A Comparative Analysis of Implicit Augmentation Techniques for Breast Cancer Diagnosis Using Multiple Views

Yumnah Hasan Talhat khan Darian Reyes Fernández de Bulnes Juan F H Albarracín
Conor Ryan
University of Limerick
{yumnah.hasan, khan.talhat, darian.reyesfernandezdebulnes, juan.albarracin, conor.ryan} @ul.ie

Abstract

The Design of effective deep-learning methods for medical image analysis represents a great challenge given the scarcity of balanced datasets, leading to biased results and overfitting. Data augmentation mitigates these limitations due to its effectiveness in increasing the diversity and quantity of training data, but the selection of an appropriate augmentation method strongly depends on the problem domain. In this study, we investigate the effects of various feature-level augmentation methods on the performance of Deep-Learning-based Breast Cancer (BC) diagnosis using mammographic images of Craniocaudal (CC) and Mediolateral Oblique (MLO) views. Through quantitative performance evaluations, we systematically assess the impact of augmentation techniques on classification using two feature extraction techniques, namely, Haralick features and deep GoogleNET features. Our experiments, conducted on the Digital Database for Screening Mammography (DDSM) and the Wisconsin Breast Cancer (WBC) datasets, reveal that Mixup, when combined with STEM, outstands as the most promising in a wide range of scenarios.

1. Introduction

Breast Cancer (BC) is the leading cause of mortality in women worldwide. According to the World Health Organization, 2.3 million new reported cases were recorded in 2020, and 685,000 deaths occurred worldwide [33]. By the end of that year, there were 7.8 million cases of women with this deadly disease. The mortality rate due to BC is dominant in 12 out of 20 regions of the world as described by Ferlay et al. [12] and, due to the increased ageing of the general population, the BC incidence is steadily rising. However, detection at the initial stages significantly improves treatment outcomes, underscoring the critical importance of timely screening and diagnosis initiatives [10].

Modern Machine Learning (ML) techniques, in particular, those that rely on Deep Neural Networks (DNN), constitute the state of the art in automatic medical diagnosis from images [4]. For these methods, the quality and quantity of the training data are key factors influencing the performance and robustness of models. However, in the medical diagnosis domain, the datasets are often imbalanced, so certain classes or categories are underrepresented compared to others. This class imbalance poses significant challenges for DL-based classification algorithms [8], turning them prone to biased predictions and reduced accuracy [46].

Data augmentation is the most widely-used method to improve the class distribution of an unbalanced dataset. Significant advancements in this field have been achieved in leveraging deep learning (DL) models for the tasks of image classification and segmentation [27]. These techniques have been used extensively in medical diagnosis applications, particularly in tasks involving predicting and segmenting cancerous masses [28].

Implicit augmentation approaches, i.e., those performed on the feature vectors, instead of the data itself, have emerged as valuable tools to address this issue [42]. These approaches include over-sampling (to generate minority-class samples), under-sampling (to drop samples from the majority class), and hybrid sampling (combination of both).

This work focuses only on implicit augmentation approaches, and conducts a comparative analysis of nine methods devised to address the issue of class imbalance in the domain of BC diagnosis. We assess the robustness of these techniques based on two distinct feature sets: deep GoogleNET features [32] and handcrafted Haralick features [19] over different mammogram perspectives, namely Craniocaudal (CC) and Mediolateral Oblique (MLO). The augmented features are used to train two Neural Network (NN) based classifiers, along with statistical significance testing. As a result, we offer recommendations for the optimal combination of features and augmentation methods to enhance the accuracy of BC diagnosis, providing valuable insights for improving detection and treatment strategies, and then filling a gap in the literature regarding how re-

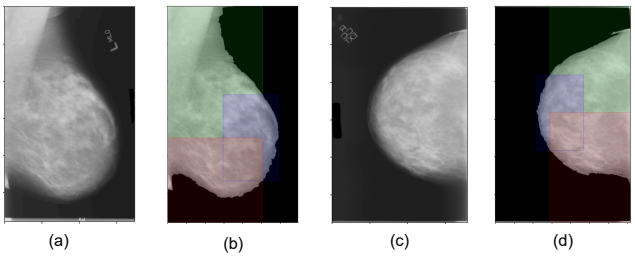

Figure 1. (a) Left MLO View (b) Segmentation of Left MLO image (c) Right CC View (d) Segmentation of Right CC image.

sponsive these augmentation techniques are to certain NN models, views, and feature extraction methods.

The structure of this paper is as follows: Section 2 provides the background and a review of the existing literature, Section 3 delves into the applied methods, and Section 4 outlines the experimentation process along with the results and statistical tests. Finally, Section 5 provides our insights on the obtained results, and Section 6 presents the closing remarks and future work.

2. Background and Related Work

Data augmentation involves generating synthetic samples by adding or altering features within the original dataset to increase diversity [31]. Its primary objective is to mitigate potential inaccuracies when developing diagnostic models with limited or imbalanced data.

The problem of class imbalance is typically addressed through two main approaches [2]: i) algorithmic, by modifying the learning process of classifiers (e.g., employing cost-sensitive learning algorithms) to prioritize the minority group; and ii) data-level, by resampling the data space to balance the class distribution (e.g., oversampling minority classes or undersampling majority ones). Data-level approaches are more commonly used because they are independent of the classifiers [14].

While traditional approaches to data augmentation involve explicit transformations applied directly to the input data (e.g., operations such as cropping, flipping, or), techniques that over-sample in the feature space [15] are widely used alternatives. Implicit augmentation operates within the feature space, leveraging the inherent structure and relationships among features to generate augmented instances [34].

The Synthetic Minority Over Sampling Technique (SMOTE) [9] is a data-level method used to increase the minority class samples with synthetic data. For example, for diabetes prediction, 99.64% accuracy is noted when SMOTE is combined with deep Long Short-Term Memory (LSTM) classifier [5]. Research [25] integrating feature selection and oversampling through SMOTE recently demonstrated the synergy of Information Gain (IG) and SMOTE, achieving the highest Area Under the Curve (AUC) score of 0.788 as compared to the other combinations of Genetic Algorithm (GA) with SMOTE, and IG. This outcome was demonstrated using the KDD Cup 2008 Breast Cancer Dataset. Additionally, SMOTE attains the highest AUC of 0.962 when applied to the Wisconsin Breast Cancer (WBC) dataset using the Support Vector Machine (SVM) classifier. Another oversampling approach, Adaptive Synthetic Oversampling (ADASYN) [22] method, generates minority samples based on the local distribution within the dataset. A recent study employed the ADASYN to address class imbalance when analyzing the INbreast dataset [30]. This study used five different feature extraction types and the ReliefF algorithm for feature selection. Remarkably, the findings revealed an impressive accuracy rate of 99.5% with their proposed methodology. The literature also shows promising results in handling class imbalance with other oversampling methods, including Borderline SMOTE (BSMOTE) [17] and SVM-SMOTE (SVM-S) [37] [1] [6].

Hybrid data resampling methods consist of both oversampling and undersampling techniques. By combining the strengths of different resampling methods, the model's generalization and presentation of the minority class can be improved. For instance, the work by Kabir and Ludwig [29] reports maximum recall and F1 scores of 0.87 and 0.43, respectively, for the minority class, when SMOTE Edited Nearest Neighbour (S-ENN) [44] is applied and XGBoost is used for classification. That work used the Breast Cancer Surveillance Consortium (BCSC) database, which contains a high class imbalance. Furthermore, the classification performance of breast instances is compared where SMOTE Tomek (S-Tomek) [7] combined with correlation feature selection using SVM classifier outperforms when used with Naive Bayes by securing an accuracy of 96.80% [39].

Previous work shows promising results on class-ratio balance with data augmentation. However, selecting a suitable augmentation method depends tightly on the problem domain. Therefore, we present a comparative analysis of augmentation methods on the features of mammographic images to identify the combination of features and augmentation techniques that provide higher true predictions.

3. Methodology

The proposed workflow analyses the effects of augmentation approaches using different data variants based on their perspectives. Two datasets, Digital Database for Screening Mammography (DDSM) [23] and WBC dataset [45], are employed in this study. For the DDSM dataset, images of CC and MLO views undergo preprocessing and segmentation into three segments, and then Haralick and deep features are extracted. Dataset configurations are then created based on the S_{CC}, S_{MLO}, and S_{CC+MLO} views. In contrast, only one setup of WBC, containing pre-extracted fea-

tures, is utilized for evaluation. Nine augmentation methods are applied to the training set described in Section 3.2, while the test set remains unaltered. Finally, breast instances are classified using NN classifiers as described in Section 4, including Multilayer Perceptron (MLP) and 1 Dimensional Convolutional Neural Network (1D-CNN). The complete pipeline is shown in Figure 2.

3.1. Dataset Details

The DDSM data set consists of 43 volumes of mammographic images. Each volume contains two case types: normal cases belong to patients without cancer, while cancer cases come from exams detecting at least one proven cancer. In this study, to construct a more imbalanced and thus realistic data set, we employ the **Cancer 02** set for positive cases and three negative sets, **Normal 01-03**.

The number of images used in this work is 197 abnormal and 3026 normal, resulting in a positive-to-negative class ratio of approximately 6:94. This breakdown closely reflects real-life scenarios in medical imaging, where the prevalence of abnormal cases is typically much lower compared to normal cases. Each image is segmented into three regions using the approach described in [41]. The upper segment refers to the top part, the central part of the breast comes in the middle segment, and the lower segment includes the bottom part of the breast as depicted in Figure 1 for each CC and MLO views.

We preprocess the segmented images with the median filter to smooth out irregular pixel values, for cleaner and more uniform examples. Moreover, subsequent Otsu thresholding for noise reduction and artefact removal are applied. Afterwards, Haralick features are extracted using Grey Level Co-occurrence Matrix (GLCM) with four orientations of 0°, 45°, 90°, and 135° using the method described in [21]. A total of 52 features per segment are extracted for each image. The deep features are extracted using the pretrained GoogleNet architecture [32]. This results in a 1023-dimensional feature vector per segmented region. Each set of features can form a different dataset on which the comparative analysis will be performed independently.

For each dataset, we further create three setups: i) S_{CC}, containing only segmented images from the CC view; ii) S_{MLO}, with only segmented images from the MLO view, and iii) S_{CC+MLO} combining both views.

The WBC dataset contains pre-extracted Fine Needle Aspiration (FNA) features from breast masses. A total of 30 features are present for malignant and benign categories. There are 212 malignant samples and 357 benign samples provided. The ratio between positive and negative classes is noted as 37:63. Integrating the WBC dataset enriches our analysis by providing clinically relevant data for evaluating augmentation methods and enhances the study's applicability to real-world scenarios. The details of positive and negative samples used for training and testing the DDSM and WBC datasets are shown in Table 1.

Table 1. The DDSM dataset is organized into three distinct setups based on two mammogram views: CC and MLO. The WBC dataset comprises a solitary setup. "Tr Pos" and "Tr Neg" denote training positive and negative samples, respectively. The positive test samples are designated as "Ts Pos", while "Ts Neg" is used for the negative test instances.

Setups	Tr Pos	Tr Neg	Ts Pos	Ts Neg
S_{CC}	98	1216	19	308
S_{MLO}	99	1204	20	298
S_{CC+MLO}	158	2420	39	606
WBC	170	286	42	71

3.2. Augmentation Methods

In this work, we compare nine different implicit augmentation methods:

1. **SMOTE** [9]: For each minority sample denoted as $\mathbf{M}$, another minority sample, denoted as $\mathbf{m}$, is selected at random from their k nearest neighbours. Next, a random point is selected between $\mathbf{m}$ and $\mathbf{M}$. This process yields a newly created sample named $(\mathbf{S_{new}})$, which is added to the dataset. The balance parameter N regulates this generation, with $N = 1$ indicating an equal representation of both minority and majority [2].
2. **BSMOTE** [17]: This is an adaptation of the original SMOTE algorithm. It first categorizes the sample into noise and border classes. The data point whose k nearest neighbours belong to the majority class is considered noise. However, if at least half of the k nearest neighbour samples are from the minority class, then the data point is classified as a border instance. After identifying the border samples, BSMOTE exclusively generates samples for the border instances, focusing efforts on areas critical for improving classification accuracy.
3. **ADASYN** [22]: Inspired by the principles of BSMOTE, ADASYN generates varying numbers of synthetic examples for the minority class based on its distribution. It determines the number of synthetic examples for each minority example by considering the number of its nearest neighbours from the majority class. Specifically, the greater the number of majority nearest neighbours, the more synthetic examples are generated [26].
4. **SVM-S** [37]: This method uses the SVM algorithm to locate the support vector samples within the minority class, utilizing them as a reference for generating synthetic instances. These support vector samples are noise-free attributes as they reside closest to the boundary, restricting the majority and minority classes [35].
5. **S-ENN** [44]: This combines SMOTE and ENN in a two-step process. First, it leverages SMOTE to create

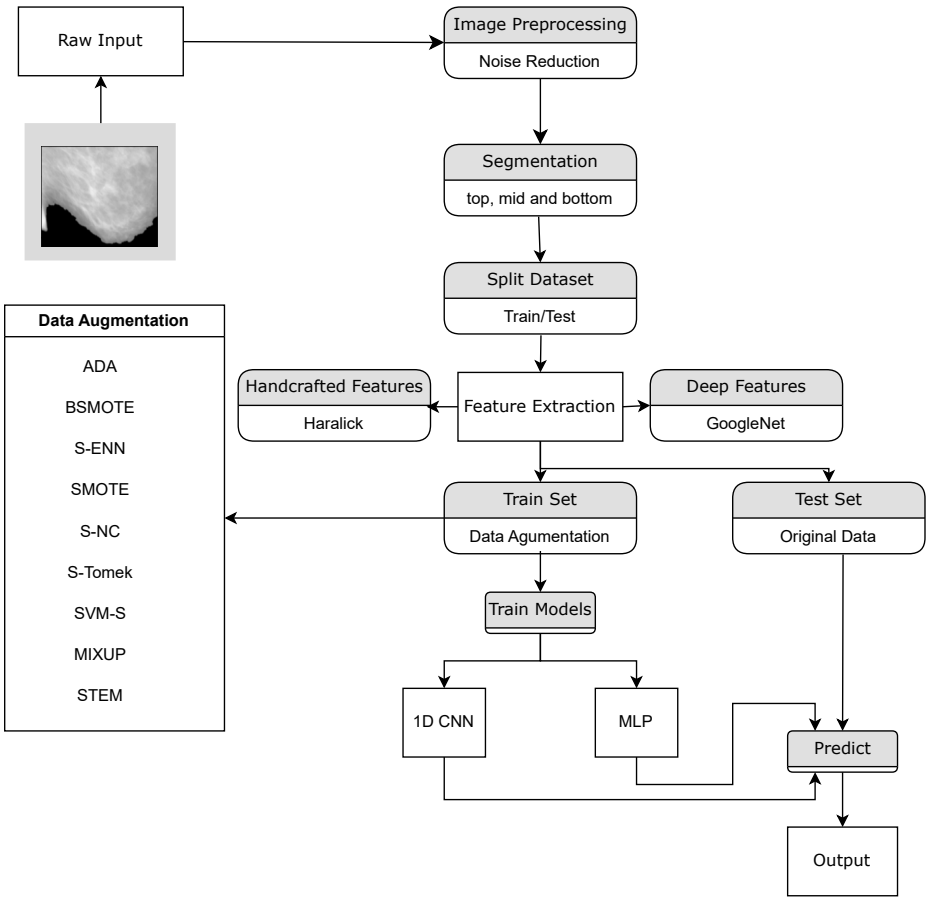

Figure 2. The workflow illustrating our comparative analysis of data-level augmentation approaches for Breast Cancer classification.

synthetic samples before employing ENN to cleanse the dataset by eliminating noisy instances. ENN eliminates the majority class instances that differ from their k nearest neighbours, facilitating smoother decision boundaries, resulting in a refined and more reliable dataset for subsequent analysis [48].

6. **S-Tomek** [7]: The primary aim of this method is to reduce the overlapping data points within each class's sample space. Following SMOTE oversampling, clusters of different classes might overlap upon each other's space. Tomek links come into play, identifying pairs of instances from distinct classes close to each other. By systematically eliminating these instances from both classes, the technique yields a balanced dataset characterized by distinct and well-defined class clusters [43].
7. **Mixup** [47]: This operates by combining two input samples, I_x and I_y, along with their corresponding labels, O_x and O_y, to create a new sample I_{new} and label O_{new} using linear interpolation [18]. This is achieved as shown below:

$$I_{\text{new}} = \lambda I_x + (1 - \lambda) I_y \tag{1}$$

$$O_{\text{new}} = \lambda O_x + (1 - \lambda) O_j \tag{2}$$

Here, λ is a randomly sampled mixing coefficient from a beta distribution. The interpolated sample I_{new} lies along the line connecting I_x and I_y in the feature space. When applied to tabular data, Mixup involves linearly interpolating feature values between pairs of rows in the dataset, along with their corresponding labels.

8. **STEM** [20]: This hybrid method operates by employing S-ENN on minority and majority-class samples, leveraging the overall distribution of both classes to alleviate both between-class and within-class imbalances. By integrating Mixup, STEM ensures a balanced generation of sample points.
9. **STEM/Mixup**: In addition to the methods described above. We also explore the combination of STEM and Mixup in this work. In STEM/Mixup, half of the samples generated by STEM are selected, alongside half of the samples produced by Mixup, and merged into a single dataset. By merging samples from both STEM and Mixup, this approach creates a balanced combination of synthetic samples. This balance aims to leverage the

strengths of each approach: the diversity and balance provided by STEM and the robustness and generalization capabilities offered by Mixup.

3.3. Classifiers

The augmented samples were used as input to two classifiers: a One-dimensional Convolutional Neural Network (1D-CNN) and a Multilayer Perceptron (MLP).

For MLP, we employed a sequential model structure consisting of the input layer accepted data with dimensions corresponding to the shape of the training dataset, ensuring compatibility with the input data format. Subsequent dense layers were added to the model, each with 64 and 32 units and Rectified Linear Unit (ReLU) activation functions. Dropout regularization with a rate of 0.5 was applied after each dense layer to mitigate overfitting by randomly dropping a fraction of the units during training. Finally, the output layer consisted of a single unit with a sigmoid activation function, suitable for binary classification tasks, providing probability predictions for the positive class.

1D-CNNs is a type of CNN designed to process one-dimensional data. They process convolutional operations to extract local features from input sequences, capturing patterns across neighbouring elements. The structure of 1D-CNNs consists of an input layer, convolutional layers, pooling layers, fully connected layers, and an output layer. The input to a 1D-CNN is a one-dimensional sequence fed to the network through the input layer [40].

The architecture of the 1D-CNN consists of an input layer that accepts one-dimensional sequences, which are then processed through convolutional layers with 128 filters and a kernel size of 3. Max-pooling layers were added to down-sample the feature maps, reducing the spatial dimensions while preserving the most important features. The output of the pooling layers was then flattened and passed through fully connected layers with 64 units and ReLU activation functions. Dropout regularization with rates of 0.5 and 0.2 was applied after the fully connected layers to prevent overfitting. Additionally, batch normalization was employed to standardize the inputs to the network, accelerating the training process and improving model performance. The output layer utilized a sigmoid activation function, providing probability predictions for binary classification.

4. Experimental Setup

All experiments involving 1D-CNN and MLP models were conducted utilizing the scikit-learn library [38] and TensorFlow framework [3]. To ensure comprehensive statistical analysis, the AutoRank library [24] was employed to evaluate the performance of implemented approaches.

Both models were trained for five epochs and 30 runs with a batch size of 64. The Adam optimizer was used with a learning rate 0.0001 and binary cross-entropy loss. Early stopping was employed to prevent overfitting.

The experiments were conducted on the DDSM and WBC datasets described in Section 3.1. Before training the models, the dataset was split into training and testing sets using an 80:20 ratio. We do not require a validation set, as we are performing a comparative analysis. Additionally, standardization was applied to normalize the data.

4.1. Results

We use AUC as the performance metric. AUC is a valuable evaluation criterion for binary classifiers because it represents the likelihood of ranking positive predictions higher than negative ones at random [16]. Moreover, it is a comprehensive evaluation, encapsulating sensitivity and specificity across a wide range of potential threshold values [11].

The initial experiments were conducted using eight augmentation approaches, and the results, as presented in Table 3, show that Mixup and STEM are the best-performing strategies. Detailed analysis of the table findings is provided in Section 4.2. Consequently, another round of experiments was undertaken, incorporating a combined version of both techniques. The results reported in Table 2 include the STEM/Mixup approach and the other eight methods.

In Table 2, the three setups of DDSM datasets are categorized under deep and Haralick feature sets. When using deep features and the S_{CC} setup, the highest AUCs for both 1D-CNN and MLP were achieved using the STEM/Mixup strategy, with scores of 0.852 and 0.888, respectively. When using Haralick features on that setup, the best-performing augmentation strategy for both classifiers is still STEM/Mixup, although it disimproves for 1D-CNN to 0.741, while it improves to 0.929 for MLP.

For S_{MLO}, the best results for 1D-CNN are, once again, with the STEM/Mixup strategy, scoring 0.942 AUC, while the best-performing strategy for MLP was ADASYN , with 0.983, which is tied with the best score of all methods. Mixup performs best for 1D-CNN when using the Haralick feature set, scoring 0.808, while STEM/Mixup is again the best strategy for MLP, scoring 0.888.

When using the setup with both CC and MLO views, i.e.. S_{CC+MLO}, the highest AUC for 1D-CNN with deep features was, once again, STEM/Mixup, with 0.942, the best result for 1D-CNN across all setups and augmentation strategies. STEM was the best-performing strategy for MLP, with a result of 0.923. Both classifiers disimproved when using Haralick features; the best score for 1D-CNN was 0.808, with Mixup, while for MLP, the best score was 0.888 with STEM/Mixup.

The highest AUC achieved by 1D-CNN on the WBC dataset was 0.719, using STEM, while both STEM and STEM/Mixup scored 0.983 with MLP.

In Figure 3, we compare the performance of the 1D-

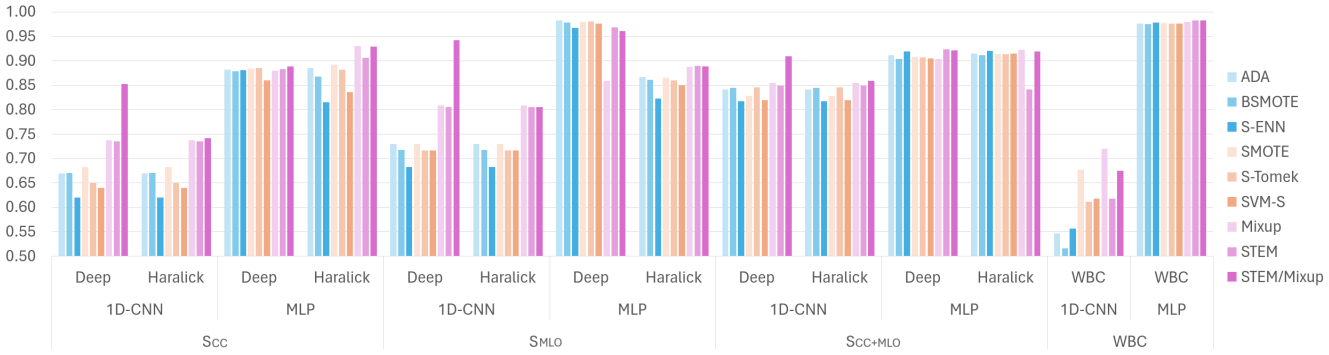

Figure 3. AUC of 1D-CNN and MLP classifiers using Haralick and deep features with different balancing techniques for all the data setups. The x-axis contains the data setups categorized into deep and Haralick features for 1D-CNN and MLP classifiers. The y-axis contains the obtained AUC by each classifier. The legend displayed the nice augmentation methods used in this work.

Table 2. The AUCs using deep and Haralick features for each DDSM data setup are compared using the different augmentation methods. Moreover, the WBC dataset contains pre-extracted features. Both 1D-CNN and MLP classifiers are presented. The highest AUCs achieved by employing the augmentation method for each classifier are emphasized in the following results.

Features	Deep						Haralick						Pre-extracted	
Augmentation	S_{CC}		S_{MLO}		S_{CC+MLO}		S_{CC}		S_{MLO}		S_{CC+MLO}		WBC	
Methods	1D-CNN	MLP	1D-CNN	MLP	1D-CNN	MLP	1D-CNN	MLP	1D-CNN	MLP	1D-CNN	MLP	1D-CNN	MLP
ADASYN	0.669	0.882	0.729	**0.983**	0.841	0.911	0.669	0.884	0.729	0.866	0.841	0.914	0.547	0.977
BSMOTE	0.670	0.878	0.717	0.978	0.845	0.904	0.670	0.867	0.717	0.861	0.845	0.912	0.516	0.975
S-ENN	0.620	0.881	0.682	0.967	0.817	0.919	0.737	0.815	0.682	0.822	0.817	0.920	0.576	0.978
SMOTE	0.682	0.884	0.729	0.979	0.828	0.908	0.620	0.892	0.729	0.865	0.828	0.914	0.556	0.976
S-tomek	0.649	0.885	0.716	0.981	0.845	0.907	0.649	0.881	0.716	0.859	0.845	0.914	0.677	0.976
SVM-S	0.640	0.859	0.716	0.976	0.819	0.905	0.682	0.905	0.716	0.850	0.819	0.908	0.611	0.975
Mixup	0.737	0.880	0.808	0.858	0.854	0.903	0.735	**0.929**	**0.808**	0.887	0.854	0.914	0.618	0.979
STEM	0.735	0.883	0.805	0.968	0.849	**0.923**	0.640	0.835	0.805	0.889	0.849	**0.922**	**0.719**	**0.983**
STEM/Mixup	**0.852**	**0.888**	**0.942**	0.961	**0.909**	0.921	**0.741**	**0.929**	0.805	**0.888**	**0.858**	0.919	0.675	**0.983**

CNN and MLP classifiers. Across all five data setups, the MLP classifier consistently performs better than the 1D-CNN classifier. The Haralick and deep feature sets are individually analyzed using nine augmentation setups; in each case, the MLP classifier consistently outperforms the 1D-CNN classifier by a distinct margin.

4.2. Statistical Tests

The non-parametric Friedman test [13] was performed on the mean values. AUCs for all 30 runs are used for both MLP and 1D-CNN classifiers. After rejecting the null hypothesis, a comprehensive understanding of the specific intergroup differentiation was identified using the Nemenyi post hoc test [36]. The significance level is set at $\alpha = 0.05$. An in-depth examination of these results shows, as depicted in Figure 4 (a), that the S_{CC+MLO} setup with deep features was the highest ranking approach for the 1D-CNN classifier. Deep features also performed best for 1D-CNN for the other two setups, achieving a higher rank than Haralick features in all cases. However, the horizontal bar connecting the two approaches shows no statistical difference.

Figure 4 (b) shows that the best result for MLP occurred when using the MLO view with deep features. This result is significantly better than the other views. The second position is shared between deep and Haralick features, both using the S_{CC+MLO} setup, and there is no significant difference between them. For S_{CC}, Haralick performed significantly better than deep features.

The results in Figure 4 were produced only using the DDSM dataset, as WBC contains pre-extracted features that cannot be directly compared to deep or Haralick features.

Another configuration was designed for the Friedman-Nemenyi post hoc test to gain more insights into the augmentation approaches. The augmentation methods for each feature type, i.e. deep and Haralick, are compared using the 1D-CNN and MLP classifiers separately. Table 3 shows that Mixup was the best-performing strategy six times for the 1D-CNN classifier and twice for the MLP; the test shows that this is a statistically significant result with 95% confidence level. The WBC dataset containing pre-extracted features was the only setup where it wasn't the best performer. Furthermore, STEM outperforms other methods in the WBC setup for both classifiers. Additionally, STEM exhibits promising performance in the $S_{MLO} - H$ and $S_{CC+MLO} - D$ scenarios when used with the MLP classifier. In the $S_{CC} - D$ scenario, S-Tomek secures the highest

Table 3. The post hoc Nemenyi test reveals the best-performing augmentation approaches for the DDSM data setups using both 1D-CNN and MLP classifiers. Mixup secured the first rank.

Setups	1D-CNN	MLP
$S_{CC} - D$	Mixup	S-Tomek
$S_{CC} - H$	Mixup	Mixup
$S_{MLO} - D$	Mixup	ADASYN
$S_{MLO} - H$	Mixup	STEM
$S_{CC+MLO} - D$	Mixup	STEM
$S_{CC+MLO} - H$	Mixup	Mixup
WBC	STEM	STEM

rank when samples are classified using MLP. Conversely, ADASYN is the optimal choice for the S_{MLO} setup when employed with the MLP classifier.

The results summarized in Table 3 indicate that Mixup and STEM are the two most frequently occurring approaches that outperformed the others. Therefore, STEM/Mixup is introduced as described in Section 3.2. The results obtained from the STEM/Mixup included in this test are shown in Table 4. It can be concluded that STEM/Mixup achieved the highest rank in all setups except for WBC and $S_{MLO} - H$ when used with the 1D-CNN classifier. Additionally, it demonstrates the best performance for the setups $S_{CC} - D$ and WBC when utilized with the MLP classifier.

Moreover, Mixup and STEM emerge as the second most promising approaches. STEM achieved the highest rank in the WBC setup when utilizing the 1D-CNN classifier, alongside $S_{MLO} - H$ and $S_{CC+MLO} - D$ when classified with the MLP classifier. Mixup demonstrates superior performance when applied to samples from setups $S_{MLO} - H$ and classified using the 1D-CNN classifier. Additionally, Mixup excels as the top performer for the $S_{CC} - H$ and $S_{CC+MLO} - H$ setups using the MLP classifier. Lastly, ADASYN emerges as the best-performing approach for the $S_{MLO} - D$ setup when combined with the MLP classifier. The nemenyi plots of the Tables 3, 4 and detailed access to our code are presented in our GitHub repository[1].

5. Discussion

As pointed out in Section 4.2, we performed two sets of statistical tests to determine, in first place, the most successful setup (i.e., combination of views and feature extractor) for each classifier and, in second place, which augmentation technique excels in each setup. These tests, along with the results presented in Table 2 and Figure 3 reveal interesting patterns in the interaction between classifiers, views, feature extractors and, more importantly, augmentation methods.

The most prominent pattern is the clear superiority of

[1]https://github.com/yumnah3/-Comparative-Analysis-of-Implicit-Augmentation-Techniques-for-Breast-Cancer

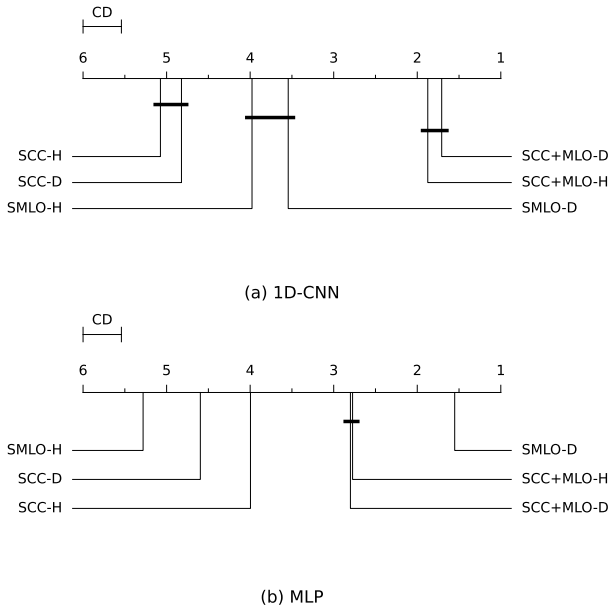

Figure 4. A Nemenyi Plot illustrates the comparative performance of dataset setups on the $DDSM$ dataset across nine augmentation methods using 1D-CNN (a) and MLP (b) as classifiers. The setups are ranked in descending order from 1 to 6, where 1 represents the highest rank (best performance), and 6 represents the lowest rank (worst performance). The methods connected by a horizontal bar lie within the Critical Distance (CD), indicating no significant difference in average ranks. The significance level is set at α = 0.05. Here, D is for deep features, and H is for Haralick features.

Table 4. The STEM/Mixup combination added in the augmentation methods to explore the diversity of generated samples from Mixup and STEM using both 1D-CNN and MLP classifiers.

Setups	1D-CNN	MLP
$S_{CC} - D$	STEM/Mixup	STEM/Mixup
$S_{CC} - H$	STEM/Mixup	Mixup
$S_{MLO} - D$	STEM/Mixup	ADASYN
$S_{MLO} - H$	Mixup	STEM
$S_{CC+MLO} - D$	STEM/Mixup	STEM
$S_{CC+MLO} - H$	STEM/Mixup	Mixup
WBC	STEM	STEM/Mixup

the MLP network over the 1D-CNN. Given that the inputs to both models are feature vectors, few or no spatial relations exist among neighbouring scalars, so dense layers are more suited than convolutional ones to extract discriminating information. For the WBC database, this difference is especially large, because the feature extraction process for FNA is the one that preserves the less spatial relations.

The comparison between Haralick and GoogleNet features shows that, in general, both extractors presented a similar performance. However, a significant difference was ob-

served in S_{MLO}, under MLP classification, in which Deep GoogleNet features clearly outperform its competitor and, additionally, obtained the best overall performance. We hypothesize that MLP could take special advantage of the deep features due to the fact that they are not as specialized as the Haralick ones. Also, S_{MLO} provides a more global view of the breast, better captured by GoogleNet while Haralick, by being a local texture descriptor, could have missed.

In our experiments, the S_{CC} view seemed to hinder the discrimination capacity of the models. Not only it presented the worst performance by itself but, when added to S_{MLO} (S_{CC+MLO}), it was worse than S_{MLO} alone.

Regarding the augmentation approaches, no significant difference was observed among them under MLP classification, suggesting a lack of responsiveness of dense layers to synthetic data obtained from the nine approaches presented. On the other hand, the 1D-CNN model is clearly benefited by the STEM/Mixup approach, as it has shown to be more sensitive to the synthetic data.

6. Conclusion

In this study, we delve into a critical aspect of Deep Learning (DL) for Breast Cancer (BC) diagnosis using mammographic images: the significance of data augmentation. We emphasize how the selection of augmentation methods is closely tied to the unique characteristics and requirements of the diagnostic task. This highlights the ongoing need for thorough investigation to determine the most effective approach for enhancing DL-based BC diagnosis. Through a systematic evaluation of various augmentation techniques on CC and MLO views from the Digital Database for Screening Mammography (DDSM) and the Wisconsin Breast Cancer (WBC) datasets, we identify that Mixup and a combined version of the Synthetic Minority Over Sampling Technique (SMOTE) with Edited Nearest Neighbour (ENN) Mixup (STEM) are the most promising approaches for 1D-CNN based architectures. A rigorous statistical analysis supports our findings.

Based on our experiments, we remark the following:

- Given the responsiveness of dense layers for feature vectors and the global perspective provided by S_{MLO}, the most promising setup involves the use of an MLP classifier with deep features and using only the MLO view. Any augmentation technique can be considered.
- When a dense architecture is not an option, an alternative like 1D-CNN can be used. For that case, deep features from only the MLO view, and augmented through STEM/Mixup (or only Mixup if STEM is possible), are preferred, given the responsiveness of convolutional layers to STEM/Mixup augmented features.
- When the features cannot be algorithmically derived from images, like the case of FNA in the WBC database, a convolutional architecture (e.g. 1D-CNN) will probably present poor performance, and the impact of implicit data augmentation will likely be precarious, as no spatial information is preserved in such representations.
- If deep features are not available but local extractors (e.g. Haralick) are, the CC view is more appropriate, and a better performance is expected with an MLP classifier, as S_{CC} provides a more local view that matches the receptive field of local extractors.
- For the same reason, when only the CC view is available, the Haralick features are the most appropriate, and better performance is expected with an MLP.

Although our work does not entail significant ethical concerns, given that we work with synthetic data at the feature level that cannot be associated to any patient, we highlight the importance of informed consent and appropriate data anonymization. Our work takes part into the efforts that constitute an important step into algorithmic fairness, as it objective is to reduce biases in the training data.

The scope of this research can be broadened in the future by incorporating alternative feature extraction methods, such as local binary pattern and wavelet transform, alongside the existing deep features. Exploring the fusion effect of handcrafted and deep features in classification tasks presents an intriguing avenue for further investigation. Additionally, delving into the comparative analysis of these diverse feature sets could offer valuable insights into the strengths and limitations of different feature extraction techniques, ultimately enhancing the robustness and accuracy of classification models in medical image analysis. On the other hand, we encourage any effort towards accountability regarding the effect of the studied augmentation methods on the decision boundaries of the classification algorithms. Finally, the results obtained from the STEM/Mixup combination suggest that simply merging the synthetic data from multiple augmentation techniques may positively impact generalization. Hence, more thoroughly exploring such combinations is a promising approach for future endeavors.

7. ACKNOWLEDGEMENTS

The Science Foundation Ireland (SFI) Centre for Research Training in Artificial Intelligence (CRT-AI), Grant No. 18/CRT/6223 and the Irish Software Engineering Research Centre (Lero), Grant No. 16/IA/4605, both provided funding for this study.

References

[1] A hybrid classifier combining borderline-smote with airs algorithm for estimating brain metastasis from lung cancer: A case study in taiwan. *Computer Methods and Programs in Biomedicine*, 119(2):63–76, 2015. 2

[2] Improving performance of classifiers for diagnosis of criti-

cal diseases to prevent covid risk. *Computers and Electrical Engineering*, 102:108236, 2022. 2, 3

[3] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[4] Mohamed A Abdou. Literature review: Efficient deep neural networks techniques for medical image analysis. *Neural Computing and Applications*, 34(8):5791–5812, 2022. 1

[5] Suja A Alex, NZ Jhanjhi, Mamoona Humayun, Ashraf Osman Ibrahim, and Anas W Abulfaraj. Deep lstm model for diabetes prediction with class balancing by smote. *Electronics*, 11(17):2737, 2022. 2

[6] Sikha S Bagui, Dustin Mink, Subhash C Bagui, and Sakthivel Subramaniam. Determining resampling ratios using bsmote and svm-smote for identifying rare attacks in imbalanced cybersecurity data. *Computers*, 12(10):204, 2023. 2

[7] Gustavo EAPA Batista, Ana LC Bazzan, Maria Carolina Monard, et al. Balancing training data for automated annotation of keywords: a case study. *Wob*, 3:10–8, 2003. 2, 4

[8] Richard A Bauder and Taghi M Khoshgoftaar. The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. *Health information science and systems*, 6:1–14, 2018. 1

[9] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 2, 3

[10] Chukwuebuka Joseph Ejiyi, Zhen Qin, Happy Monday, Makuachukwu Bennedith Ejiyi, Chiagoziem Ukwuoma, Thomas Ugochukwu Ejiyi, Victor Kwaku Agbesi, Amarachi Agu, and Chiduzie Orakwue. Breast cancer diagnosis and management guided by data augmentation, utilizing an integrated framework of shap and random augmentation. *BioFactors*, 2023. 1

[11] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006. 5

[12] Jacques Ferlay, Murielle Colombet, Isabelle Soerjomataram, Donald M Parkin, Marion Piñeros, Ariana Znaor, and Freddie Bray. Cancer statistics for the year 2020: An overview. *International journal of cancer*, 149(4):778–789, 2021. 1

[13] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937. 6

[14] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2011. 2

[15] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31, 2018. 2

[16] Chongomweru Halimu, Asem Kasem, and SH Shah Newaz. Empirical comparison of area under roc curve (auc) and mathew correlation coefficient (mcc) for evaluating machine learning algorithms on imbalanced datasets for binary classification. In *Proceedings of the 3rd international conference on machine learning and soft computing*, pages 1–6, 2019. 5

[17] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005. 2, 3

[18] Xiaotian Han, Zhimeng Jiang, Ninghao Liu, and Xia Hu. G-mixup: Graph data augmentation for graph classification. In *International Conference on Machine Learning*, pages 8230–8248. PMLR, 2022. 4

[19] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973. 1

[20] Yumnah Hasan, Fatemeh Amerehi, Patrick Healy, and Conor Ryan. Stem rebalance: A novel approach for tackling imbalanced datasets using smote, edited nearest neighbour, and mixup. In *2023 IEEE 19th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 3–9. IEEE, 2023. 4

[21] Yumnah Hasan, Allan de Lima, Fatemeh Amerehi, Darian Reyes Fernandez de Bulnes, Patrick Healy, and Conor Ryan. Interpretable solutions for breast cancer diagnosis with grammatical evolution and data augmentation. *arXiv preprint arXiv:2401.14255*, 2024. 3

[22] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008. 2, 3

[23] Michael Heath, Kevin Bowyer, Daniel Kopans, P Kegelmeyer Jr, Richard Moore, Kyong Chang, and S Munishkumaran. Current status of the digital database for screening mammography. In *Digital Mammography: Nijmegen, 1998*, pages 457–460. Springer, 1998. 2

[24] Steffen Herbold. Autorank: A Python package for automated ranking of classifiers. *Journal of Open Source Software*, 5 (48):2173, 2020. 5

[25] Min-Wei Huang, Chien-Hung Chiu, Chih-Fong Tsai, and Wei-Chao Lin. On combining feature selection and oversampling techniques for breast cancer prediction. *Applied Sciences*, 11(14):6574, 2021. 2

[26] Peng Jun Huang. *Classification of imbalanced data using synthetic over-sampling techniques*. University of California, Los Angeles, 2015. 3

[27] Zeshan Hussain, Francisco Gimenez, Darvin Yi, and Daniel Rubin. Differential data augmentation techniques for medical imaging classification tasks. In *AMIA annual symposium*

proceedings, page 979. American Medical Informatics Association, 2017. 1

[28] Xiaoyan Jiang, Zuojin Hu, Shuihua Wang, and Yudong Zhang. Deep learning for medical image-based cancer diagnosis. *Cancers*, 15(14):3608, 2023. 1

[29] Md Faisal Kabir and Simone Ludwig. Classification of breast cancer risk factors using several resampling approaches. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1243–1248. IEEE, 2018. 2

[30] Taha Muthar Khan, Shengjun Xu, Zullatun Gull Khan, et al. Implementing multilabeling, adasyn, and relieff techniques for classification of breast cancer diagnostic through machine learning: Efficient computer-aided diagnostic system. *Journal of Healthcare Engineering*, 2021, 2021. 2

[31] Cherry Khosla and Baljit Singh Saini. Enhancing performance of deep learning models with different data augmentation techniques: A survey. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, pages 79–85. IEEE, 2020. 2

[32] Rohit Kundu and Soham Chattopadhyay. Deep features selection through genetic algorithm for cervical pre-cancerous cell classification. *Multimedia Tools and Applications*, 82 (9):13431–13452, 2023. 1, 3

[33] Shaoyuan Lei, Rongshou Zheng, Siwei Zhang, Shaoming Wang, Ru Chen, Kexin Sun, Hongmei Zeng, Jiachen Zhou, and Wenqiang Wei. Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. *Cancer Communications*, 41 (11):1183–1194, 2021. 1

[34] Boyi Li, Felix Wu, Ser-Nam Lim, Serge Belongie, and Kilian Q Weinberger. On feature normalization and data augmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12383–12392, 2021. 2

[35] Tajul Miftahushudur, Halil Mertkan Sahin, Bruce Grieve, and Hujun Yin. Enhanced svm-smote with cluster consistency for imbalanced data classification. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 431–441. Springer, 2023. 3

[36] Peter Bjorn Nemenyi. *Distribution-free multiple comparisons.* Princeton University, 1963. 6

[37] Hien M. Nguyen, Eric W. Cooper, and Katsuari Kamei. Borderline oversampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21, 2011. 2, 3

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 5

[39] Lalu Ganda Rady Putra, Khairan Marzuki, and Hairani Hairani. Correlation-based feature selection and smote-tomek link to improve the performance of machine learning methods on cancer disease prediction. *Engineering & Applied Science Research*, 50(6), 2023. 2

[40] Joao Rala Cordeiro, António Raimundo, Octavian Postolache, and Pedro Sebastião. Neural architecture search for 1d cnns—different approaches tests and measurements. *Sensors*, 21(23):7990, 2021. 5

[41] Conor Ryan, Krzysztof Krawiec, Una-May O'Reilly, Jeannie Fitzgerald, and David Medernach. Building a stage 1 computer aided detector for breast cancer using genetic programming. In *Genetic Programming: 17th European Conference, EuroGP 2014, Granada, Spain, April 23-25, 2014, Revised Selected Papers 17*, pages 162–173. Springer, 2014. 3

[42] Manisha Saini and Seba Susan. Deep transfer with minority data augmentation for imbalanced breast cancer dataset. *Applied Soft Computing*, 97:106759, 2020. 1

[43] Harsh Sharma and Anushika Gosain. Oversampling methods to handle the class imbalance problem: A review. In *International Conference on Soft Computing and its Engineering Applications*, pages 96–110. Springer, 2022. 4

[44] Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 408–421, 1972. 2, 3

[45] William H Wolberg, W Nick Street, and Olvi L Mangasarian. Breast cancer wisconsin (diagnostic) data set [uci machine learning repository], 1992. 2

[46] Peng Yao, Shuwei Shen, Mengjuan Xu, Peng Liu, Fan Zhang, Jinyu Xing, Pengfei Shao, Benjamin Kaffenberger, and Ronald X Xu. Single model deep learning on imbalanced small datasets for skin lesion classification. *IEEE transactions on medical imaging*, 41(5):1242–1254, 2021. 1

[47] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization, 2018. 4

[48] Liying Zhang and Haihang Sun. Esa-gcn: An enhanced graph-based node classification method for class imbalance using enn-smote sampling and an attention mechanism. *Applied Sciences*, 14(1):111, 2023. 4