach) Synthetic Minority Oversampling Technique (SMOTE) has been acknowledged as one of the most effective data level solutions and also the ensemble learning techniques have recently emerged as effective; but can yield best results when integrated with data level solutions.

In this work, a Boosting-based oversampling technique is introduced with a customized oversampling rate within an ensemble framework through cost-sensitive error formulation.

The oversampling rate is tailored by using the Local Covariance Matrix (LCM), while the AdaBoost ensemble model with C4.5 weak learner is implemented as the ensemble framework. In this paper, the solutions proposed to solve the class imbalance problem can be categorized into two major groups: Data level solution formally known as Data Sampling which modifies data distribution and yields a revised set with balanced data distribution, and Algorithmic level solution which amends the classifier to improve the classifier accuracy and Data sampling can be either under-sampling (elimination of majority class instances) or oversampling (addition of duplicate minority class instances).

The motivation for performing is to balance the data distribution by replicating the minority class instances to a specific extent. Among the proposed oversampling techniques, the Synthetic Minority Oversampling Technique (SMOTE) is the most practiced one. SMOTE was proposed by Chawla et al. which facilitates to generate of synthetic instances along the minority class instances, by working on the feature space rather than the data space But, typically in practice the uniform oversampling rate, N of SMOTE may lead to escalated redundancy among the minority class instances when the degree of noise is high in the data space.

The data-level solutions provide to rebalance the data distribution, either by replicating the minority instances or by eliminating the majority instances. The first approach is termed oversampling, while the latter one is defined as under-sampling.

As per research, Chawla et al. defined the SMOTE algorithm to replicate minority instances that are lying near each other, and Han et al. proposed the Borderline-SMOTE method to generate synthetic instances along the borderline instances with high misclassification cost.

Barua et al.. defined MWMOTE to generate synthetic minority class instances from a set of weighted informative minority instances.

In discussing algorithmic level solutions, cost-sensitive technique defines a cost function against misclassification to reduce classification error, as Nguyen et al. define a phase cost-sensitive learning approach. Castro et al. defined a novel cost-sensitive algorithm to improve the classification capability of a multi-layer perceptron model for imbalanced data. Ensemble learning solution, the integration of ensemble learning techniques with data pre-processing methods is a popular practice followed by the researchers, Chawla et al. defined SMOTE Boost which combines SMOTE and Boosting to generate synthetic instances from the minority class.

AdaBoost ensemble model with decision tree C4.5 as the weak learner is implemented for iterative learning of the balanced training set. A misclassification ratio-based cost function is defined, which emphasizes learning the misclassified minority instances in each iteration to enforce more weightage to them in consecutive iterations. The entire process is repeated for a fixed number of iterations until it meets the stopping condition or the desired accuracy [15].

Table 1: Provides Resampling Techniques

Author(s)	Aim of Study	Imbalance Data	Applied on	Methods
Hana Babiker Nassar [5]	Presented Imbalance data for Breast cancer with different classifier models are built. Classifier algorithms (ANN, Rep tree, J48, SVM) are used as meta classifiers and, finally, ensemble models.	Yes	trained experimental algorithms	Classification algorithms (ANN, Rep tree, J48, SVM) meta algorithms (Bagging, Boosting, Random subspace).
Keerthana Rajendran, Manoj Jayabalan, Vinesh Thiruchelvam [6]	Proposed in this study, the hybrid balancing method achieved greater performance across the proposed classifiers. The breast cancer predictive model developed using the Bayesian Network was rarely explored in previous breast cancer studies, and in this study, this classifier proved to achieve the highest accuracy when compared to other works done using the BCSC dataset.	Yes	This study attempts to apply three different class balancing techniques: oversampling (synthetic minority oversampling technique), under-sampling (spread subsample), and the hybrid method (SMOTE and spread subsample) and the classification algorithms, including study Naïve babies', Bayesian Network, Random Forest, and Decision tree (C4.5). This study proved that the hybrid method with the Bayesian Network achieved the greatest in predicting breast cancer.	Oversampling (SMOTE) and under-sampling, the hybrid method also classification algorithms: Naïve Bayesian, random forest, and decision tree.
Ayat Mahmoud, Farid Ali, Ayman El-Kilany, Sherif Mazen [10]	Presents a study of the different techniques that are used to handle the imbalanced Dataset, and finally proposes a novel oversampling technique to tackle the binary classification of the imbalanced dataset problem. The proposed techniques tackle the imbalanced dataset problem using a sample of the majority to create a better new sample of minority data.	Yes	This study explored the nature of the imbalanced data and its current real-life application. At the last. This study introduced the proposed techniques for handling the imbalanced data problem. In the future, the aim is to further apply it to categorical data sets, and the direction is to apply the approach to multiclass data sets.	Three main approaches to tackle imbalanced data, Data level method, algorithm level method, and hybrid method combined both data and algorithm level method.
Val Andrei Fajardo, David Findlay, Roshanak Houmanfar, Charu Jaiswal [11]	This study used SMOTE techniques, which forces the decision region of the minority class to be more general. Altering the class distributions of a dataset does have downsides, however; under-sampling the majority class may lead to discarding useful data, and oversampling the minority class can lead to over fitting.	Yes	This study introduces a new generative modeling approach using Variational Autoencoders (VAE) to oversample. The minority class in an imbalanced dataset, with a focus on binary target variables.	SMOTE techniques and ADAYSN the algorithm uses a density distribution to determine the number of additional synthetic examples needed to be generated for each minority sample.
Michal Koziarski [12]	Proposed a novel two-stage resampling methodology, in which we initially use the oversampling techniques in the image space to leverage a large amount of data for the training of a convolutional neural network.	Yes	An imbalanced image recognition problem with an application to the multi-class texture analysis in colorectal cancer histology proposed a novel approach, two-stage data resampling, to mitigate the described deficiencies of over- and under-sampling.	Resampling methodology, random oversampling,
Roweida Mohammed, Jumanah Rawashdeh and Malak Abdullah [14]	In this study, there are experiments with two resampling adopted techniques: oversampling and undersampling. Several researchers consider it a challenging issue that needs more attention to resolve the imbalance problem.	Yes	Presented two techniques to handle the problem of class imbalance and applied them to different machine learning classification models.	Used resampling techniques: oversampling and undersampling.
Debashree Devi, Saroj K.	In this work, a Boosting-based	Yes	In this paper, the solutions	Over-sampling techniques

Biswas, Biswajit Purkayastha [15]	oversampling technique is introduced with a customized oversampling rate within an ensemble framework through cost-sensitive error formulation.		proposed to solve the class imbalance problem can be categorized into two major groups: Data level solution formally known as Data Sampling which modifies data distribution and yields a revised set with balanced data distribution, and Algorithmic level solution which amends the classifier to improve the classifier accuracy and Data sampling can be either under-sampling (elimination of majority class instances) or oversampling (addition of duplicate minority class instances).	(SMOTE) and boosting
Md Faisal Kabir, Simone A. Ludwig [7]	Proposed sampling-based approaches can be classified into three major categories - random under-sampling, random over-sampling, and a hybrid of over-sampling and under-sampling.	Yes	Applied three different classification techniques on the areal breast cancer dataset, first using specified classification techniques without any resampling techniques. Second, several resampling methods to get better performance.	Classification techniques were used in this study: Decision tree, Random Forest, and Extreme Gradient Boosting.
Faseeha Matloob, Shabib Aftab, Ahmed Iqbal. [13]	Presents a framework for software defect prediction by using feature selection and ensemble learning techniques. Software Defect Prediction (SDP) is an effective way to resolve this issue, which ensures the high quality of software with a limited number of resources.	Yes	These techniques are categorized as supervised, unsupervised, and hybrid.	The classification used in this research includes: "Naïve Bayes (NB), Multi-Layer Perceptron (MLP), Radial Basis Function (RBF), Support Vector Machine (SVM), K Nearest Neighbor (KNN), k Star (K*), One Rule (One R), PART, Decision Tree (DT), and Random Forest (RF)". It is observed that the proposed framework performed well compared to all of the base classifiers.
Ruchita Gupta Rupal Bhargava Manoj Jayabalan [9]	The purpose of this study is to explore the impact of various class balancing techniques on a model built on an imbalanced breast cancer dataset, the various methods in this study.	Yes	Applied to improve the classification of the minority class. Classifiers were built on unstamped data using the cost-sensitive version of these algorithms. Ensemble models Were also explored.	Sampling techniques – Random under-sampling (RUS), adaptive synthetic (ADAYSN) oversampling techniques, and hybrid techniques using RUS and ADAYSN and also using several ensemble model voting and stacking for improvement.

3.METHODS

We will employ methodology as a framework in this study, which consists of steps starting with a literature review and obtained dataset and preparing it well after that applying pre-processing of the dataset. We received the dataset, data transformation, and tool selection using pre-processing (missing value), cross-validation, attribute selection, resampling, base classification algorithms, Meta-learning algorithms, and resampling methods. They evaluate their performance by building a classification model and feature selection techniques.

A. Data Description

A data set is a collection of information gathered for a certain objective; data can be collected, for instance, by surveys, interviews, observations, extraction, and so forth.

The data set used in this study was taken from files and records of Khartoum State Hospital, and recorded 1144 patients made up the entire sample.

The data include Age, Tumour Size T), Node – Caps, deg – malign (Metastasis), L, and left breast\R, right breast, Irradiate, and Class are among the information.

There are 336 recurrences and 808 no recurrences.

In this imbalanced data set, the class with the most occurrences in an imbalanced data set is referred to as a major class, while the class with comparatively fewer instances is referred to as a minor class.

Table 2: illustrates the Data Description

Item	Describe	Attribute Type
(T, Tumour)	The patient's tumour in the breast	Numeric
Age	Patient's Age	Numeric
(N, Nodes)	Node is present or not in the cap of the breast	Nominal
(M, Metastasis)	Tumors spread to other parts of the body	Nominal
Deg-Malig	Stages of breast cancer	Numeric
L/R	Breast, left and right	Nominal
Irradiate	Present or not	Nominal
Class	No recurrence-events, recurrence-events (reduce the risk of breast cancer)	Nominal

B. A Selected Tools

The software framework of this work has been developed with the WEKA tool. WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, classification, regression, clustering, and association rules; it also includes visualization tools. The new machine learning schemes can also be developed with this package. WEKA is open-source software issued under General Public License [16].

C. Pre-processing

One crucial phase in the data mining process. Data pre-processing describes any kind of processing that is referred to as data pre-processing. In this study, data was entered into an Excel sheet and saved by CSV, missing values were handled, and all of the data were numeric and nominal.

D. Resampling

Sampling procedures are sometimes referred to solve imbalanced data problems with the distribution of a dataset, Sampling techniques involve artificially re-sampling the data set, also known as the data pre-processing method.

1) Under - Sampling

The most essential method in under-sampling is a random under-sampling method that tries to balance the class distribution by randomly removing the majority class sample. Figure 1 shows the random. Under-sampling method. The problem with this method is the loss of valuable information.

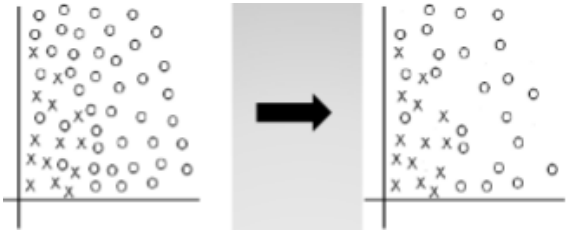

Figure 1: Randomly removes the majority of samples.

2) Over – Sampling

Random oversampling methods also help achieve balanced class distribution by replicating minatory class samples. Figure 2 shows random over-sampling.

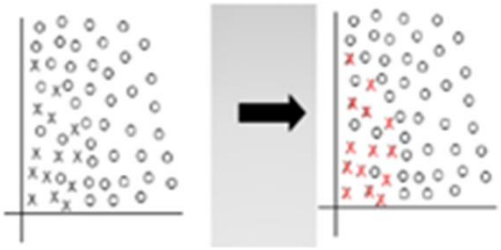

Figure 2: Randomly Removes Minority Samples.

4.RESULT AND DISCUSSION

A. Ensemble learning

To increase classification accuracy, I employed an Ensemble model in this research, the classification composite model comprising a combination of classifiers base and meta-classifiers. Ensemble methods can be used to increase overall accuracy by learning and combining a series base classifier model. Bagging, boosting, and random subspace are popular ensemble methods. The ensemble Model combined different types of classifiers

to find the optimal classification performance from the combined sub-model. It contains two layers; the first layer is made up of base classifiers, while the second layer is a Meta classifier that takes basic classifiers input and uses them to create the final prediction.

B. *The first Experiment and result*

This experiment consists of a base classifier (SVM, ANN, REP, and J48) without attribute selection; we used all base classifier algorithms, and we found the best result: SVM.

Figure 3 shows the result of the SVM to classify data to obtain the accuracy of the result without attribute selection.

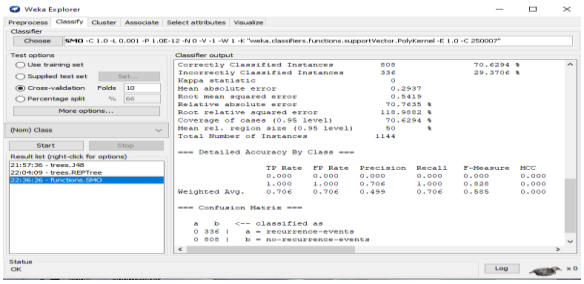

Figure 3: Result of SVM without attribute selection.

C. *Result of SVM with attribute selection*

The Experiment consists of a Base classifier (SVM, REP TREE, ANN, and J48) with Attribute Selection. We used All Base classifier Algorithms, and we found the best result was SVM.

Figure 4 shows the result of SVM to classify data to obtain the accuracy of the result with Attribute selection using Gain ratio with Ranker.

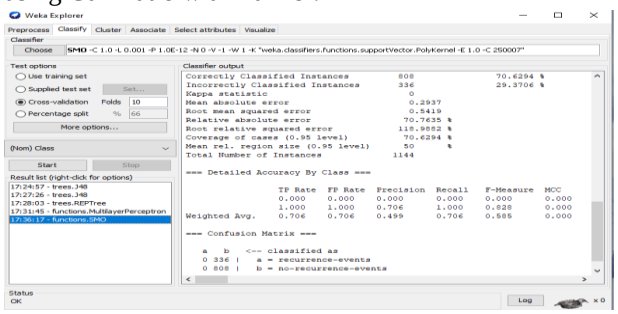

Figure 4: Result of SVM with attribute selection.

D. *Result of the rep tree with resampling*

The Experiment consists of a Base classifier (SVM, REP TREE, ANN, and J48) with a Resampling Method. We used All Base classifier Algorithms, and we found the best result was REP TREE.

Figure 5 shows the result of the Decision tree Rep Tree in WEKA to classify data of patients to obtain accuracy of results with resampling method.

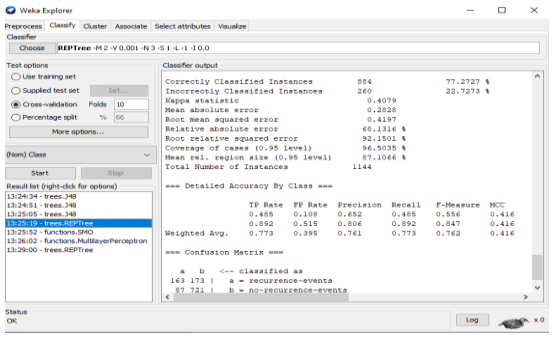

Figure 5: Result of rep tree with resampling.

E. *The second experiment boosting without attribute selection*

Consists of meta meta-classifier (Bagging, Boosting, Random subspace) without Attribute selection. We used All Meta Classifier, and the best result was Boosting and the rest of the results.

The experiment was conducted using the Boosting Meta-learning Algorithm in WEKA to classify data of patients to obtain accuracy of results without attribute selection.

Figure 6 shows the result of boosting without attribute selection.

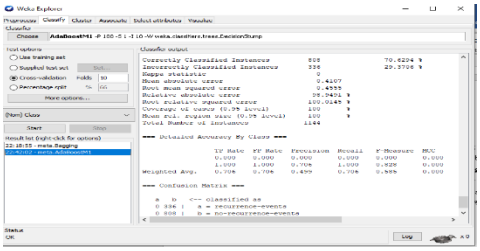

Figure 6: Result of boosting without attribute selection.

F. *Result of boosting with attribute selection*

The Experiment consists of a meta-classifier (Bagging, Boosting, Random subspace) with Attribute selection. We used the All Meta Classifier, and we found the best result was Boosting.

Figure 7 shows the result of the Boosting Meta-learning Algorithm in WEKA to classify data of patients to obtain accuracy of results with attribute selection.

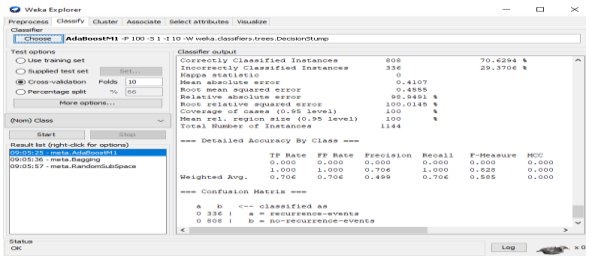

Figure 7: Result of boosting with attribute selection

G. Result of bagging with resampling

The Experiment consists of a meta-classifier (Bagging, Boosting, Random subspace) with a Resampling Method. We used the All Meta Classifier, and we found the best result was bagging the rest of the results.

Figure 8 shows the result of the Bagging Meta learning Algorithm in WEKA to classify data of patients to obtain accuracy of results with the Resampling Method.

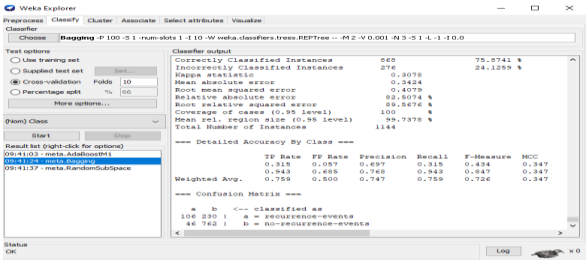

Figure 8: Result of bagging with resampling.

H. The third Experiment resulted in ensemble model boosting with j48

Meta-learning algorithms in combination with a base classifier. We used the All Base classifier (J48, SVM, ANN, REP TREE) Combination with Meta Classifier (Bagging, Boosting, Random subspace) we found the best result was Boosting with J48.

Figure 9 shows the result of boosting the ensemble learning algorithm with the J48 tree classification algorithm after the Resampling technique and Attribute Selection.

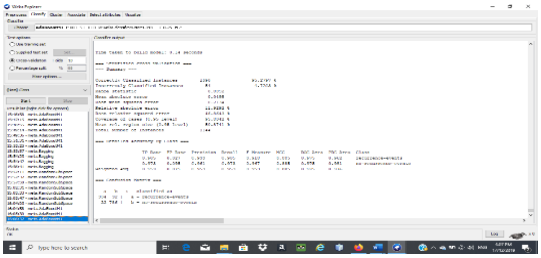

Figure 9: Result of ensemble boosting with j48

I. 3.5.1 The result of ensemble bagging with resampling combined j48

The Experiment Meta earning Algorithms combination with base classifier. We used All Base classifiers (J48, SVM, ANN, REP TREE) in Combination with the Meta Classifier (Bagging, Boosting, Random subspace). We found the best result was bagging combined with J48.

Figure 10 shows the result of Bagging with Attribute Selection and resampling Combined (J48).

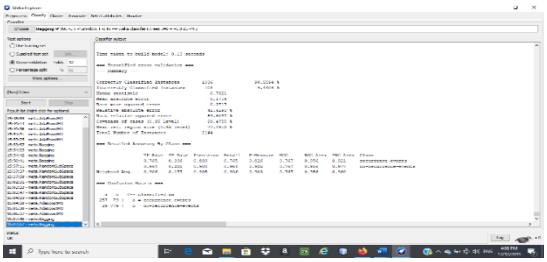

Figure 10: Result of ensemble bagging combined j48.

J. 3.5.2 The result of random subspace with resampling combined j48

The Experiment Meta earning Algorithms combination with base classifier. We used All Base classifiers (J48, SVM, ANN, REP TREE) in Combination with the Meta Classifier (Bagging, Boosting, Random subspace).. We found the best result was Random, combined with J48.

Figure 11 shows the result of Random subspace with Attribute Selection and resampling Combined (J48).

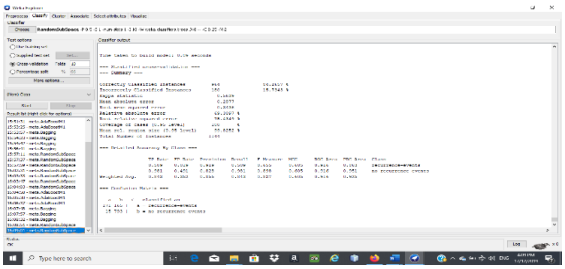

Figure 11: Result of random subspace combined j48

K. 3.5.3 Analysis of Results for Base Classifier

Comparison of base classifier algorithms between base classifiers algorithms, Meta classifiers algorithms, and ensemble learning combination As shown in Table 3 , 4 , 5.

Table 3: Comparison of base classifier

Method	Accuracy %	Precision	Recall	F-measure	Computational time
SVM with attribute selection	70.6294%	0.499	0.706	0.585	0.14 second
SVM without attribute selection	70.6294%	0.499	0.706	0.585	0.15 second
Rep tree with resampling	77.2727%	0.761	0.773	0.762	0.01 second

Table 4: Comparison of Meta classifier

Method	Accuracy %	Precision	Recall	F-measure	Computational time
boosting without attribute selection	70.6294	0.499	0.706	0.585	0.47 second
Boosting with attribute selection	70.6294%	0.499	0.706	0.585	0.17 second
Bagging with resampling	758741%	0.747	0.957	0.726	0.05 second

Table 5: Comparison of Ensemble classification model

Method	Accuracy %	Precision	Recall	F-measure	Computational time
Ada boost+j48	95.2797%	0.953	0.953	0.953	0.14 second
Bagging + j48	90.559%	0.905	0.906	0.903	0.13 second
Random subspace + j48	84.2657%	0.855	0.843	0.827	0.09 econd

5.CONCLUSION

Building a classification model to categorize the imbalanced data about breast cancer, which is available in the IT departments of Sudanese hospitals, is the goal of this study. This will assist us in forecasting whether cancer will recur or not. The study concluded that the optimum classification model for breast cancer data to improve the ensemble learning algorithm using a single classifier J48 is the accuracy of the classification methods assessed. In this study, a single classifier is used, and a combination ensemble meta-algorithm that combines three well-known Meta-learning algorithms (bagging, boosting, and random subspace). Additionally, the resampling methodology is used to assess the accuracy of the categorization system.

Using techniques like gain ratio and Ranker, the study also demonstrates the identification of the most crucial feature of breast cancer survival. To achieve the best result with an accuracy of 95.279% and a low error rate and performance, the Ad Boost meta-learning combination with a single classifier is recommended for the classification of breast cancer.

REFERENCES

[1] J. Huang, Q. Qiao, and Y. Zhou, "Review of Breast Cancer Detection Method," 2023.

[2] S. Bhise, S. Bepari, S. Gadekar, D. Kale, A. Singh Gaur, and S. Aswale, "Breast Cancer Detection using Machine Learning Techniques," 2021. [Online]. Available: https://www.researchgate.net/publication/353285629

[3] A.Thilaka and E.Sundaravalli, "Breast Cancer Forecasting Using Machine Learning Algorithms," International Journal of Data Informatics and Intelligent Computing, vol. 2, no. 3, pp. 11–20, 2023, doi: 10.59461/ijdiic.v2i3.72.

[4] A. E. Karrar, "Investigate the Ensemble Model by Intelligence Analysis to Improve the Accuracy of the Classification Data in the Diagnostic and Treatment Interventions for Prostate Cancer," International Journal of Advanced Computer Science and Applications, vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130122.

[5] R. Yasir, A. Elsharif Karrar, W. I. Osman, M. Mutasim, and R. Y. Eltayeb, "Handling Imbalanced Data through Re-sampling: Systematic Review," Indonesian Journal of Electrical Engineering and Informatics (IJEEI), vol. 11, no. 2, pp. 503–514, 2023, doi: 10.52549/ijeei.v11i2.

[6] Rajendran, Keerthana, et al.., Predicting breast cancer via supervised machine learning methods on class imbalanced data, International Journal of Advanced Computer Science and Applications, (2020), 54-63, 11(8).

[7] Kabir, Faisal, et al.., Classification of Breast Cancer Risk Factors Using Several Resampling Approaches.

[8] Ustyannie, Windyaning, et al., Oversampling Method to Handling Imbalanced Datasets Problem In Binary Logistic Regression Algorithm, IJCCS (Indonesian Journal of Computing and Cybernetics Systems), (2020), 1, 14(1).

[9] Gupta, Ruchita, et al.., Diagnosis of Breast Cancer on Imbalanced Dataset Using Various Sampling Techniques and Machine Learning Models, Proceedings International Conference on Development in e Systems Engineering, De SE, (2021), 162-167.

[10] Steglich, Mike, et al., Proceedings of the 34th International ECMS Conference on Modelling and Simulation, ECMS 2020: June 2020, United Kingdom.

[11] Fajardo, Val Andrei, et al., VOS: a Method for Variational Oversampling of Imbalanced Data (2018).

[12] Koziarski M, Two-Stage Resampling for Convolutional Neural Network Training in the Imbalanced Colorectal Cancer Image Classification (2020).

[13] Matloob, Faseeha, at el., A Framework for Software Defect Prediction Using Feature Selection and Ensemble Learning Techniques, International Journal of Modern Education and Computer Science, (2019), 14-20, 11 (12).

[14] Mohammed, Roweida, et al., Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results, 2020 11th International Conference on Information and Communication System, ICICS 2020 (2020), 243-248.

[15] Debashree Devi, at el., 2019 International Conference on Computer Communication and Informatics (ICCCI), (2019).

[16] Nassar, Hana Babiker Classification for Imbalanced Breast Cancer Dataset Using Resampling Methods, IJCSNS International Journal of Computer Computer Science and Network Security, (2023) 23 (1)