Balancing Act- Addressing Class Imbalance in Medical Image Classification

Sarika Gadupudi- 030822244
sarika.gadupudi01@student.csulb.edu

Simran Sarawagi- 030824584
simran.sarawagi01@student.csulb.edu

Instruction to run the code:

1. Ensure you have Python installed on your system.
2. Install the required libraries by running:
   pip install numpy torch torchvision scikit-learn imbalanced-learn matplotlib
3. To access the dataset module install the following library:
   pip install datasets
4. Install the following libraries for running plots and images:
   pip install torchcam seaborn

Google Colab Link: **Final.ipynb**

Abstract

This project addresses imbalanced datasets in medical image classification by exploring various preprocessing techniques, including data augmentation, undersampling, oversampling [9], alongside advanced methods like SMOTE and Borderline SMOTE. We stress the need to maintain image quality while addressing class imbalance, keeping in mind the risk of producing unrealistic images using SMOTE-like techniques. Additionally, we highlight the significance of class activation maps in understanding model behavior and feature learning. Through experimentation, our aim is to enhance model performance, particularly in minimizing false negatives in medical diagnoses.

1) Introduction:

Medical image analysis plays a crucial role in diagnosing various diseases and conditions, including pneumonia detection through chest X-ray imaging. However, one common challenge in medical image analysis is dealing with imbalanced datasets [10], where one class significantly outnumbers the other. This imbalance can lead to biased models favoring the majority class, resulting in poor detection of the minority class. Such misclassifications, whether false positives or false negatives, can have serious repercussions, including delayed treatment or misdiagnosis, impacting patient outcomes. Achieving a balance between minimizing false positives and maximizing true positives is vital in medical classification tasks.

This study explores various data preprocessing techniques to address the imbalance in chest X-ray pneumonia datasets, aiming to enhance pneumonia detection accuracy and reliability,

thereby improving patient care and outcomes.

2) Related Work:

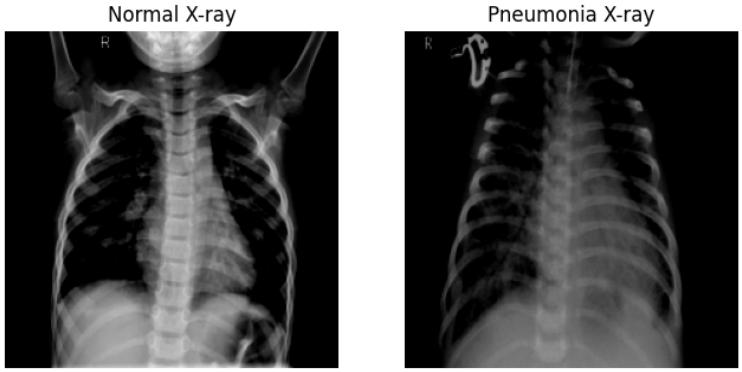
Class imbalance is a pressing concern in medical imaging, crucial for accurate diagnosis. Traditional methods like under-sampling and oversampling address imbalances but often reduce dataset diversity. The Synthetic Minority Over-sampling Technique (SMOTE) by Chawla et al. [2] has notably improved class balance by creating synthetic samples, yet it raises realism of training images. Studies by Hasan, Srwa & Sagheer, Ali & Veisi, Hadi [3] have demonstrated SMOTE's effectiveness in breast cancer classification, highlighting a need for domain-specific validation.

Despite challenges, Class Activation Maps (CAMs) have become crucial for enhancing model interpretability. Moore's work [1] on CAMs in diagnosing retinal diseases showcases their role in validating model decisions, improving diagnostic accuracy, and fostering trustworthy machine learning applications in healthcare.

BorderlineSMOTE[8] improves upon SMOTE by exclusively oversampling minority class instances near the class boundary, under the assumption that those further from the boundary are less prone to misclassification. Furthermore, synthetically augmented datasets can lead to model overfitting by generating non-photorealistic images, as highlighted by Shorten and Khoshgoftaar [4]. A more effective approach could involve initially passing the image data through a pre-trained CNN to extract feature representations, followed by applying SMOTE to these representations.

3) Dataset:

The dataset utilized in this study was obtained from the Hugging Face model repository under the name "hf-vision/chest-xray-pneumonia"[6]. This dataset comprises a total of 5,863 chest X-ray images [7] in JPEG format, categorized into two classes: Pneumonia and Normal. The dataset was divided into a training set and a test set. The training set consisted of 5,216 samples, while the test set contained 624 samples.



3.1 Class Imbalance

Upon examining the distribution of classes within the training dataset, it was observed that there exists a class imbalance, with a significantly higher number of pneumonia cases compared to normal cases. Specifically, there were 1,341 Normal cases and 3,875 Pneumonia cases.
This class imbalance poses a challenge for machine learning models, as they may tend to be biased towards the majority class, leading to suboptimal performance in detecting the minority class.

4) Methodology:

4.1 ResNet Model

ResNet18[12], pre-trained on the ImageNet dataset, was adapted for pneumonia detection. The last fully connected layer of the ResNet18 model was replaced with a new linear layer with two output units to accommodate the binary classification task of pneumonia detection.
Cross-entropy loss was utilized as the training objective in the ResNet18 model, aiding in minimizing the disparity between predicted and actual class probabilities. It serves as a crucial component in optimizing the model's parameters to enhance classification accuracy and convergence during training.

4.2 Undersampling

In the undersampling technique, the majority class, which represents pneumonia cases, was selectively reduced to match the minority class count. This balancing act is crucial to preventing the model from favoring the majority class during training. Using the np.random.choice function, samples from the majority class are randomly selected without replacement, creating a representative subset. These samples are then combined with the original minority class samples, ensuring a balanced class distribution. The combined dataset is shuffled to maintain randomness and loaded into a DataLoader, configured with a specified batch size and shuffling enabled, to ensure each training batch presents a balanced mix of classes, facilitating more equitable learning.

4.3 Oversampling

Oversampling addresses class imbalance by replicating instances of the minority class (normal cases), using random sampling to ensure adequate exposure during model training. This process involves using the np.random.choice function to select minority class instances with replacement, effectively increasing their presence. These oversampled minority samples are then concatenated with the original majority class samples to form a balanced training dataset.

4.4 SMOTE (Synthetic Minority Over-sampling Technique) [2]

SMOTE (Synthetic Minority Over-sampling Technique) is employed to alleviate class imbalance by generating synthetic samples for the minority class through interpolation between existing samples. This process begins by collecting and concatenating training data using a DataLoader, ensuring image dimensions align with model requirements. SMOTE is then applied to the flattened image data, effectively balancing the class distribution by augmenting the minority class. The enhanced dataset is restructured into PyTorch tensors and loaded into a new DataLoader for training. The model is then trained using this SMOTE-augmented data.

4.5 Borderline SMOTE [8]

Borderline SMOTE is utilized to enhance class distribution by generating synthetic samples near the decision boundary between classes, aiming to boost model generalization. This process begins with data loading and transformation, where medical images and labels are prepared in DataLoader format suitable for neural network processing. After converting image tensors into NumPy arrays, Borderline SMOTE is applied to create balanced training data. The synthetic data is then converted back into tensors and used for model training, incorporating techniques such as forward passes and backpropagation to improve learning from both original and synthetic samples.

4.6 Data Augmentation Techniques

Data augmentation techniques were employed to further enhance the robustness of the model and improve its generalization performance. These techniques include:

i)Random Rotation: Randomly rotating the images by a certain degree to simulate variations in image orientation.

ii)Horizontal Flip: Randomly flipping the images horizontally to simulate mirror images.

These augmentation techniques help increase the diversity of the training dataset and enable the model to learn from a wider range of image variations, thereby improving its ability to generalize to unseen data.

5) Experiments:

5.1 Experimental Setup

The experiments were conducted using Google Colab with a batch size of 32 and a fixed number of epochs set to 5. All input images were resized to a standard size of 224 x 224 pixels. Additionally, grayscale images were converted to 3-channel grayscale to match the input format expected by ResNet18. The dataset was preprocessed using various techniques, including undersampling, oversampling, SMOTE, Borderline SMOTE, and data augmentation. Each preprocessing technique was evaluated using the ResNet18 model as the baseline architecture, trained with the Adam optimizer and a learning rate of 0.005.

5.2 Metrics Used

Each preprocessing technique's performance was assessed using three vital metrics: F1 score, balanced accuracy, and recall. These metrics are well-suited for medical image classification tasks due to the inherent class imbalance and the critical need to minimize false negatives (missed diagnoses). F1 score offers a balanced evaluation of precision and recall, while recall specifically measures the model's capability to accurately identify positive cases, crucial for medical diagnosis. Balanced accuracy, averaging recall and specificity, provides a holistic evaluation of the model's performance across both classes, indicating its ability to maintain balance in predictions despite class imbalance.

5.3 Recall and Accuracy Plots

We calculated the balanced accuracy and recall across 5 epochs and visualized them using Matplotlib:

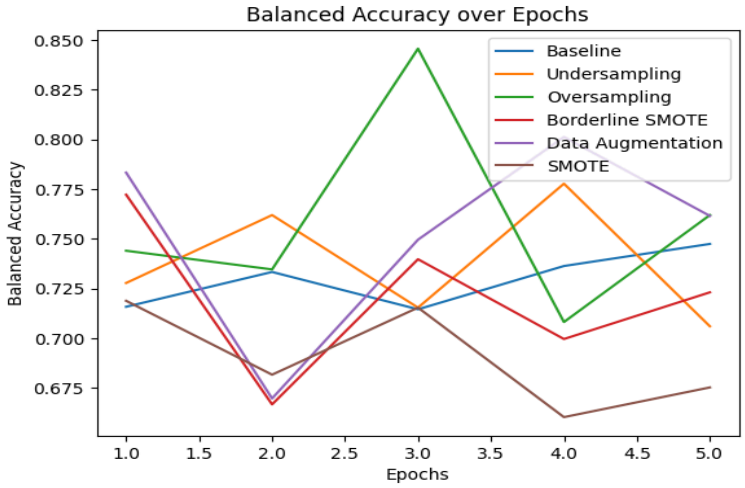


Fig1. Balanced Accuracy vs Epochs

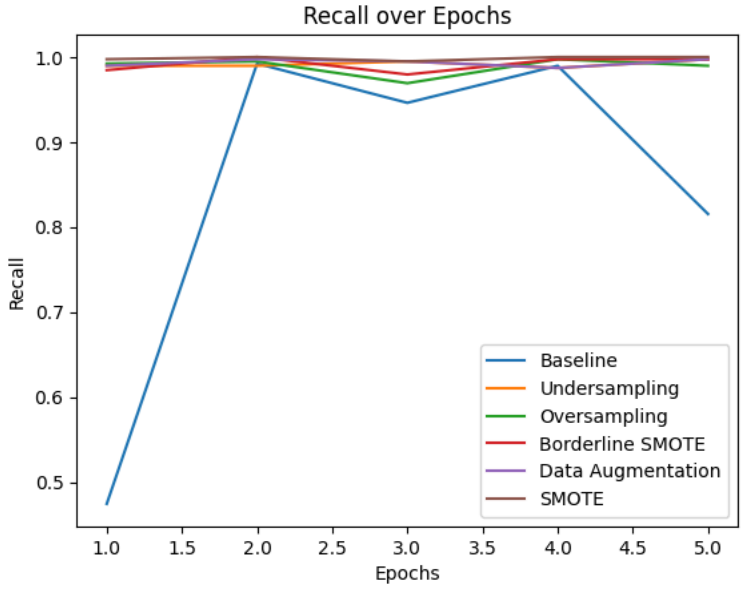


Fig2. Recall vs Epochs

5.4 Results

We compared the Accuracy, Recall and F1 score for different preprocessing techniques against our baseline model:

Experiment	Balanced Accuracy	Recall	F1 Score
Baseline	0.74	0.81	0.81
Undersampling	0.70	0.99	0.84
Oversampling	0.76	0.98	0.87
Data Augmentation	0.76	0.99	0.87
SMOTE	0.67	1.0	0.83
Borderline-SMOTE	0.72	0.99	0.85

Furthermore, confusion matrices were examined to understand the predictive performance of the model and the distribution of true positives, false positives, true negatives, and false negatives across various techniques:

Experiment	TP	TN	FP	FN
Baseline	318	159	75	72
Undersampling	389	97	137	1
Oversampling	386	125	109	4
Data Augmentation	389	123	111	1
SMOTE	390	82	152	0
Borderline-SMOTE	389	105	129	1

5.5 Class Activation Maps

We employed SmoothGradCAMpp[11] to generate Class Activation Maps, which highlight influential regions in X-rays guiding the model's predictions, providing deeper insights into

model behavior. We observed that the baseline model(Fig 3) tended to focus predominantly on a specific corner of the image, neglecting the central part of the lung.
Undersampling(Fig 4) emphasized the corner areas, but it also produced few CAMs attempting to extract features from the central area of the lung.

Fig3. Baseline Model

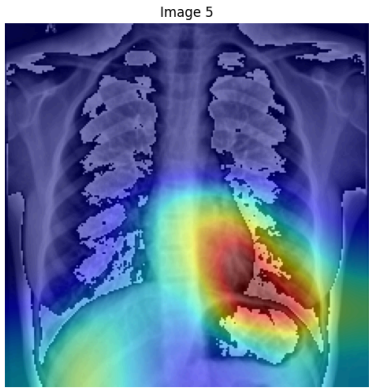


Fig4. Undersampling

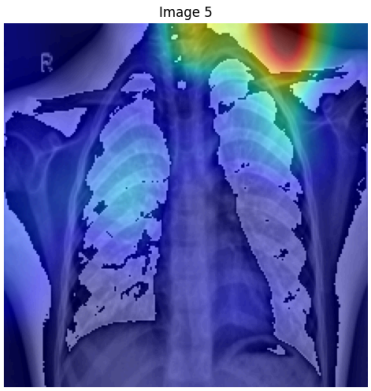


Fig5. Borderline SMOTE

Fig6. SMOTE

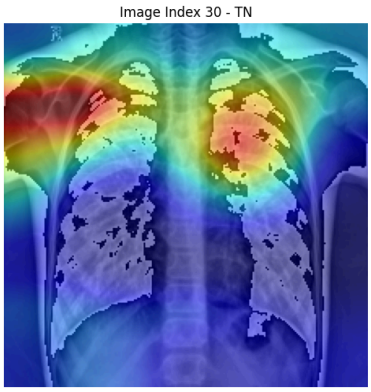


Fig7. Oversampling(upper region)

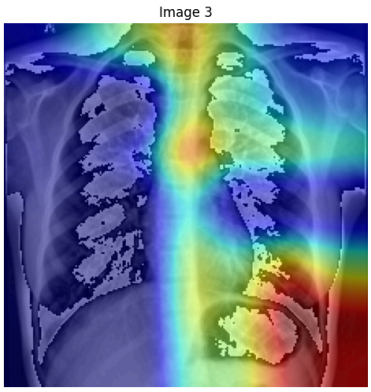


Fig8. Oversampling(right region)

Fig9. Oversampling(left region)

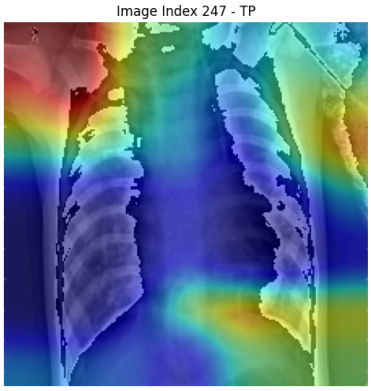


Fig10. Data Augmentation

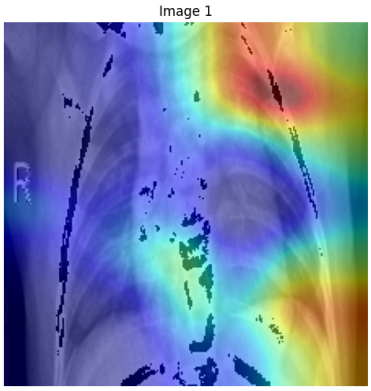


Fig11. Data Augmentation

For Borderline SMOTE(Fig 5) and SMOTE(Fig 6), the model focused more on the background, generating CAMs that were not relevant.

In contrast, with oversampling(Fig 7,8&9) and data augmentation(Fig10,11), we observed a variety of CAMs being generated, each focusing on different parts of the lung, aiming to cover the entire area and extract features from it.

5.6 Discussion

In this study, we investigated the effectiveness of various data preprocessing techniques in addressing class imbalance within medical image classification. We assessed the performance of each technique using metrics such as F1 Score, Recall, and Balanced Accuracy. Below is a comprehensive discussion of the performance of each technique and the significance of these metrics in medical image classification.

The Baseline Model exhibited moderate performance but displayed limitations in effectively managing class imbalance, evident from its lower recall and F1 Score. This highlights the necessity of employing resampling techniques to tackle the inherent dataset imbalance.

Undersampling notably enhanced recall (0.9974) and F1 Score (0.8493) but at the expense of Balanced Accuracy (0.7060). This suggests that while the model became highly sensitive to positive cases, it suffered in accurately classifying negative cases, potentially leading to an increase in false positives.

Oversampling yielded improvements across all metrics compared to the baseline, achieving the highest Balanced Accuracy (0.7620) and a robust F1 Score (0.8723). This implies that oversampling facilitated the creation of a more balanced training set, enabling the model to perform well across both classes.

The Augmented Model, incorporating techniques such as random rotations and flips, attained the highest F1 Score (0.8742) and maintained high recall (0.9974) and Balanced Accuracy (0.7615). This highlights the efficacy of augmentation in providing diverse training samples and aiding the model in generalization.

The SMOTE Model achieved the highest recall (1.0), ensuring no positive cases were overlooked. However, it exhibited the lowest Balanced Accuracy (0.6752) among the tested methods, indicating challenges in accurately classifying negative cases.

Borderline-SMOTE also demonstrated high recall (0.9974) and improved F1 Score (0.8568) but displayed lower Balanced Accuracy (0.7231) compared to the Oversampling Model. This suggests that while effective, Borderline-SMOTE may not achieve as balanced results as simple oversampling in this scenario.

6)Conclusion:

The Augmented and Oversampling Models performed exceptionally well, displaying the highest F1 Scores and Balanced Accuracies, effectively managing class imbalance. The high recall of the SMOTE Model makes it suitable for scenarios intolerant to missing positive cases, underscoring the importance of selecting the right preprocessing technique, especially in medical image

classification where false negatives can be costly.

Recall is crucial in medical image classification, measuring the model's ability to identify all relevant instances. The SMOTE Model achieved a perfect recall of 1.0, successfully identifying all positive cases. Similarly, other models like Undersampling, Borderline-SMOTE, and Augmented also demonstrated high recall values of 0.9974, effectively capturing nearly all positive cases.

Balanced Accuracy, accounting for class imbalance, is more informative than plain accuracy. The Oversampling Model achieved the highest Balanced Accuracy of 0.7620, indicating commendable performance across both classes. Conversely, the Baseline Model, despite lower recall and F1 Score, exhibited a relatively high Balanced Accuracy of 0.7474, suggesting it maintained a reasonable balance between the classes without advanced resampling techniques.

The F1 Score, balancing precision and recall, is valuable where both false positives and false negatives are concerns. The Augmented Model attained the highest F1 Score of 0.8742, closely followed by the Oversampling Model with 0.8723, highlighting the effectiveness of these techniques in balancing precision and recall.

References:

[1] Moore C, Class activation mapping (CAM). Reference article, Radiopaedia.org (Accessed on 17 May 2024) https://doi.org/10.53347/rID-72380

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research*, 16, 321-357. DOI:10.1613/jair.953

[3] Hasan, Srwa & Sagheer, Ali & Veisi, Hadi. (2021). Improving Breast Cancer Classification Using (SMOTE) Technique and Pectoral Muscle Removal in Mammographic Images. Mendel. 27. 8. 10.13164/mendel.2021.

[4] Shorten, Connor and Taghi M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning." *Journal of Big Data* 6 (2019): 1-48.

[5] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

[6] https://huggingface.co/datasets/hf-vision/chest-xray-pneumonia

[7] Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, V2, doi: 10.17632/rscbjbr9sj.2

[8] Han, H., Wang, WY., Mao, BH. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang, DS., Zhang, XP., Huang, GB. (eds) Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, vol 3644. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11538059_91

[9] Wongvorachan, T.; He, S.; Bulut, O. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Information 2023, 14, 54. https://doi.org/10.3390/info14010054

[10]Gao L, Zhang L, Liu C, Wu S. Handling imbalanced medical image data: A deep-learning-based one-class classification approach. Artif Intell Med. 2020 Aug;108:101935. doi: 10.1016/j.artmed.2020.101935. Epub 2020 Aug 7. PMID: 32972664; PMCID: PMC7519174.

[11] https://github.com/frgfm/torch-cam

[12]https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html