electronics

Article

Fusion Model for Classification Performance Optimization in a Highly Imbalance Breast Cancer Dataset

Sapiah Sakri and Shakila Basheer *

Department of Information Systems, College of Computer and Information Science, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia
* Correspondence: sbbasheer@pnu.edu.sa

Abstract: Accurate diagnosis of breast cancer using automated algorithms continues to be a challenge in the literature. Although researchers have conducted a great deal of work to address this issue, no definitive answer has yet been discovered. This challenge is aggravated further by the fact that most available datasets have imbalanced class issues, meaning that the number of cases in one class vastly outnumbers those of the others. The goal of this study was to (i) develop a reliable machine-learning-based prediction model for breast cancer based on the combination of the resampling technique and the classifier, which we called a 'fusion model'; (ii) deal with a typical high-class imbalance problem, which is posed because the breast cancer patients' class is significantly smaller than the healthy class; and (iii) interpret the model output to understand the decision-making mechanism. In a comparative analysis with three well-known classifiers representing classical learning, ensemble learning, and deep learning, the effectiveness of the proposed machine-learning-based approach was investigated in terms of metrics related to both generalization capability and prediction accuracy. Based on the comparative analysis, the fusion model (random oversampling techniques dataset + extreme gradient boosting classifier) affects the accuracy, precision, recall, and F1-score with the highest value of 99.9%. On the other hand, for ROC evaluation, the oversampling and hybrid sampling techniques dataset combined with extreme gradient boosting achieved 100% performance compared to the models combined with the undersampling techniques dataset. Thus, the proposed predictive model based on the fusion strategy can optimize the performance of breast cancer diagnosis classification.

Keywords: breast cancer prediction model; class-imbalanced data; resampling techniques; ensemble learning; deep learning; classical learning; Breast Cancer Surveillance Consortium dataset

Citation: Sakri, S.; Basheer, S. Fusion Model for Classification Performance Optimization in a Highly Imbalance Breast Cancer Dataset. *Electronics* **2023**, *12*, 1168. https://doi.org/10.3390/electronics12051168

Academic Editor: Maria Evelina Fantacci

Received: 23 December 2022
Revised: 11 February 2023
Accepted: 20 February 2023
Published: 28 February 2023

1. Introduction

Breast cancer is the leading cause of cancer death in women globally, accounting for an estimated 685,000 deaths in 2020 [1]. Many countries have implemented breast screening programs, which can significantly reduce mortality rates combined with early treatment [2]. Radiologists must review mammograms (breast X-ray images) generated by these screening programs for diagnosis; this can be time-consuming, costly, and laborious [3].

Artificial intelligence and data mining have recently been comprehensively used to predict the survivability of breast cancer patients [4]. Data mining and machine-learning-based systems could improve cancer diagnosis capability and reduce diagnosis errors [5]. Many studies have applied data mining algorithms on different datasets to classify and diagnose breast cancer. These algorithms indicated good classification results and thus encouraged many researchers to take up the challenging task [6]. However, in recent years, the rampant class-imbalanced data problem has been perceived as a data mining challenge [7]. The issue of imbalanced data, particularly in the medical domain, has always been the focal point for researchers in machine learning and data mining [8]. Class imbalance refers to the imbalanced property of many real healthcare datasets. A class imbalance occurs when the majority of class instances usually outnumber the minority

class instances. The imbalanced classification problem in the healthcare domain occurs when the data are often highly skewed due to individual heterogeneity and diversity, such as in cancer diagnostics [9]. Learning from class imbalance is crucial in data mining and knowledge delivery problems. The cost of misclassification causes higher damages in the minority class (class of interest) compared to the majority [10]. One real-world example is misclassifying a breast cancer patient as a non-breast-cancer patient, which can be fatal because treatment may be delayed. Although existing methods have improved the accuracy of breast cancer diagnosis, these methods still aim to maximize accuracy, assuming that the data are well balanced [11]. The accuracy achieved from an imbalanced dataset will degrade the model performance, which is usually biased towards the majority class. Hence, it consequently leads to the misclassification of minority class instances, known as "noise", where the cost of incorrectly categorizing a minority group exceeds that of an incorrectly categorized majority group [12]. In breast cancer diagnosis, misclassifying a breast cancer patient as non-breast cancer incurs high costs and has fatal effects [13]. We presented the previous studies' review analysis and identified the gap to be addressed as the study's contribution in Section 2.

The remaining four sections have been organized in this paper. Section 2 elaborates on previous works related to this study, followed by Section 3, which explains the study's methodology. The experimental results are discussed in Section 4. Limitations of the study are elaborated on in Section 5, and we conclude the study, including some suggestions for future work, in Section 6.

2. Literature Review

This section aims to analyze previous literature on class imbalance in breast cancer clinical datasets and investigate the solutions to address this issue. This section describes the overview of the class-imbalanced issue in breast cancer datasets, an overview of techniques handling class imbalance, and an overview of the class-imbalanced issue in breast cancer studies in Sections 2.1–2.3, respectively. Section 2.4 details this study's contribution.

2.1. Overview of Class-Imbalanced Issue in Breast Cancer Dataset

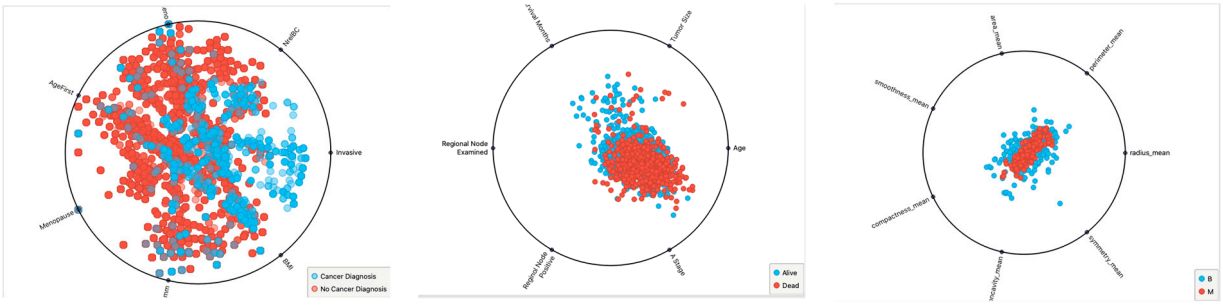
Imbalanced issues are currently a prevalent research topic and a comparatively recent area of research interest in machine learning [14]. If the sum of the data for the majority group is much more significant than that of the minority group, we have a data imbalance. The ideal dataset for most classifiers is one in which the proportion of examples from each class is approximately equal. A balanced environment is required to ensure that the classifier performs at its best. As a result, when unequal data exist, an imbalanced data problem occurs [15]. The existence of underrepresented minority groups also causes imbalanced data. It can also happen if the dataset is skewed. Most classifiers are biased toward the majority class in a balanced class [16]. All data from the real world have biases. Typically, real-world data can be divided into highly and lowly imbalanced [17].

When the imbalance ratio (IR) is excessively high compared to low imbalanced data, there is a highly imbalanced data problem. Moderate imbalance exists when the minority-to-majority ratio is 50:1; in contrast, an extreme imbalance exists when the IR is 1000:1. According to Triguero et al. [18], a dataset is considered to be moderately imbalanced if the IR is between 50:1 and 100:1. Another well-known IR standard states that if IR is 1000:1 up to 10,000:1, the dataset is considered extremely imbalanced [18]. The formula to calculate the IR is shown in Equation (1). Table 1 shows the benchmark degree of imbalanced data. Figure 1 visualization of class imbalance in breast cancer datasets. Table 2 presents the overview of commonly deployed breast cancer dataset characteristics and their imbalance ratio.

$$\text{IR} = (\text{No. of Majority Class Samples})/(\text{No. of Minority Class Samples}) \quad (1)$$

Table 1. Benchmark of the degree of imbalance in data.

Class Imbalance Degree	IR (Majority–Minority)	The Proportion of Minority Class
Extremely Imbalanced (Extreme)	10,000:1, 1000:1	<1% of the dataset
Moderately Imbalanced (Moderate)	100:1, 50:1, 10:1, 5:1	1–19% of the dataset
Mildly Imbalanced (Mild)	4:1, 2:1	20–40% of the dataset



BCSC dataset; IR 25.66:1 SEER dataset; IR 5.53:1 Prognostic dataset; IR 3.21:1

Figure 1. Illustration of class imbalance in breast cancer datasets.

Table 2. Class-imbalanced characteristics in public open breast cancer datasets (clinical data).

Dataset Name	#Instances	Class Distribution	#Attributes	Imbalance Ratio	Class Imbalance Degree
Breast Cancer	286	0:201, 1:85	9	2.36	Mild
Breast Cancer Wisconsin (Original)	699	0:458, 1:241	10	1.9	Mild
Breast Cancer Wisconsin (Prognostic)	198	0:151, 1:47	34	3.21	Mild
Breast Cancer Wisconsin (Diagnostic)	569	0:357, 1:212	32	1.69	Mild
Breast Cancer from OpenML	286	0:201, 1:85	10	2.36	Mild
Breast Cancer Coimbra	116	0:52, 1:65	10	1.25	Mild
SEER Breast Cancer Dataset	4024	Alive: 3408, Dead: 616	15	5.53	Moderate
Breast Cancer Surveillance Consortium-Risk Factor	180,465	No Risk: 173,696, Risk: 6769	13	25.66	Moderate

The most pressing issue caused by data imbalanced classification is the misdiagnosis of the minority class, which is more probable and cost-sensitive than the majority class [19]. Many machine learning algorithms have been developed to improve classification accuracy. However, this design principle leads to errors in minority class classifications. The algorithm's primary benefit is improved classification accuracy for samples from the majority class, which are taken to be more representative of the whole. For this reason, learning algorithms favor classes with more data to work with [19]. When solving classification problems, most algorithms assume or expect that the costs for each class are roughly equal. These algorithms are ineffective in dealing with complex imbalanced data sets, which are common in the real world, particularly in the medical field. Most machine learning techniques count on there being roughly the same number of examples in each category.

Consequently, problems arise for majority-class-favored learning algorithms due to the unequal number of samples. Minority groups are more critical despite their smaller representation in the overall data set. Therefore, the learning algorithms' diagnosis accuracy must be increased [20]. Misclassifying a minority group can have disastrous results, especially in the medical field, where it can delay treatment or even cause unnecessary patient pain. In addition, the most effective classification scheme should have a higher proportion of correctly diagnosed minority class diseases [20].

An imbalanced learning problem is illustrated through a real-world example, a biomedical application of a dataset obtained from a sequence of mammograms from multiple patients. Patients are classified as either positive (those with cancer) or negative (those without cancer) using binary image analysis, which generates normal classes (labels). Previous data suggest there will be more patients without cancer than those with it. For instance, Table 2 shows that there are 173,696 "negative" (majority class) samples in the BCSC-risk factor dataset, while there are 6769 "positive" samples (minority class). The best classification would yield a moderate prediction rate that reliably identified the most common and rare groups in the data (ideally 100%). However, the actual classification shows a vast disparity in accuracy, with 100% for the majority class and 0–10% for the minority class. In other words, if the minority group only had a 10% accuracy rate, 6769 patients would be mistakenly placed in the majority group.

This way, 6769 people at high risk for cancer will be given negative diagnoses. Misclassifying a healthy patient as cancerous is more expensive for healthcare providers [19]. Incorrect identification of a noncancerous cell type may require further clinical evaluation. However, a false positive for cancerous cells can devastate health and even result in death [19]. The literature shows that imbalanced data are crucial in medical diagnosis, where prediction is typically prioritized over treatment [18]. Therefore, studying the issue of class division is essential. Disparities in socioeconomic status are pervasive and impact many data-related disciplines. Most learning classifier systems have also been criticized for failing to address class imbalance adequately. The following are examples of problems with imbalanced medical data that hamper classifier learning [18]:

1. Samples with low information and low training data density: The number of samples with a stable imbalance rate is crucial in determining the efficacy of a classifier model when dealing with the class imbalance problem. The main rules discovered in a small class are suspect. Modeling classification and separating minority from majority samples also benefit from more data and information [19]. Minority groups often have inadequate data due to data bias. One article claims that a decrease in the error rate attributable to class-imbalanced distribution can be attained by collecting sufficient samples from the minority class (correcting the imbalance rate);
2. Class overlap: makes it challenging to apply separation rules, and it is challenging to categorize samples from underrepresented groups. Simple classifiers can learn correct classification without class overlap [20];
3. Small disjuncts: occur when the minority group's concept includes ancillary ideas. These auxiliary ideas, with their inequivalent class samples, further complicate the problem;
4. Noisy data: When a classifier encounters a small cluster of minority classes, it may ignore them as noise [21]. The evaluation measures used to direct the learning process may be the source of the sample's ignorance (assumed noise). Noisy samples are more likely to be found in the "safe" regions of the other classes, as defined by their labels [22];
5. Borderline samples: These are discovered where the majority and minority classes intersect. Learning algorithms struggle when dealing with borderline and noisy samples. During the training process, most classification algorithms strive to learn the most accurate borderline of each class to achieve a better prediction. Borderline samples are more likely to be misdiagnosed than nonborderline samples, making them more important for classification [22]. The difficulties associated with medical datasets are shown in Figure 2. The literature review shows that many methods have been developed to deal with the issue of imbalanced data, as discussed in Section 2.3.

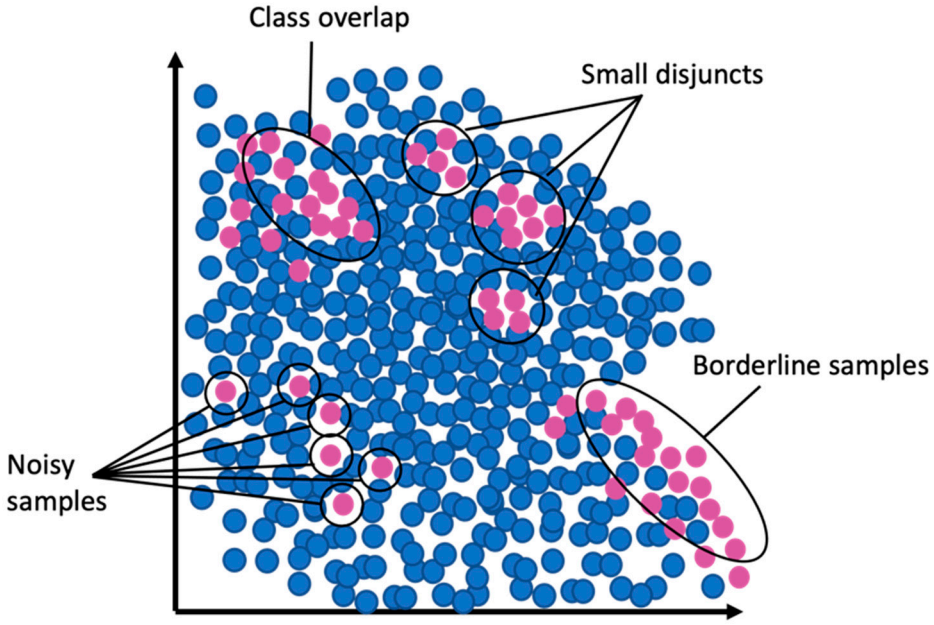


Figure 2. Imbalanced data classification problem in the medical dataset.

2.2. Overview of Techniques Handling Class-Imbalanced

There are three main approaches to dealing with class imbalance: (1) a data-level approach, (2) an algorithm-level approach, and (3) a hybrid approach. The data-level strategy for resolving the imbalanced class is preprocessing input data that balance by redistributing items across the data space. Balancing reduces the size of the majority class while increasing the size of the minority class. The methods used in this tactic can be classified as either random undersampling, oversampling, or a hybrid of the two. Class data balancing is more effective at the data level [23]. The algorithm-level strategy may develop or update existing algorithms and assess the effects of minor classes [24]. The hybrid strategy combines data-level and algorithm-level strategies to address the class imbalance problem.

A common method for addressing the issue of class differences in data is resampling. The original data set's construction is modified to achieve the optimal balance (50:50, for instance) [25]. A standard learning algorithm can be applied with no changes when using a resampled training data set. These methods have greater practicality without requiring a specific learning algorithm [26]. Correct sampling is essential for achieving a simple, even distribution of classes in a system [22]. More specifically, this paper is concerned with oversampling, undersampling, and preprocessing the data in this setting [27].

Resampling methods that change the data distribution do not care which classifier is used. Class balance is adjusted using several methods, each tailored to the specific characteristics of the sample. There are three different types of resampling: (1) oversampling: by creating new samples or reusing old ones, it validates and represents the minority group [28]. (2) Undersampling: minimizing data and removing majority class samples to confirm an equal number of samples for both classes. (3) Hybrid sampling utilizes both resampling strategies [8]. In Table 1, we present the full titles, acronyms, and citations of the balance methods discussed in this paper. What follows is a brief overview of each technique.

2.2.1. Oversampling

When dealing with an unbalanced dataset, oversampling methods add artificial samples to the underrepresented group. This process requires either resampling previously collected data from minority groups or generating new data sets. Methods of repetition include picking at random or taking representative samples from minority and majority

groups. When it does so, the classifier gives these ambiguous locations to the underrepresented sample group. Critics of oversampling say it does nothing more than rebalance minority and majority samples without providing any new insight. Methods for developing new synthetic samples to address this deficiency are being developed in promising areas [28]. Popular oversampling methods include the ones listed below:

1. ADAptive SYNthetic Sampling (ADASYN): The main idea behind this method is to assign different values to different subsets of the minority group based on their relative degree of academic challenge. Synthetic data generation is more significant for harder-to-learn samples than simpler ones [29];
2. Adjusting the Direction Of the synthetic Minority clasS examples (ADOMS): Synthetic samples are generated along the first principal component axis of geographically dispersed data [30];
3. Agglomerative Hierarchical Clustering (AHC): Selects a subset category randomly and returns the made-up samples to serve as a prototype [31];
4. Synthetic Minority Oversampling Techniques (SMOTE): Based on feature space similarities between minority class samples, the SMOTE algorithm generates synthetic data. The central idea is to draw representative samples from underrepresented groups by looking at their nearest neighbor. The feature vector of the considered sample is compared to those of its nearest neighbors, and the resulting difference is used to create new models [32];
5. Borderline SMOTE (BLSMOTE): In improving their predictions, classified algorithms work hard to acquire the most accurate learning possible. Incorrectly classifying samples closer to the boundary is more costly than misclassifying samples further away. All K neighbors of a marginal sample are included in the majority class. This equivocal approach pre-samples the underrepresented demographic using the statistical methodology known as SMOTE [33];
6. Random Oversampling (ROS): Ensures that both groups have an equal number of samples by randomly generating additional samples for the underrepresented group [34];
7. Safe-Level-SMOTE (SL-SMOTE): Before creating synthetic samples, this method determines the SMOTE-recommended threshold for minority class samples. K-nearest-neighbors calculates an appropriate sample size for underrepresented groups. Then, each synthetic sample is placed close to the lowest possible level. Therefore, all artificial samples are produced within acceptable limits. The maximum tolerable value is considered close to K, while the minimum is believed to be close to zero [35].

2.2.2. Undersampling

The original data set is undersampled to meet a specified ratio of missing to nonmissing values. Elimination can be performed randomly or with the help of more efficient specific standards, such as eliminating borderline samples. The correct classification of minority samples is achieved using the second method, which efficiently decreases the space allocated to the majority class. This, however, runs the risk of wiping out some crucial data [24]. These are some of the most common methods of sampling:

1. Condensed Nearest Neighbor (CNN): This method may be less useful for learning because it discards the vast majority of class samples on the cusp of a decision. We can make the dataset smaller by omitting some aspects from the original data set while leaving the classifier NN unchanged [36];
2. Tomek Link (TL): Specifically, the method employs Tomek links x and y, where x is a member of the majority class, and y is a member of the minority class, and where their distance is smaller than that of any other sample, such as z. Tomek-linked samples are either at the class boundary or contain spurious information. Many studies have employed TL as a method of guided undersampling by excluding data from the dominant group [37];
3. Condensed Nearest Neighbor–TL (CNN-TL): As with the OSS algorithm, this method combines CNN and TL. The mutually beneficial subset is identified in advance of

applying TL. As described in Tomek's linked article [38], the goal is to prune the data set by omitting particular elements using CNN, which significantly affects the performance of the NN classifier;

4. Edited Nearest Neighbor (ENN): Another way to eliminate examples. Using k = 3 nearest neighbors, you can identify misclassified cases in a dataset and remove them from further analysis [38];
5. Neighborhood Cleaning Rule (NCL): In this strategy, we eliminate the samples from the majority class by applying the ENN rule. Class labels of the eliminated samples differ from those of 3–5 neighbors [39];
6. One-Sided Selection (OSS): For TL detection, 1-NN chooses all samples from the minority class and some misclassified samples from the majority class. Most of the Tomek link's encapsulated class samples are omitted [40];
7. Random Undersampling (RUS): In maintaining parity between groups, most class samples are purged randomly [41];
8. Undersampling Based on Clustering (SBC): The theory assumes that any given data set will contain subsets with varying degrees of similarity. First, we cluster all of the training samples. The samples would then establish the purposes of the clusters. The functions of the clusters are equivalent regardless of whether the samples come from the minority or the majority. Similarly, the SBC method achieves class balance by randomly selecting a large number of samples from clusters belonging to the majority class based on the ratio of the two classes [42].

2.2.3. Hybrid Sampling

This sampling strategy is a hybrid between oversampling and undersampling. There are several popular hybrid sampling methods, including:

1. SMOTE-ENN: This technique is a hybrid, built on the foundation of SMOTE by ENN, and filters out unwanted background noise. To further narrow down the sample space, ENN can filter out data that do not fit either classification. Mislabeled samples are weeded out by comparing them to their three nearest counterparts [38];
2. SMOTE-TL: This hybrid approach employs SMOTE to eliminate data containing the Tomek link selectively. To create a Tomek connection, you need to find two samples nearest to each other but do not share the same category. Before detecting and eliminating Tomek links, SMOTE oversamples the original data set. Thus, a uniform dataset with clearly delineated class clusters is generated [38];
3. Selective Processing of Imbalanced Data (SPIDER): This method employs intricate sample filtering on the majority group and oversampling the local minority group. Misclassified or noisy samples can be identified with the help of the K-nearest neighbor (KNN). Then, depending on what you choose (weak, firm, or relabeling), the noisy objects are either repeated or removed [39];
4. Selective Processing of Imbalanced Data 2 (SPIDER2): Preprocessing samples from both the majority and the minority classes is the first step in this method. The majority class characteristics are defined, and then the noisy samples are found and either discarded or relabeled according to the relabeling options available (reclassified as the minority class). The same thing happens to the minority group, and their noisy samples are repeated after being uncovered [40].

The class imbalance issue is typically resolved through either oversampling or undersampling methods. Although oversampling helps preserve the large dataset of the majority class, it can make the classifier-training process more time-consuming and introduce the possibility of overfitting. However, under selective sampling's sampling means that it can be trained in a relatively short amount of time. On the other hand, undersampling can cause significant samples to be lost [41].

2.3. Overview of Class-Imbalanced Issues in Breast Cancer Studies

There are numerous works on breast cancer forecasting. However, few studies have focused on methods of dealing with class imbalance. Even though, to our knowledge and as evidenced by the review (see Table 2), most breast cancer datasets were found to have a class-imbalanced issue. As a result, our review only includes studies that have provided evidence on how to deal with this issue specifically to address the class overlap, small disjuncts, borderline samples, and noisy sample issues discussed and their success with the proposed methods or combination of methods. The previous studies are summarized in Table 3. The following paragraph contains a discussion of the review.

San et al. [42] identified and classified breast cancer risk factors using four classifiers: LR, RF, SVM, and MLP. These four classifiers are trained and tested using data splits of 80–20, 70–30, and 60–40. The SMOTE data resampling technique significantly improves the precision–recall rate. The combination of SMOTE and RF classifiers produces the best fusion model.

Huang et al. [43] used SMOTE in conjunction with information gain (IG) and genetic algorithm (GA) feature selection methods to resample the class-imbalanced. For highly class-imbalanced datasets, the experimental results based on two breast cancer datasets show that the combination outperforms either feature selection or oversampling alone. SMOTE + GA + SVM is the most effective fusion model.

Vuttipittayamongkol and Elyan [44] investigated an undersampling method based on the recursive neighborhood (URNS) for classifying imbalanced data by exploring neighborhood instances using the recursive process. Experiments' outcomes demonstrate URNS's efficacy in classifying poorly balanced breast cancer datasets. URNS outperformed well-established and state-of-the-art methods on most datasets by achieving the highest sensitivity and G-mean. Sensitivity and G-mean were significantly improved over the baseline across all datasets (RF with no resampling).

Wang et al. [45] proposed a new framework for entropy and confidence-based undersampling boosting (ECUBoost). ECUBoost is distinguished by three features that set it apart from previous works. The incorporation of dynamic resampling methods with confidence into the boosting ensemble; the use of entropy and confidence of instances as evaluation standards; instance selection; and an effective undersampling-based ensemble learning system, which extends the concept of dynamic resampling methods with the boosting ensemble along with a novel data preprocessing technique to a broader ground.

Al-Shamaa et al. [46] presented the Hellinger distance undersampling (HDUS) method to resample the data using three classification algorithms (DT, SVM, and KNN). They compared it to the baseline model (without resampling method) and three state-of-the-art undersampling methods (Tomek link, RUS, and ENN).

Desuky and Hussain [47] combined a novel simulated annealing (SA) strategy with machine learning classifiers for the first time to balance imbalanced datasets. SA is a better hybrid method for dealing with unbalanced settings and thus improving overall performance. The process of achieving the best solution for a problem (selecting an optimal subset of majority class instances) is known as optimization; in this work, the simulated annealing optimization technique aids in improving the objective function (classification performance) value. They employ undersampling techniques such as simulated annealing, discriminant analysis (DA), SVM, DT, and kNN.

Zhang et al. [48] used two standard breast cancer data sets and 12 representative imbalanced data sets to assess the model effectiveness of AK-Boosted C5.0. Furthermore, the statistical test analysis shows that AK-Boosted C5.0 is effective. The results show that the proposed AK-Boosted C5.0 method can significantly improve classification performance without feature selection algorithms and in less time. Furthermore, a cluster-based undersampling algorithm may be a better resampling alternative.

Koziarski [49] presented radial-based undersampling, a novel undersampling algorithm based on the previously introduced concept of mutual class potential. The proposed algorithm was motivated primarily by the desire to apply the idea of non-nearest

neighbor-based resampling, which had once been used in radial-based oversampling, to the undersampling procedure. The proposed method is conceptually more straightforward and more computationally efficient than the radial-based oversampling algorithm.

Zhang and Chen [50] used a sample-based method to reduce the imbalanced effects in breast cancer diagnosis (a hybrid method based on random oversampling examples (ROSE), K-means, and SVM). The performance metrics evaluation results showed that the proposed K-means, ROSE, and SVM methods are the best fusion models for dealing with imbalanced datasets.

Rajendran et al. [51] solved the class imbalance problem by combining SpreadSample and SMOTE with a SpreadSubsample/SMOTE hybrid. Classification models were built using C4.5, Bayesian network, and RF based on this resampling method. The results showed that the Bayesian network generated by the hybrid sampling methods was a better model for a decision support system, especially for the early diagnosis and treatment of breast cancer patients.

Tran et al. [52] investigated an engineered upsampling (ENUS) method for dealing with imbalanced data to improve the predictive performance of machine learning models. When the minority–majority class ratio is less than 20%, ENUS training models improve balanced accuracy by 3.74 percent, sensitivity by 8.36 percent, and F1-score by 3.83 percent. In addition, our research discovered that the XGBoost tree (XGBTree) using ENUS produced the best results.

To address data class imbalance problems, Ibrahim [53] proposed a salp swarm optimization-based undersampling technique (SSBUT). Using the proposed SSBUT, the similarity relationship among the majority class samples is thoroughly examined, and samples that do not affect the classification algorithm's accuracy are removed from the majority class. The proposed SSBUT's performance was tested on benchmark medical imbalanced datasets, and the results were compared to state-of-the-art undersampling techniques. Regarding various evaluation criteria, the experimental results show that the proposed SSBUT consistently outperforms state-of-the-art undersampling methods.

On five unbalanced clinical datasets (Breast Cancer Disease, Coronary Heart Disease, Indian Liver Patient, Pima Indians Diabetes Database, and Coronary Kidney Disease), Kumar et al. [54] compared the empirical performance of seven class balancing techniques (decision tree, K-nearest neighbor, logistic regression, artificial neural network, support vector machine, and Gaussian naïve Bayes). The SMOTE-ENN balancing method achieves 99.8%, 99.5%, 99.1%, and 99.1% accuracy when using KNN, SVM, LR, and ANN, respectively.

The synthetic minority oversampling technique (SMOTE) was used by Mahesh et al. [55] to deal with the problem of imbalanced data in the class and noise. The suggested task consists of two steps. SMOTE is used in the first phase to reduce the impact of imbalanced data issues. Then, data are classified using the naïve Bayes, decision trees classifier, random forest, and their ensembles in the second phase. The XGBoost-random forest ensemble classifier outperforms with 98.20 percent accuracy in the early detection of breast cancer, according to the experimental results.

In CDSMOTE, Elyan et al. [56] used integrated class decomposition (CD) for the majority class and SMOTE for the minority class. The majority of instances are grouped into clusters based on similarity in the first phase to reduce the dominance of the majority class without losing information. Following that, oversampling is performed to ensure that the distribution of people in the minority class is even.

Based on previous research, this study aims to provide a detailed analysis of imbalanced data and its characteristics. The purpose of this study is to examine the effect of data preprocessing on the performance of classifiers in the interest of improving early breast cancer diagnosis and treatment by studying data that is imbalanced in several ways. Table 3 presents a detailed summary of research works on breast cancer prediction. A combination of techniques and a predictive classifier aid in the improvement of machine learning results. When compared to a single model, this approach produces better predictive performance. As shown in Figure 2, this study proposes a new hybrid model that combines ROS and

integrates boosting techniques (XGBoost) to develop a robust early breast cancer prediction model to fill the research gap. The fusion model is expected to resolve the imbalanced data problem of class overlap, small disjuncts, borderline samples, and noisy samples.

Table 3. Summary of the reviewed studies using resampling techniques in Breast Cancer prediction.

Study	Year	Proposed Resampling Technique	Resampling Type	Classifiers	Best Fusion Model with the Highest Accuracy
[43]	2022	SMOTE	Oversampling	LR, RF, SVM, MLP	SMOTE + RF
[44]	2021	SMOTE + feature selection (IG and GA)	Oversampling	SVM	SMOTE + GA + SVM
[45]	2020	Overlap-based undersampling (URNS)	Undersampling	KNN RF	URNS + RF
[46]	2020	Undersampling boosting (ECUBoost)	Undersampling	RF	ECUBoost + RF
[47]	2020	Hellinger distance undersampling (HDUS)	Undersampling	kNN, SVM, DT	HDUS + DT
[48]	2021	Undersampling using simulated annealing (SA)	Hybrid sampling	SVM, DT, kNN, and discriminant analysis (DA)	SA + kNN
[49]	2021	Cluster-based undersampling	Undersampling	Boosted C5.0	Cluster-based undersampling + Boosted C5.0
[50]	2020	Radial-based undersampling (RBU)	Undersampling	CART, kNN, NB, SVM	RBU + NB
[51]	2019	ROSE + K-means	Oversampling	SVM	ROSE + K-means + SVM
[52]	2020	SpreadSample + SMOTE	Hybrid sampling	C4.5, Bayesian network, and RF	SpreadSample + SMOTE + Bayesian network
[53]	2022	Engineered upsampling method (ENUS)	Oversampling	XGBoost tree (XGBTree), kNN, DT, RF, ANN, SVM	ENUS + XGBoost
[54]	2022	Salp swarm optimization-based undersampling technique (SSBUT)	Undersampling	C4.5, SVM, NB	SSBUT + C4.5
[55]	2022	undersampling, random oversampling, SMOTE, ADASYN, SVM-SMOTE, SMOTEEN, and SMOTETOMEK.	Hybrid sampling	DT, KNN, LR, ANN, SVM, and NB	SMOTE-ENN + kNN
[56]	2022	SMOTE	Oversampling	NB, DT, RF, XGBoost, XGBoost-NB, XGBoost-DT, XGBoost-RF	SMOTE + XGBoost-RF
[57]	2021	SMOTE, ADASYN, CD, CDSMOTE	Oversampling	Boosting, SVM, RF	CDSMOTE + RF

2.4. Research Contribution

The following contributions are the result of systematic experimental work in this study. Overall, we present an end-to-end machine-learning-based model classification in dealing with extremely class-imbalanced datasets, which consists of the steps listed below:

1. To reduce the imbalance ratio of the BCSC dataset, nine different state-of-the-art resampling techniques, which include Random undersampling (RUS), edited nearest neighbor undersampling (ENN), Tomek links undersampling (TL), random oversampling (ROS), SMOTE, borderline SMOTE (BLSMOTE), SMOTE + edited nearest neighbor undersampling (SMOTE-ENN), SMOTE + Tomek link undersampling (SMOTETomek) and SPIDER, are harnessed;
2. We constructed 27 different fusion models via pretraining of three renowned classifiers, namely, extreme gradient boosting (XGBoost), artificial neural network (ANN), and support vector machine (SVM);
3. In evaluating the efficiency and effectiveness of the models, the study deployed six performance metrics such as confusion matrix, accuracy, precision, recall, F-score, and "area under the curve" (AUC) of "receiver characteristic operator" (ROC) (AUC-ROC);
4. Assessing the performance of the 27 fusion models by performing a comparative analysis between the proposed algorithm with the rest of the algorithms;
5. To validate the results by analyzing the differences between all classifiers and indicating the best classifier;
6. To compare the performance of the new proposed fusion model with the state-of-the-art predictive models applied to the BCSC dataset as a benchmark dataset for experimental validation.

3. Materials and Methods

3.1. Proposed Method

The proposed method in the study predicts breast cancer by combining various resampling techniques with several selected machine-learning-based classifiers. The goal is to find the best combination (fusion model) using the confusion matrix, accuracy, precision, recall, F1-score, and ROC. We used three undersampling, three oversampling, and three hybrid sampling techniques to balance the dataset. Three well-known classifiers are used for model classification. Figure 3 depicts the research's conceptual framework.

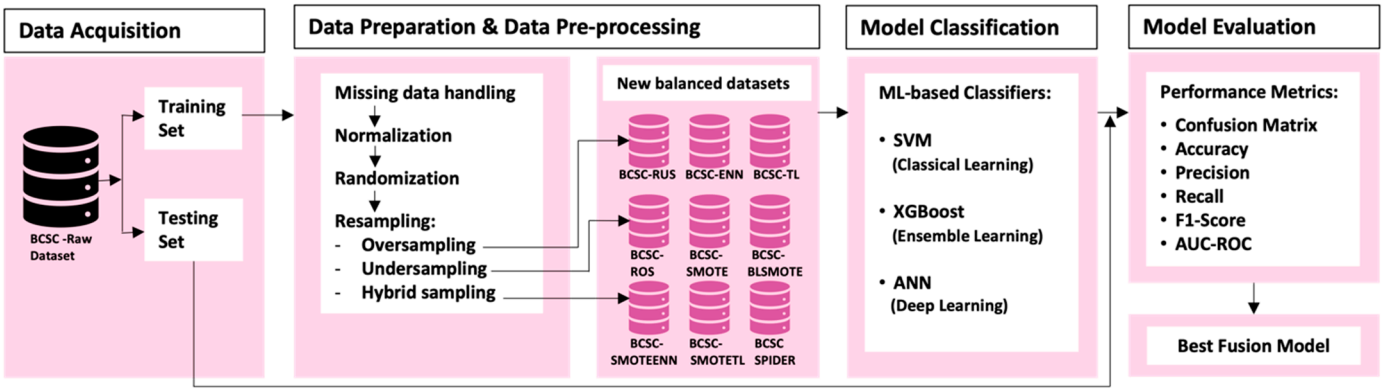


Figure 3. Proposed research method.

3.2. Dataset Acquisition

The BCSC dataset was acquired by downloading it from https://www.bcsc-research.org/data/rfdataset/risk-estimation-dataset-download (accessed on 14 February 2022). The dataset consisted of 280,660 breast cancer mammography data records among women aged between 35 and 54 years (the mammography image was not included). Of the 280,660 data samples, 180,465 samples were trained and validated. Only trained and validated data was used in this study to ensure a more accurate result. A total of 173,696 were labeled as 'no cancer', while 6769 were labeled as 'cancer' Based on the data distribution, the dataset is moderately imbalanced based on the IR benchmark (see Section 2.1). Table 4 presents the dataset attributes and description. The distribution is shown in Figure 4.

Table 4. BCSC dataset description.

S#	Variable	Short Name
1	Menopausal status	menopause
2	Age group	agegrp
3	Breast density	density
4	Race	race
5	Hispanic	hispanic
6	Body mass index	bmi
7	Age at first birth	Agefirst
8	Number of first-degree relatives with breast cancer	nrelbc
9	Previous breast procedure	brstproc
10	Result of the last mammogram before the index mammogram	lastmm
11	Surgical menopause	surgmeno
12	Current hormone therapy	hrt
13	Diagnosis of invasive breast cancer within one year of the index screening mammogram	invasive
14	Diagnosis of invasive or ductal carcinoma in situ breast cancer within one year of the index screening mammogram	cancer
15	Training data	training
16	Frequency count of this combination of covariates and outcomes (all variables 1 to 14)	count

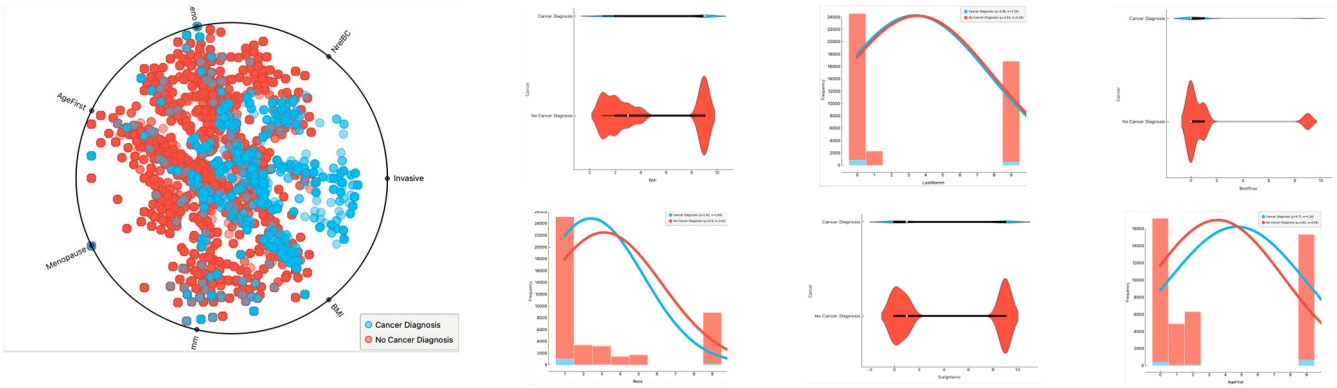
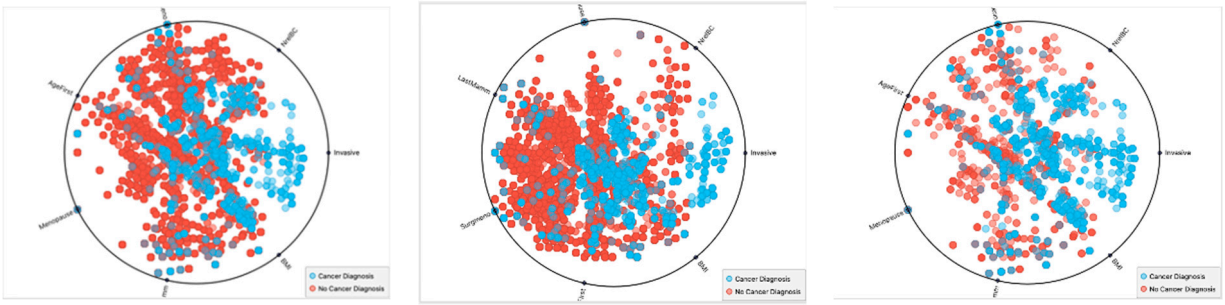


Figure 4. Illustration of BCSC dataset data properties.

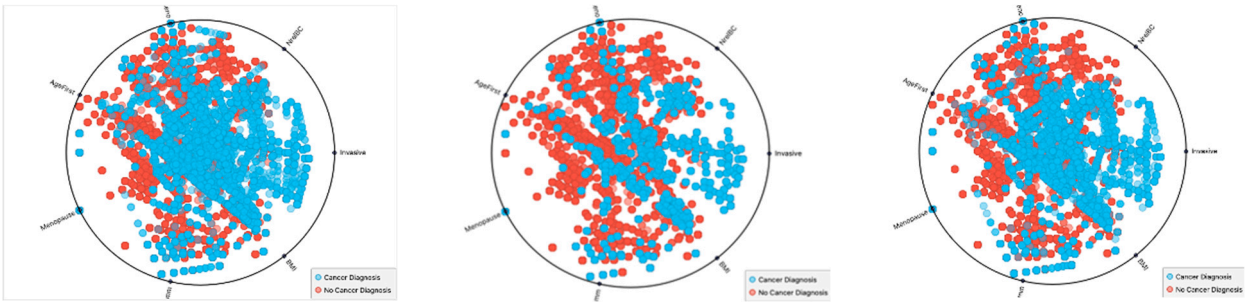
3.3. Data Preprocessing: Resampling of BCSC Dataset

The most crucial step in achieving the best classification results is preprocessing. It is frequently applied to data before classification to ensure that the desired results are obtained. Preprocessing strategies for the breast cancer dataset are being researched to improve detection model accuracy, reduce computational time, and accelerate training. Furthermore, by normalizing the data, the optimizer may achieve a mean (μ) = 0 and a standard deviation (σ) = 1, allowing it to converge more quickly. To balance the BCSC dataset, we used five undersampling techniques (RUS, ENN, TL), four oversampling techniques (SMOTE, BLSMOTE, ROS), and two hybrid sampling techniques (SMOTE-ENN, SMOTE-TL, SPIDER). Section 2 goes over all of the techniques. The visualization of the new balanced dataset (after using undersampling, oversampling, and hybrid sampling techniques) is shown in Figures 5–7.



BCSC-TL BCSC-ENN BCSC-RUS

Figure 5. Effect of undersampling techniques on the BCSC dataset.



BCSC-SMOTE BCSC-ROS BCSC-BLSMOTE

Figure 6. Effect of oversampling techniques on the BCSC dataset.

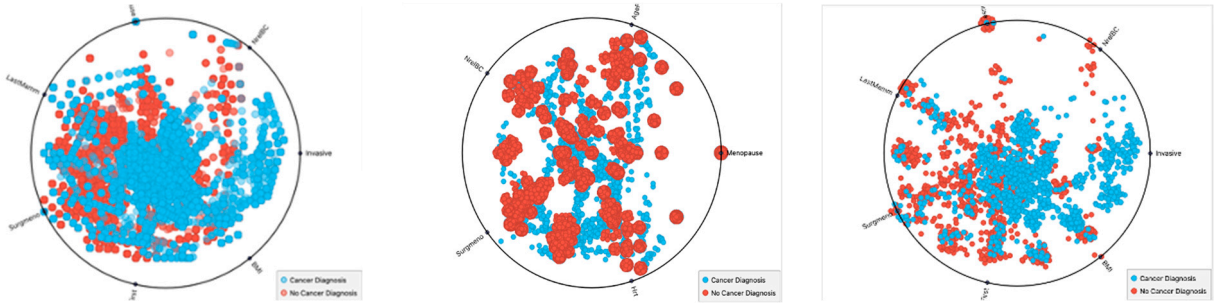


Figure 7. Effect of hybrid sampling techniques on the BCSC dataset.

Ref. [57] discovered that the standard resampling techniques deployed in the health domain for (1) oversampling techniques are SMOTE, BLSMOTE, ESL-SMOTE, IQR-SMOTE, KMPP-SMOTE; (2) undersampling techniques are ENN and KMPP-US; and (3) hybrid sampling techniques are SMOTE-ENN, SPIDER, and AHC-OS + KM-US. Figure 8 presents the outcome of their review. Based on the result, we proposed using the above-mentioned resampling techniques in this study because, according to the literature review and to the best of our knowledge, not all of the techniques presented in this paper have been used to balance the BCSC dataset. This study will determine which resampling technique improved the model's accuracy, precision, recall, F-measure, and ROC.

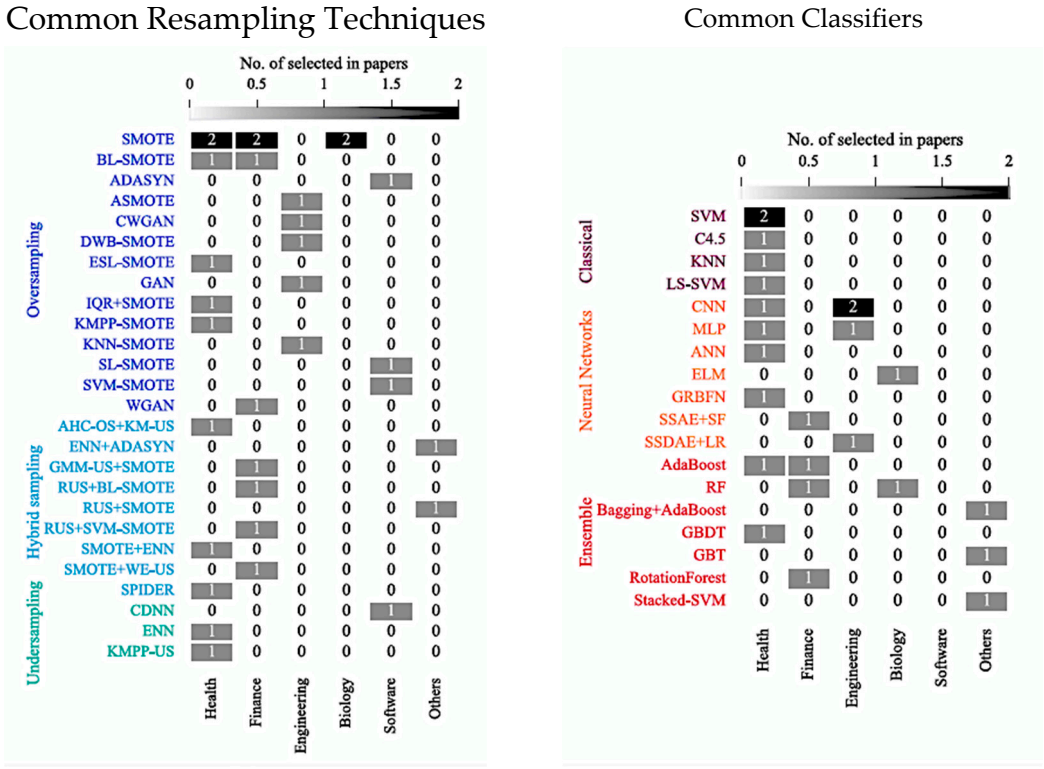


Figure 8. Common resampling techniques and classifiers deployed in previous studies [57].

3.4. Model Classification

In assessing the classification accuracy of a breast cancer model, this study used support vector machine (SVM) to represent classical machine learning, artificial neural network (ANN) to represent deep learning, and extreme gradient boosting (XGBoost) to represent ensemble learning. The classifiers were chosen based on the findings of Werner et al. [57], who identified SVM as the most commonly used classifier in health domain

research, specifically in class-imbalanced research. ANN and XGBoost are two other popular classifiers used in this research domain. Figure 8 depicts the summary result of their review. The following are the deployed classifiers in this study:

1. XGBoost: This algorithm is a highly efficient version of a gradient-boosted decision tree. By providing a wrapper class, XGBoost facilitates using models in the scikit-learn framework as either classifiers or regressors. XGBoost's classification model goes by the name XGB Classifier. Since it was designed and developed with model performance and computational speed in mind, XGBoost is an efficient and accurate execution of gradient boosting machines that have confirmed to push the boundaries of computing power for upgraded tree algorithms. With the tree-boosting algorithm in mind, it was built to use all available memory and processing powerfully. XGBoost is popular because it can be applied to many machine learning and data mining problems. In 2015, for instance, 17 out of the 29 winning challenge solutions published on the ML competition site Kaggle all made use of XGBoost;
2. ANN: The algorithm for artificial neural networks is inspired by the structure and function of real neurons, down to the minor details of the dendrites, somas, and axons. Each ANN comprises a network of artificial neurons, each of which performs an essential mathematical operation. An artificial neural network consists of input, hidden, and output layers, each a collection of linked neurons. This network acquires the ability to perform tasks by observing a large enough sample of similar instances. Classification and regression issues are well within the neural networks' capabilities. Advanced perception versions of ANNs, called multilayer ANNs, can be used to tackle difficult classification and regression issues. When it comes to binary classification, perception ANNs are by far the most popular. In our classification work, we used a very similar set of ANNs. In an ANN, the number of neurons in the input layer is proportional to the number of features in the data set used for training. Even a network's hidden layer can be considered an independent entity. The research uses an input layer with 31 neurons connected to the first hidden layer's nine neurons. The links between the first and second secret layers have been mapped out to the extent of 9-9. It is a binary classification problem, so there's only one neuron in the output layer;
3. SVM: Using hyperplanes, SVM separates data for classification purposes. SVM is based on using hyperplanes to classify data into similar groups. Data with nonregularities and unknown distributions are ideal candidates for the SVM method. Our research uses the caret and kern lab packages to construct and tune the SVM model's hyperparameters. Specifically, we used a grid search algorithm to determine the best values for our model's hyperparameters.

In this study, we proposed to use XGBoost, ANN, and SVM as machine-learning-based classifiers to be trained on a balanced dataset that was resampled using various techniques. The experiment aims to see how effectively each combination technique improves the models' ability to predict breast cancer.

3.5. Performance Metrics of Classifiers

We use standard machine learning performance metrics to evaluate the classification performance of the prediction models. To describe the performance of the models, we use the confusion matrix, as shown in Figure 9. It consists of the following metrics:

1. TP (true positive)—The number of observations that the model classified as 'positive' and that are actually 'positive';
2. FP (false positive)—The number of observations that the model classified as 'Positive' but are actually 'Negative'. It is also called a Type-1 error;
3. TN (true negative)—The number of observations that the model classified as 'Negative' and are actually 'Negative';
4. FN (false negative)—The number of observations that the model classified as 'Negative' but are actually 'Positive'. It is also called a Type-2 error.

	Predicted: Cancer	Predicted: No Cancer
Actual: Cancer	TP	FN (Type-2 error)
Actual: No Cancer	FP (Type-1 error)	TN

Figure 9. Confusion Matrix.

The following metrics are used to measure the effectiveness and efficiency of the prediction models:

$$\text{Accuracy (overall performance)} = (TP + TN)/(TP + FP + TN + FN); \quad (2)$$

$$\text{Precision (model predictive power)} = TP/(TP + FP); \quad (3)$$

$$\text{Recall (hit rate)} = TP/(TP + FN); \quad (4)$$

$$\text{F1-Score} = (2 * \text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}); \quad (5)$$

ROC (receiver operating characteristic)—A quick assessment of the model's quality. The nearer to 1, the better the model. (6)

4. Results and Discussion

We used Python for data visualization and the Orange3 data mining tool to implement the experiments. We first conduct the model classification involving SVM, XGBoost, and ANN classifiers on nine balanced datasets balanced by nine resampling techniques. The results of the model classification are presented and discussed in Section 4.1. The confusion matrix analysis of all the datasets is presented and discussed in Section 4.2. The AUC-ROC analysis results for all the datasets and classifiers are presented and discussed in Section 4.3. The efficiency analysis of the time taken to build the models is presented in Section 4.4. Lastly, the comparative analysis between models in terms of all the performance measures is presented in Section 4.5.

4.1. Model Classification Evaluation

The experimentation was carried out using three machine-learning-based models: the SVM (classical learning), the XGBoost (ensemble learning), and the ANN (deep understanding) model. The model classification results are presented in Tables 5–7. Table 8 shows the models' overall performance.

Table 5. Effect of resampling techniques on SVM classifier performance.

	Undersampling			**Oversampling**			**Hybrid Sampling**		
Metrics	**RUS**	**ENN**	**TL**	**ROS**	**SMOTE**	**BLSMOTE**	**SMOTE-TL**	**SMOTE-ENN**	**SPIDER**
Accuracy	0.817	0.914	0.897	0.688	0.549	0.596	0.575	0.607	0.879
Precision	0.817	0.967	0.967	0.706	0.612	0.611	0.630	0.621	0.881
Recall	0.817	0.914	0.897	0.688	0.549	0.596	0.575	0.607	0.879
F1-Score	0.817	0.934	0.924	0.681	0.475	0.581	0.525	0.587	0.879
ROC	0.882	0.843	0.876	0.780	0.728	0.713	0.762	0.743	**0.931**

Table 5 observes that the ENN technique improved the SVM classifier by obtaining the highest accuracy (91.4%%), highest precision (96.7%), highest recall (91.4%), and highest F1-score (93.4%). On the other hand, SPIDER improved the model by achieving the highest ROC with 93.1%.

Table 6 indicates that ROS improved the XGBoost classifier by attaining the highest accuracy (99.9%), precision (99.9%), recall (99.9%), F1-score (99.9%), and AUC-ROC (100%). SMOTE, BLSMOTE, SMOTE-TL, and SMOTE-ENN improved the model by achieving the highest ROC of 100%.

Table 6. Effect of resampling techniques on XGBoost classifier performance.

	Undersampling			Oversampling			Hybrid Sampling		
Metrics	**RUS**	**ENN**	**TL**	**ROS**	**SMOTE**	**BLSMOTE**	**SMOTE-TL**	**SMOTE-ENN**	**SPIDER**
Accuracy	0.934	0.996	0.996	0.999	0.994	0.997	0.994	0.997	0.983
Precision	0.935	0.996	0.996	0.999	0.994	0.997	0.994	0.997	0.983
Recall	0.934	0.996	0.996	0.999	0.994	0.997	0.994	0.997	0.983
F1-Score	0.934	0.996	0.996	0.999	0.994	0.997	0.994	0.997	0.983
ROC	0.980	0.984	0.984	1.000	1.000	1.000	1.000	1.000	0.998

Table 7. Effect of resampling techniques on ANN classifier performance.

	Undersampling			Oversampling			Hybrid Sampling		
Metrics	**RUS**	**ENN**	**TL**	**ROS**	**SMOTE**	**BLSMOTE**	**SMOTE-TL**	**SMOTE-ENN**	**SPIDER**
Accuracy	0.915	0.994	0.993	0.995	0.983	0.990	0.983	0.991	0.958
Precision	0.915	0.994	0.993	0.995	0.983	0.990	0.983	0.991	0.958
Recall	0.915	0.994	0.993	0.995	0.983	0.990	0.983	0.991	0.958
F1-Score	0.915	0.994	0.993	0.995	0.983	0.990	0.983	0.991	0.958
ROC	0.963	0.972	0.967	0.999	0.998	0.999	0.998	0.999	0.965

Table 7 observes that the ROS technique improved the ANN classifier in all the metrics at 99.5% and the highest ROC at 99.9%. BLSMOTE and SMOTE-ENN improved the classifier by achieving the highest ROC of 99.9%.

Table 8. Overall performance of model classification.

Metric	Highest Value	Classifier	Resampling Strategy	The Best Fusion Model
Accuracy	99.9%	XGBoost	ROS	ROS + XGBoost
Precision	99.9%	XGBoost	ROS	ROS + XGBoost
Recall	99.9%	XGBoost	ROS	ROS + XGBoost
F1-Score	99.9%	XGBoost	ROS	ROS + XGBoost
			ROS	ROS + XGBoost
			SMOTE	SMOTE + XGBoost
ROC	100%	XGBoost	BLSMOTE	BLSMOTE + XGBoost
			SMOTE-ENN	SMOTE-ENN + XGBoost
			SMOTE-TL	SMOTE-TL + XGBoost

Table 8 presents that XGBoost outperformed the rest of the classifiers by achieving the highest value of 99.9% in terms of accuracy, precision, recall, F1-score, and ROC at 100%. The result also indicates that the dataset balanced by ROS produced appropriate useful samples for training compared to the rest of the datasets. The best fusion model for accuracy, precision, recall, and F1-score is XGBoost + ROS. On the other hand, the best fusion model for ROC is XGBoost + ROS, XGBoost + SMOTE, XGBoost + BLSMOTE, XGBoost + SMOTE-ENN, and XGBoost + SMOTE-TL.

4.2. Confusion Matrix Analysis

To check the model's classification performance, we implemented a confusion matrix. The confusion matrix classifies breast cancer as 'cancer diagnosed' or 'no cancer diagnosed.' We presented the confusion matrix for each of the balanced datasets to observe which classifiers perform better in which dataset. The comparative analysis of the confusion matrix between the datasets based on the resampling techniques is presented in Tables 9–11 for datasets resampled by oversampling, undersampling, and hybrid sampling techniques, respectively.

The overall performance of the confusion matrix indicates that the fusion model ROS + XGBoost outperformed the rest of the fusion model predicting correctly the highest number

of 'cancer diagnoses' instances (42,237) and the highest number of 'no cancer diagnose' instances (42,146). The result supports the overall performance of the model classification, as shown in Table 8.

Table 9. Confusion matrix of the fusion model for datasets balanced using oversampling techniques.

		Predicted		
		Cancer Diagnosis	No Cancer Diagnosis	Σ
ROS + XGBoost	Actual Cancer Diagnosis	42237	0	42237
	Actual No Cancer Diagnosis	91	42146	42237
	Σ	42328	42146	84474
		Predicted		
		Cancer Diagnosis	No Cancer Diagnosis	Σ
ROS + ANN	Actual Cancer Diagnosis	42237	0	42237
	Actual No Cancer Diagnosis	421	41816	42237
	Σ	42658	41816	84474
		Predicted		
		Cancer Diagnosis	No Cancer Diagnosis	Σ
ROS + SVM	Actual Cancer Diagnosis	35284	6953	42237
	Actual No Cancer Diagnosis	19378	22859	42237
	Σ	54662	29812	84474
		Predicted		
		Cancer Diagnosis	No Cancer Diagnosis	Σ
SMOTE + XGBoost	Actual Cancer Diagnosis	42039	198	42237
	Actual No Cancer Diagnosis	304	41933	42237
	Σ	42343	42131	84474
		Predicted		
		Cancer Diagnosis	No Cancer Diagnosis	Σ
SMOTE + ANN	Actual Cancer Diagnosis	41694	543	42237
	Actual No Cancer Diagnosis	935	41302	42237
	Σ	42629	41845	84474
		Predicted		
		Cancer Diagnosis	No Cancer Diagnosis	Σ
SMOTE + SVM	Actual Cancer Diagnosis	39031	3206	42237
	Actual No Cancer Diagnosis	34909	7328	42237
	Σ	73940	10534	84474
		Predicted		
		Cancer Diagnosis	No Cancer Diagnosis	Σ
BLSMOTE + XGBoost	Actual Cancer Diagnosis	42121	116	42237
	Actual No Cancer Diagnosis	176	42061	42237
	Σ	42297	42177	84474
		Predicted		
		Cancer Diagnosis	No Cancer Diagnosis	Σ
BLSMOTE + ANN	Actual Cancer Diagnosis	42014	223	42237
	Actual No Cancer Diagnosis	609	41628	42237
	Σ	42623	41851	84474
		Predicted		
		Cancer Diagnosis	No Cancer Diagnosis	Σ
BLSMOTE + SVM	Actual Cancer Diagnosis	32987	9250	42237
	Actual No Cancer Diagnosis	24903	17334	42237
	Σ	57890	26584	84474

Table 10. Confusion matrix of the fusion model for datasets balanced using undersampling techniques.

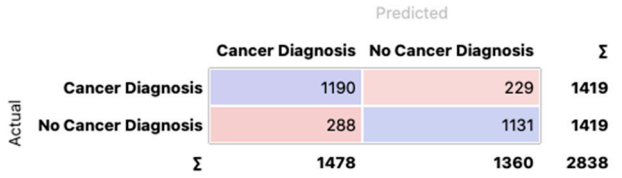
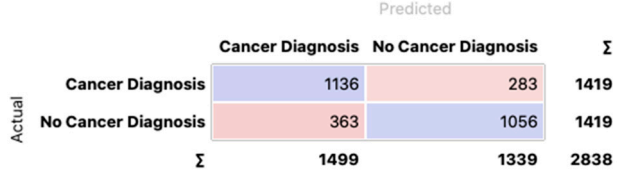
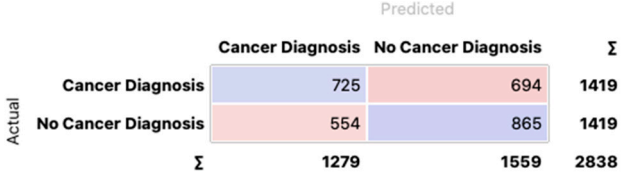
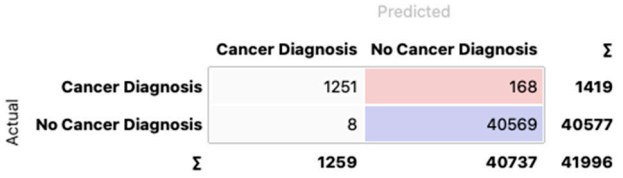
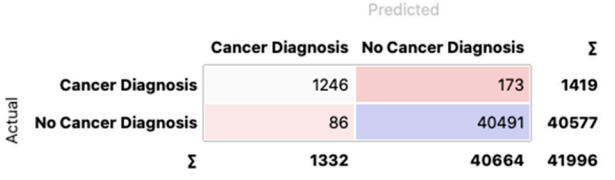
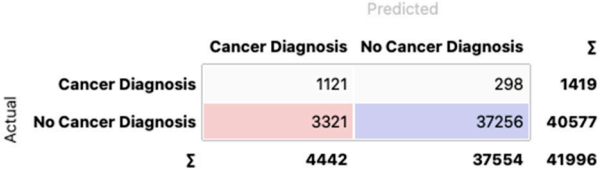
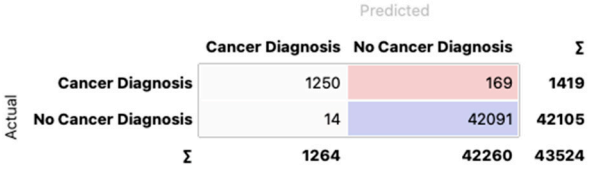
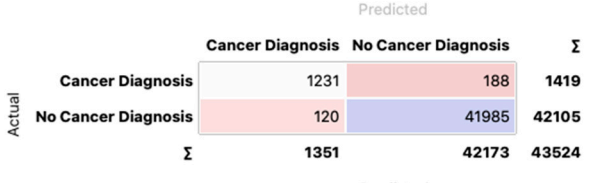
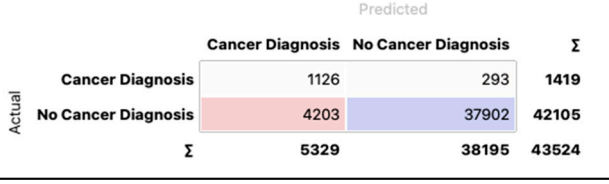
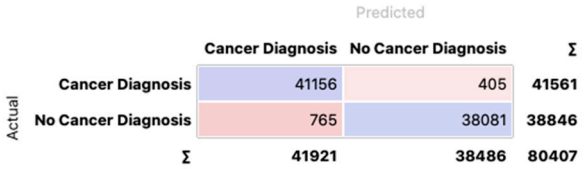
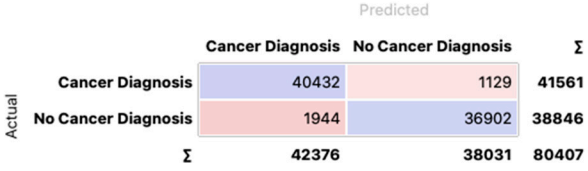
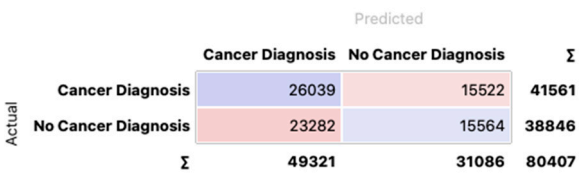
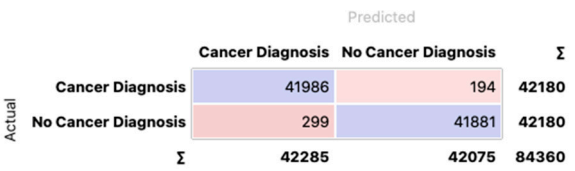
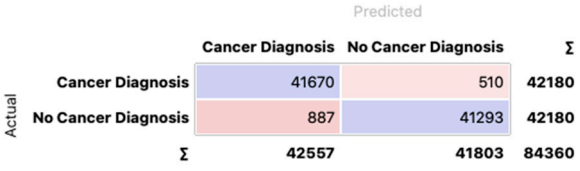
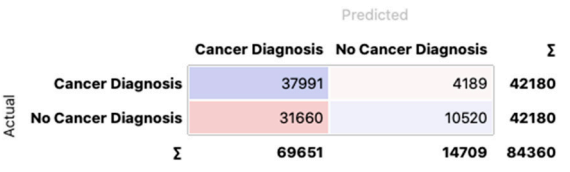
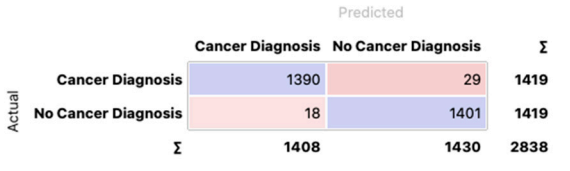
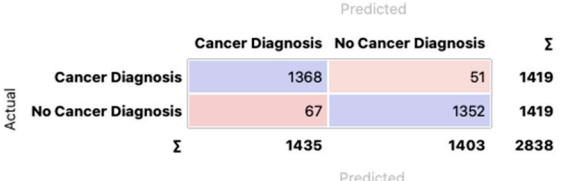
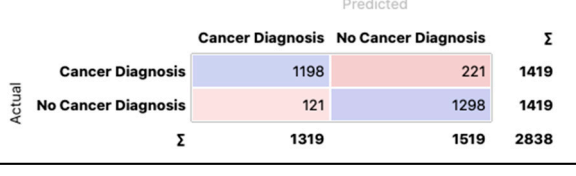
RUS + XGBoost	
RUS + ANN	
RUS + SVM	
ENN + XGBoost	
ENN + ANN	
ENN + SVM	
TL + XGBoost	
TL + ANN	
TL + SVM	

Table 11. Confusion matrix of the fusion model for datasets balanced using hybrid sampling techniques.

SMOTE-ENN + XGBoost	
SMOTE-ENN + ANN	
SMOTE-ENN + SVM	
SMOTE-TL + XGBoost	
SMOTE-TL + ANN	
SMOTE-TL + SVM	
SPIDER + XGBoost	
SPIDER + ANN	
SPIDER + SVM	

4.3. ROC Analysis

Classification results for the models are shown in a graphical format called the receiver operating characteristic (ROC) curve, which also includes total classification thresholds. The ROC curve is a visual representation of a comparison between the true-positive rate (TPR) on the *y*-axis and the false-positive rate (FPR) on the *x*-axis (FPR). The proposed methods outperformed the rest of the model in classification, as shown in the ROC plot of all the fusion models in Tables 12–14 for datasets resampled by oversampling, undersampling, and hybrid sampling techniques, respectively.

Table 12. ROC analysis of all the models based on the datasets balanced using oversampling techniques.

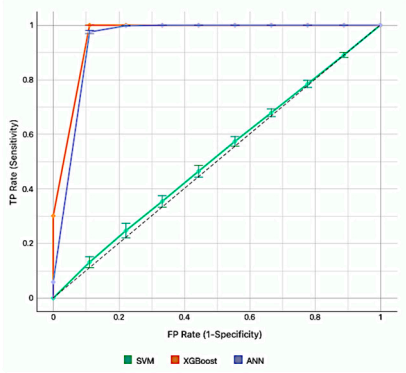
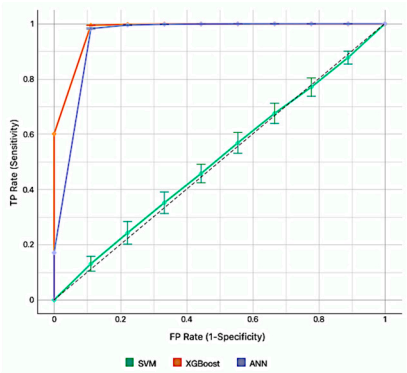
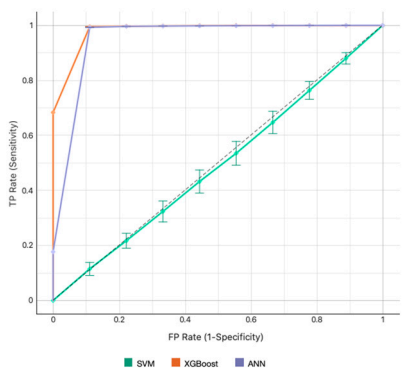
		
BCSC-ROS dataset	BCSC-SMOTE dataset	BCSC-BLSMOTE dataset

Table 13. ROC analysis of all the models based on the datasets balanced using undersampling techniques.

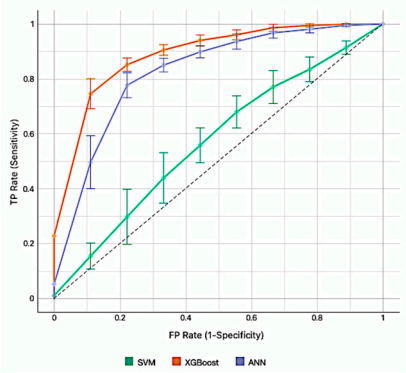
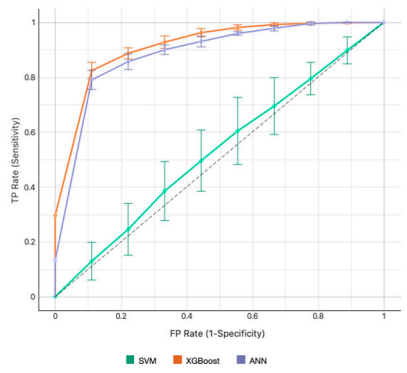
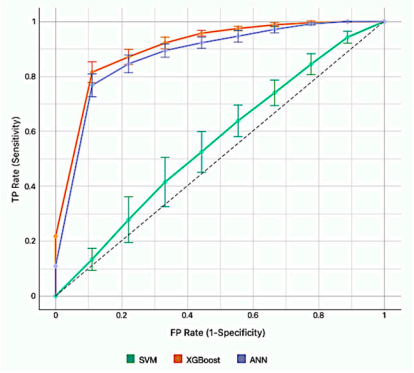
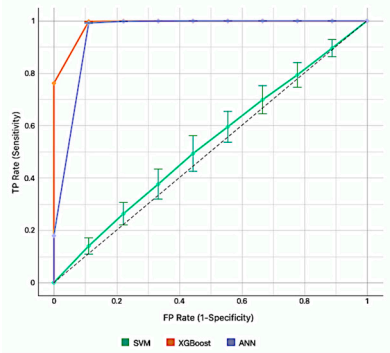
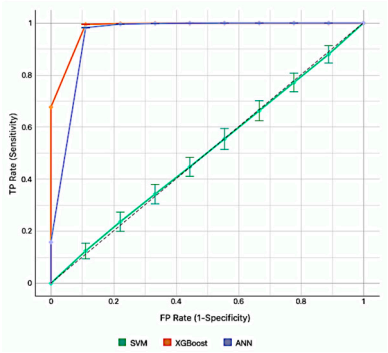
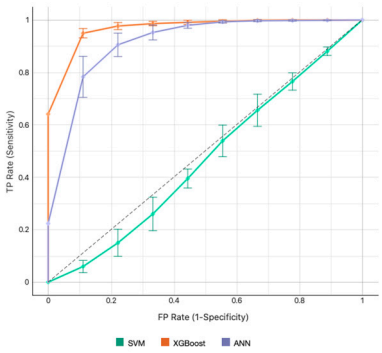
		
BCSC-RUS dataset	BCSC-ENN dataset	BCSC-TL dataset

Table 14. ROC analysis of all the models based on the datasets balanced using hybrid sampling techniques.

		
BCSC-SMOTE-ENN dataset	BCSC-SMOTE-TL dataset	BCSC-SPIDER dataset

4.4. Model Efficiency

The time complexity (s) measures the model's effectiveness during the model classification and evaluation process. Figures 9 and 10 show the training and testing time of all the fusion models, respectively. The fact that the majority of the training was performed offline was not considered during the experiment. As shown in Figure 10, RUS + SVM has the shortest training time of 0.934 s, and SMOTE + XGBoost has the longest training time of 2781.5 s. On the other hand, ROS + XGBoost has a training time of 93.85 s, which is considered moderately short and efficient. However, the testing time for all the fusion models, as shown in Figure 11, indicates that all the fusion models have relatively short time compared to their training time. RUS + ANN has the shortest testing time (0.043 s), and BLSMOTE + SVM is the longest to test (2.607 s).

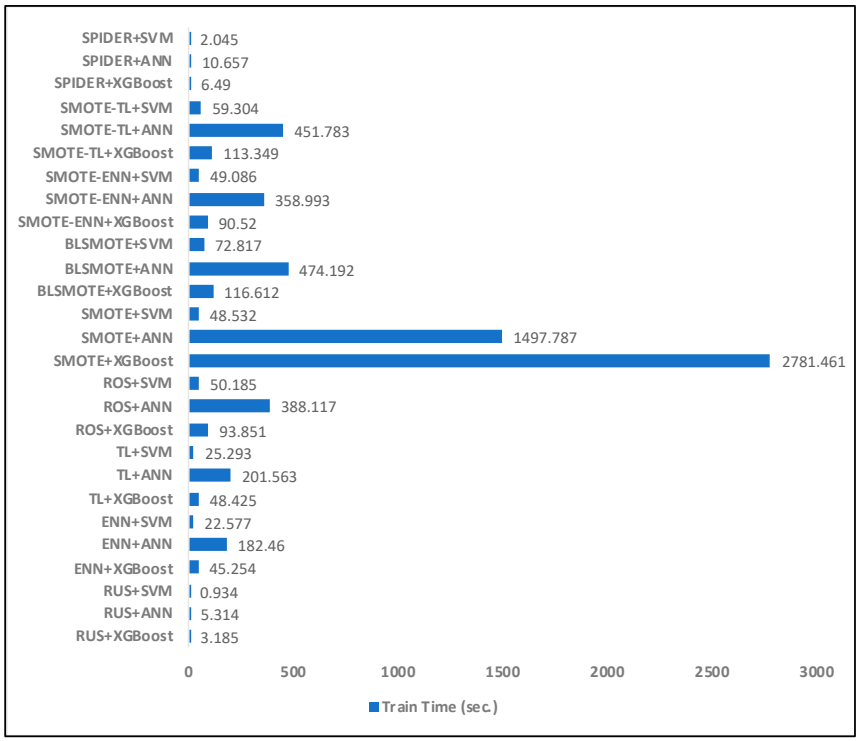


Figure 10. Training time (s) for all the fusion models.

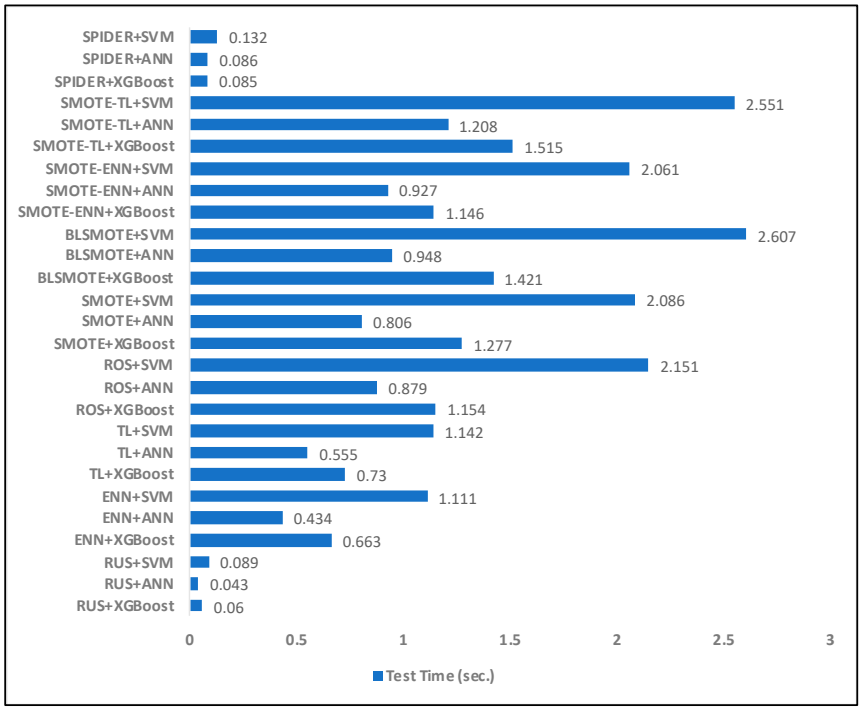
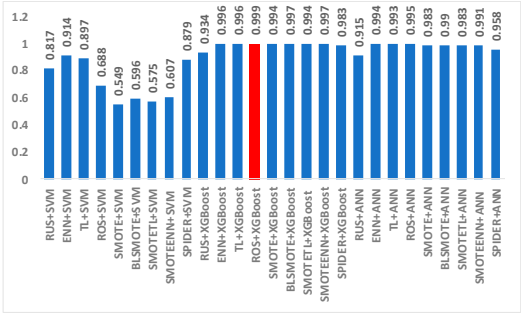


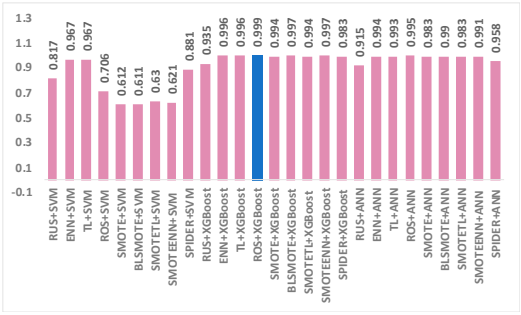
Figure 11. A testing time for all the fusion models.

4.5. Comparative Analysis between the Fusion Models

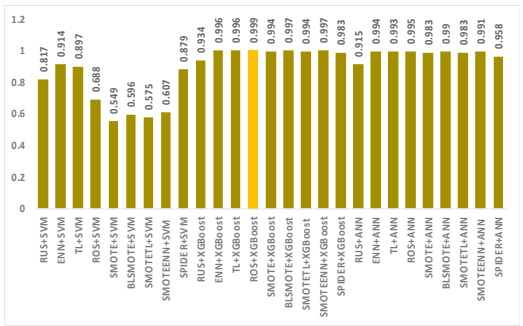
To further investigate the best fusion performance for model classification, we compared and correlated it with the rest of the model using performance measures (Acc, Pres, recall, and F1-score). The analysis revealed that the ROS + XGBoost model outperformed these models in classification measures. The SVM-based models have the least fundamental performance matrix in breast cancer detection, as presented in Figure 12.



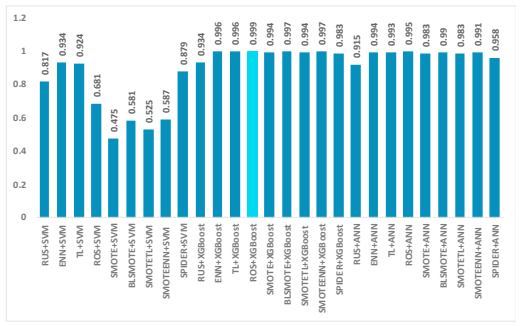
Accuracy performance of the fusion models



Precision performance of the fusion models



Recall performance of the fusion models



F1–score performance of the fusion models

Figure 12. The overall critical performance of the fusion models.

To assess the study's novelty, we compared the findings to those of previous studies using the same dataset, including those by Kabir and Ludwig [58], San Yee et al. [42], and Rajendran et al. [51]. (BCSC). Table 15 summarizes the findings. The current study's accuracy performance outperformed the rest of the previous studies.

Table 15. Comparative analysis of the current and previous studies on the BCSC dataset.

	Kabir and Ludwig [58]	**Rajendran et al. [51]**	**San Yee et al. [42]**	**The Current Study**
Year	2018	2020	2022	2023
Resampling Techniques deployed	RUS, ROS, SMOTE, ENN, SMOTE-ENN, SMOTE-TL	SMOTE SpreadSubsampling	SMOTE	RUS, ENN, TL, ROS, SMOTE, BLSMOTE, SMOTE-ENN, SMOTE-TL, SPIDER
Classifiers deployed	DT, RF, XGBoost	Bayesian network, NB, LR, SVM, MLP	LR, RF, SVM, MLP	XGBoost, ANN, SVM
Best Fusion Model	ENN + XGBoost	SMOTE + Bayesian network	SMOTE + RF	ROS + XGBoost
Accuracy	0.9149	0.9910	0.8200	0.9999

5. Limitations

Extensive hyperparameter tuning of the resampling techniques and the classifiers was very challenging due to the many experiments and the time required for each training evaluation. While batch normalization and dropout have increased robustness in hyperparameters, a more in-depth investigation, particularly optimizing different resampling techniques and classifiers, may have yielded better results for each experiment. For simplicity and fairness, we used the same setup for all experiments. The classifiers were investigated using classical, ensemble, and deep learning classifiers, representing only the single most well-known classifier. Another limitation is that only the most well-known resampling techniques with a fixed sampling rate are evaluated using a single classification algorithm. Over and above simple resampling technique selection, a more ambitious future research direction is the development of automatic methods for classifying imbalanced data. Enhanced techniques may produce different results on the same datasets.

Numerous proposed solutions to the class imbalance exist, and only a subset of the most widely used ones are discussed here. Other, more complex methods, such as SMOTE and its variants, may, on the other hand, improve performance for classical machine learning problems. More complex processes are, by definition, more difficult to implement and tune. As a result, a single technique may take some time to gain traction over more direct sampling and weighting strategies.

The data used to train the classifiers significantly impacts the accuracy of any prediction. As a result, obtaining high-quality, balanced data is crucial in model classification. In a class-imbalanced dataset, any resampling technique will face challenges in dealing with issues such as class overlap, small disjuncts, and borderline and noisy samples. Furthermore, most learning classifier systems have been reported to be inadequate in coping with the class imbalance problem.

6. Conclusions and Future Work

This paper investigated three undersampling, three oversampling, and three hybrid sampling techniques on the BCSC dataset, which has a moderate IR. According to the study, unbalanced data reduces the functional efficiency of default classifiers. As a result, techniques for preprocessing data were used to optimize algorithm functions. The best balancer and classifier for the breast cancer dataset were discovered in this study by examining the impact of class imbalance on classifier performance and comparing the functions of preprocessing techniques and classification on the dataset. An extensive empirical study was conducted in which 27 balanced datasets from a moderately class-imbalanced BCSC dataset were resampled using three oversampling, three undersampling,

and three hybrid sampling methods to optimize classification models and determine which strategy is best for improving the prediction model. Three machine-learning-based classifiers were used in the model classification: classical, ensemble, and deep learning classifiers. Based on previous work, as shown in Table 15, experiments on the same dataset produce different best models. Performance was assessed using six performance metrics tailored to the specific problem of imbalanced data. These models provide valuable patterns for determining the most appropriate resampling strategy for handling class-imbalanced datasets. Despite the extensive research, some limitations to work have been discussed in Section 5, and some other limitations will be addressed in future work.

As for future works, the current approach (assessed at the data level) could be evaluated together with the algorithm-level and the hybrid techniques to optimize the effectiveness and performance of the methods in improving the model classification on the class-imbalanced BCSC dataset. A comprehensive investigation of cost-sensitive and ensemble algorithms on various cancer datasets involving clinical and image data should be explored.

Author Contributions: Conceptualization, S.S. and S.B.; methodology, S.S.; software, S.S.; validation, S.S. and S.B.; formal analysis, S.B.; investigation, S.S.; resources, S.B.; data curation, S.S.; writing—original draft preparation, S.S.; writing—review and editing, S.B.; visualization, S.S.; supervision, S.B.; project administration, S.B.; funding acquisition, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia grant number PNU-DRI-RI-20-008 and the APC was funded by the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia.

Informed Consent Statement: Written informed consent has been obtained from the patient(s) to publish this paper.

Acknowledgments: The authors thank the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia, for funding this research through project number PNU-DRI-RI-20-008.

Conflicts of Interest: The authors declare they have no conflict of interest to report regarding the present study.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics **2020**: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef]
2. Mandelblatt, J.S.; Stout, N.K.; Schechter, C.B.; Broek, J.J.V.D.; Miglioretti, D.L.; Krapcho, M.; Trentham-Dietz, A.; Munoz, D.; Lee, S.J.; Berry, D.A.; et al. Collaborative modeling of the benefits and harms associated with different U.S. Breast cancer screening strategies. *Ann. Intern. Med.* **2016**, *164*, 215–225. [CrossRef]
3. Geller, B.M.; Bowles, E.J.A.; Sohng, H.Y.; Brenner, R.J.; Miglioretti, D.L.; Carney, P.A.; Elmore, J.G. Radiologists' Performance and Their Enjoyment of Interpreting Screening Mammograms. *AJR Am. J. Roentgenol.* **2009**, *192*, 361. [CrossRef]
4. Alqahtani, W.S.; Almufareh, N.A.; Domiaty, D.M.; Albasher, G.; Alduwish, M.A.; Alkhalaf, H.; Almuzzaini, B.; Al-Marshidy, S.S.; Alfraihi, R.; Elasbali, A.M.; et al. Epidemiology of cancer in Saudi Arabia thru 2010–2019: A systematic review with constrained meta-analysis. *AIMS Public Health* **2020**, *7*, 679. [CrossRef]
5. Kaya Keleş, M. Breast cancer prediction and detection using data mining classification algorithms: A comparative study. *Tehnicki Vjesnik* **2019**, *26*, 149–155. [CrossRef]
6. Yadavendra; Chand, M. A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method. *Mach. Vision Appl.* **2020**, *31*, 46. [CrossRef]
7. Zhang, J.; Chen, L.; Abid, F. Prediction of breast cancer from imbalance respect using cluster-based undersampling method. *J. Healthc. Eng.* **2019**, *2019*, 7294582. [CrossRef]
8. Guan, H.; Zhang, Y.; Xian, M.; Cheng, H.D.; Tang, X. SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling. *Appl. Intell.* **2021**, *51*, 1394–1409. [CrossRef]
9. Fotouhi, S.; Asadi, S.; Kattan, M.W. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J. Biomed. Inf.* **2019**, *90*, 103089. [CrossRef]
10. Ali, H.; Li, H.; Afele Retta, E.; Haq, I.U.; Guo, Z.; Han, X.; Feng, J. Representation of Differential Learning Method for Mitosis Detection. *J. Healthc. Eng.* **2021**, *2021*, 6688477. [CrossRef]
11. Jayatilake, S.M.D.A.C.; Ganegoda, G.U. Involvement of machine learning tools in healthcare decision making. *J. Healthc. Eng.* **2021**, *2021*, 6679512. [CrossRef]

12. Awan, M.J.; Mohd Rahim, M.S.; Salim, N.; Rehman, A.; Nobanee, H. Machine Learning-Based Performance Comparison to Diagnose Anterior Cruciate Ligament Tears. *J. Healthc. Eng.* **2020**, *2022*, 2550120. [CrossRef]
13. Kang, Q.; Chen, X.S.; Li, S.S.; Zhou, M.C. A noise-filtered under-sampling scheme for imbalanced classification. *IEEE Trans. Cybern.* **2016**, *47*, 4263–4274. [CrossRef]
14. Rodríguez-Torres, F.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A. An Oversampling Method for Class Imbalance Problems on Large Datasets. *Appl. Sci.* **2022**, *12*, 3424. [CrossRef]
15. Jedrzejowicz, J.; Jedrzejowicz, P. GEP-based classifier for mining imbalanced data. *Expert Syst. Appl.* **2021**, *164*, 114058. [CrossRef]
16. Zhou, F.; Gao, S.; Ni, L.; Pavlovski, M.; Dong, Q.; Obradovic, Z.; Qian, W. Dynamic self-paced sampling ensemble for highly imbalanced and class-overlapped data classification. *Data Min. Knowl. Discov.* **2022**, *36*, 1601–1622. [CrossRef]
17. Zhao, J.; Jin, J.; Chen, S.; Zhang, R.; Yu, B.; Liu, Q. A weighted hybrid ensemble method for classifying imbalanced data. *Knowl.-Based Syst.* **2020**, *203*, 106087. [CrossRef]
18. Triguero, I.; Del Río, S.; López, V.; Bacardit, J.; Benítez, J.M.; Herrera, F. ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition: An extremely imbalanced big data bioinformatics problem. *Knowl.-Based Syst.* **2015**, *87*, 69–79. [CrossRef]
19. Chen, Z.; Duan, J.; Kang, L.; Qiu, G. Class-imbalanced deep learning via a class-balanced ensemble. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 5626–5640. [CrossRef]
20. Ebenuwa, S.H.; Sharif, M.S.; Alazab, M.; Al-Nemrat, A. Variance ranking attributes selection techniques for binary classification problem in imbalance data. *IEEE Access* **2019**, *7*, 24649–24666. [CrossRef]
21. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Progr. Artif. Intell.* **2016**, *5*, 221–232. [CrossRef]
22. Xie, Y.; Qiu, M.; Zhang, H.; Peng, L.; Chen, Z. Gaussian distribution based oversampling for imbalanced data classification. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 667–679. [CrossRef]
23. Du, G.; Zhang, J.; Jiang, M.; Long, J.; Lin, Y.; Li, S.; Tan, K.C. Graph-based class-imbalance learning with label enhancement. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–15. [CrossRef]
24. Koziarski, M.; Woźniak, M.; Krawczyk, B. Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise. *Knowl.-Based Syst.* **2020**, *204*, 106223. [CrossRef]
25. Tsai, C.F.; Lin, W.C. Feature selection and ensemble learning techniques in one-class classifiers: An empirical study of two-class imbalanced datasets. *IEEE Access* **2021**, *9*, 13717–13726. [CrossRef]
26. Mishra, S.; Mallick, P.K.; Jena, L.; Chae, G.S. Optimization of skewed data using sampling-based pre-processing approach. *Front. Public Health* **2020**, *8*, 274. [CrossRef]
27. Jung, I.; Ji, J.; Cho, C. EmSM: Ensemble mixed sampling method for classifying imbalanced intrusion detection data. *Electronics* **2022**, *11*, 1346. [CrossRef]
28. Guzmán-Ponce, A.; Valdovinos, R.M.; Sánchez, J.S.; Marcial-Romero, J.R. A new under-sampling method to face class overlap and imbalance. *Appl. Sci.* **2020**, *10*, 5164. [CrossRef]
29. Alamri, M.; Ykhlef, M. Survey of Credit Card Anomaly and Fraud Detection Using Sampling Techniques. *Electronics* **2022**, *11*, 4003. [CrossRef]
30. Yang, F.; Wang, K.; Sun, L.; Zhai, M.; Song, J.; Wang, H. A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis. *BMC Med. Inf. Decis. Mak.* **2022**, *22*, 344. [CrossRef]
31. Dos Santos, R.P.; Silva, D.; Menezes, A.; Lukasewicz, S.; Dalmora, C.H.; Carvalho, O.; Giacomazzi, J.; Golin, N.; Pozza, R.; Vaz, T.A. Automated healthcare-associated infection surveillance using an artificial intelligence algorithm. *Infect. Prev. Pract.* **2021**, *3*, 100167. [CrossRef]
32. Tarawneh, A.S.; Hassanat, A.B.; Almohammadi, K.; Chetverikov, D.; Bellinger, C. Smotefuna: Synthetic minority over-sampling technique based on furthest neighbour algorithm. *IEEE Access* **2020**, *8*, 59069–59082. [CrossRef]
33. Pei, X.; Mei, F.; Gu, J. The real-time state identification of the electricity-heat system based on Borderline-SMOTE and XGBoost. *IET Cyber-Phys. Syst. Theory Appl.* **2022**, 1–11. [CrossRef]
34. Lin, E.; Chen, Q.; Qi, X. Deep reinforcement learning for imbalanced classification. *Appl. Intell.* **2020**, *50*, 2488–2502. [CrossRef]
35. De Freitas, R.C.; Naik, G.R.; Valença, M.J.S.; Bezerra, B.L.D.; de Souza, R.E.; dos Santos, W.P. Surface electromyography classification using extreme learning machines and echo state networks. *Res. Biomed. Eng.* **2022**, *38*, 477–498. [CrossRef]
36. Solanki, Y.S.; Chakrabarti, P.; Jasinski, M.; Leonowicz, Z.; Bolshev, V.; Vinogradov, A.; Jasinska, E.; Gono, R.; Nami, M. A hybrid supervised machine learning classifier system for breast cancer prognosis using feature selection and data imbalance handling approaches. *Electronics* **2021**, *10*, 699. [CrossRef]
37. Kraiem, M.S.; Sánchez-Hernández, F.; Moreno-García, M.N. Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. an approach based on association models. *Appl. Sci.* **2021**, *11*, 8546. [CrossRef]
38. Rasool, A.; Bunterngchit, C.; Tiejian, L.; Islam, M.R.; Qu, Q.; Jiang, Q. Improved machine learning-based predictive models for breast cancer diagnosis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3211. [CrossRef]
39. Jadhav, A.; Mostafa, S.M.; Elmannai, H.; Karim, F.K. An Empirical Assessment of Performance of Data Balancing Techniques in Classification Task. *Appl. Sci.* **2022**, *12*, 3928. [CrossRef]
40. Rendon, E.; Alejo, R.; Castorena, C.; Isidro-Ortega, F.J.; Granda-Gutierrez, E.E. Data sampling methods to deal with the big data multi-class imbalance problem. *Appl. Sci.* **2020**, *10*, 1276. [CrossRef]

41. Tasci, E.; Zhuge, Y.; Camphausen, K.; Krauze, A.V. Bias and Class Imbalance in Oncologic Data—Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets. *Cancers* **2022**, *14*, 2897. [CrossRef]
42. San Yee, W.; Ng, H.; Yap, T.T.V.; Goh, V.T.; Ng, K.H.; Cher, D.T. An Evaluation Study on the Predictive Models of Breast Cancer Risk Factor Classification. *J. Logist. Inform. Serv. Sci.* **2022**, *9*, 129–145. [CrossRef]
43. Huang, M.W.; Chiu, C.H.; Tsai, C.F.; Lin, W.C. On combining feature selection and over-sampling techniques for breast cancer prediction. *Appl. Sci.* **2021**, *11*, 6574. [CrossRef]
44. Vuttipittayamongkol, P.; Elyan, E. Neighbourhood-based undersampling approach for handling imbalanced and over-lapped data. *Inf. Sci.* **2020**, *509*, 47–70. [CrossRef]
45. Wang, Z.; Cao, C.; Zhu, Y. Entropy and Confidence-Based Undersampling Boosting Random Forests for Imbalanced Problems. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 5178–5191. [CrossRef]
46. Al-Shamaa, Z.Z.R.; Kurnaz, S.; Duru, A.D.; Peppa, N.; Mirnezami, A.H.; Hamady, Z.Z.R. The Use of Hellinger Distance Undersampling Model to Improve the Classification of Disease Class in Imbalanced Medical Datasets. *Appl. Bionics Biomech.* **2020**, *2020*, 8824625. [CrossRef]
47. Desuky, A.S.; Hussain, S. An Improved Hybrid Approach for Handling Class Imbalance Problem. *Arab. J. Sci. Eng.* **2021**, *46*, 3853–3864. [CrossRef]
48. Zhang, J.; Chen, L.; Tian, J.X.; Abid, F.; Yang, W.; Tang, X.F. Breast cancer diagnosis using cluster-based undersampling and boosted C5. 0 algorithm. *Int. J. Control Autom. Syst.* **2021**, *19*, 1998–2008. [CrossRef]
49. Koziarski, M. Radial-based undersampling for imbalanced data classification. *Pattern Recognit.* **2020**, *102*, 107262. [CrossRef]
50. Zhang, J.; Chen, L. Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. *Comput. Assist. Surg.* **2019**, *24*, 62–72. [CrossRef]
51. Rajendran, K.; Jayabalan, M.; Thiruchelvam, V. Predicting breast cancer via supervised machine learning methods on class imbalanced data. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 54–63. [CrossRef]
52. Tran, T.; Le, U.; Shi, Y. An effective up-sampling approach for breast cancer prediction with imbalanced data: A machine learning model-based comparative analysis. *PLoS ONE* **2022**, *17*, e0269135. [CrossRef]
53. IBRAHIM, M.H. A Salp Swarm-Based Under-Sampling Approach for Medical Imbalanced Data Classification. *Avrupa Bilim ve Teknoloji Dergisi* **2022**, *34*, 396–402. [CrossRef]
54. Kumar, V.; Lalotra, G.S.; Sasikala, P.; Rajput, D.S.; Kaluri, R.; Lakshmanna, K.; Shorfuzzaman, M.; Alsufyani, A.; Uddin, M. Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques. *Healthcare* **2022**, *10*, 1293. [CrossRef]
55. Mahesh, T.R.; Vinoth Kumar, V.; Muthukumaran, V.; Shashikala, H.K.; Swapna, B.; Guluwadi, S. Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast. *Cancer* **2022**, *2022*, 4649510. [CrossRef]
56. Elyan, E.; Moreno-Garcia, C.F.; Jayne, C. CDSMOTE: Class decomposition and synthetic minority class oversampling tech-nique for imbalanced-data classification. *Neural Comput. Appl.* **2020**, *33*, 2839–2851. [CrossRef]
57. Werner de Vargas, V.; Schneider Aranda, J.A.; dos Santos Costa, R.; da Silva Pereira, P.R.; Victória Barbosa, J.L. Imbalanced data pre-processing techniques for machine learning: A systematic mapping study. *Knowl Inf Syst* **2023**, *65*, 31–57. [CrossRef]
58. Kabir, M.F.; Ludwig, S. Classification of breast cancer risk factors using several resampling approaches. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018. [CrossRef]