Addressing Data Imbalance Issue in Medical Images

Saroj Bhatta
Department of Computer Science
Wake Forest University
Winston-Salem, NC 27109
sbhatta24@wfu.edu

Mohammad Marufur Rahman
Department of Computer Science
Wake Forest University
Winston-Salem, NC 27109
rahmm224@wfu.edu

Abstract

Data imbalance, which occurs when a dataset contains significantly larger number of samples of data in one class than the others, is a prevalent issue in the field of machine learning. Data imbalance can pose serious problems like biased predictions and poor accuracy and generalization for minority class. While various data sampling strategies, like Random Under Sampling (RUS), Random Over Sampling (ROS), Adaptive Synthetic Sampling (ADASYN), Synthetic Minority Oversampling Technique (SMOTE), and Synthetic Minority Oversampling Technique Tomek (SMOTETomek) etc. have been employed to address this issue, they can often introduce challenges like over-fitting, information loss, and can be computationally complex and expensive. To address these limitations, we proposed a new strategy that involves creating pairs of data instances and classifying each pair as being from same class or not. A Convolutional Neural Network (CNN) model was then trained using these paired instances. During testing, unseen instances were also paired up with known positive and negative instances allowing the model to predict whether the pair belonged to the same class. While this proposed strategy yielded increase in recall by 20% and slight increase in F1 score by 3.4%, the overall accuracy decreased by 7% and gain in AUC was only 0.6, which was not significant. Moreover, some basic image augmentation techniques like masked augmentation and interpolation were also explored and employed, but they failed to provide any significant improvement from baseline model. Therefore, more advanced image augmentation techniques, including generative methods such as diffusion models, were also explored. Despite generating images using both a simple diffusion model and pre-trained diffusion models, the performance of the classification model did not show improvement. This may be due to the generated images failing to preserve the specific features of hernia X-ray images. Therefore, further fine-tuning of the model as well as a deeper understanding of the biomedical characteristics of hernia for image augmentation could be beneficial for further study. The code repository for this study can be accessed on Github at https://tinyurl.com/cs674dataimb

1 Introduction

In a classic classification problem, a dataset contains distinct target labels. When all the classes contain an equal number or same proportion of samples in the dataset, it is called a balanced dataset. In contrast, when different classes have largely varying numbers of samples, meaning difference in the proportion of samples from different classes, it creates a imbalance in the dataset.

A machine learning model trained on an imbalanced dataset perform poorly as they develop bias towards majority class, and in some worst cases, they completely disregard the minority class (6).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

This sort of biased model will have poor generalization and low accuracy towards minority class. Moreover, evaluating model's performance will be difficult since the traditional performance metrics may not correctly reflect the effectiveness of the model across imbalanced dataset.

Imbalanced datasets are often common in the medical domain, as most people tend to be generally healthy (2). Therefore, the data collected from ill patients constitutes the minority class. However, machine learning models that are designed for disease detection focus on identifying and classifying ill patients rather than healthy ones. This emphasis on the minority class is crucial for early diagnosis of illness and treatment of those ill patients.

In the medical domain, misclassification of a disease can have severe consequences. For example, a life-threatening disease or medical condition incorrectly being classified as normal medical condition has a far higher cost than a normal medical condition being incorrectly being classified as a serious medical condition. Misclassifying normal medical condition may lead to additional tests, but misclassifying serious medical condition will lead to serious health risk (6).

Therefore, more research on developing better strategies to address and tackle the challenges posed by data imbalance emerges as an important undertaking. Investigating and developing robust methods to address imbalance can foster advancements for machine learning applications in medical domain, leading to more accurate detection of medical conditions.

2 Related work

Even though there has been recent advancements in handling imbalanced data, there is still relatively sparse research on class balancing approaches. This study (2) used credit card fraud detection data and analyzed different approaches of class balancing including RUS, ROS, ADASYN, SMOTE, SMOTETomek and trained Support Vector Machine model. The study found that SMOTETomek method produced best performance among all other data sampling strategies applied. In another study (3), a CNN model is trained using contrastive learning where the model learns representation from unlabeled data through similarities between plant diseases images, achieving 87.42% accuracy which was higher than contemporary supervised learning models. One of the studies (5) presented a hybrid approach of using both unsupervised learning through cluster analysis and supervised learning through decision tree as a promising solution to deal with imbalanced class, but it does not provide a detailed study, and the effectiveness of other machine learning approaches are not analyzed.

A majority weighed minority oversampling method, BIRCH and Boundry Midopoint Centroid Synthetic Minority Over-Sampling Technique (BI-BMCSMOTE), is also proposed which considers the boundaries of minority samples and their cluster density functions (7). It tries to reduce the shortcomings of traditional over-sampling methods, which often overlook crucial information of boundary samples and high similarity between existing and newly generated samples. Other oversampling strategies include oriented oversampling with spatial information (OOSI) to deal with noisy dataset by generating high quality artificial data (8). Data under-sampling is also another practice used to reduce the imbalance in dataset. In the study (9), a combination of data augmentation and a conventional under-sampling method yield huge improvement in accuracy and F1-score. Most of these studies and papers explored different over-sampling and under-sampling strategies, their variations, and hybrid approaches incorporating sampling strategies with other data strategies or algorithm based strategies.

A comprehensive study of image data augmentation approaches (10) lays out a taxonomy for all relevant image data augmentation approaches. Several image manipulations, including geometric (rotation, translation, shearing, flipping) and non-geometric manipulation (cropping, noise, color-space, jitter), as well as image erasing techniques are discussed under basic image data augmentation methods. However, these basic techniques can incur information loss through erasing, cutouts, masking, excessive geometric manipulations, etc. especially for sensitive medical images like x-ray images, and requires a detailed understanding of the dataset to obtain optimal degree of geometric manipulation for the best results. The study also uncovered and discussed many advanced data augmentation techniques like image mixing, auto augmentation, feature augmentation, neural style transfer, and diffusion data augmentation. It also showed that diffusion methods produce realistic data variations contributing to data diversity and high model robustness and performance gain.

Recently, generative models, especially diffusion models, have made great strides in creating realistic images across different fields. However, there is still limited research on using them for medical

image generation (12). This study also adjudges that, compared to other models, like Generative Adversarial Networks (GANs), diffusion models have shown better performance, stability, and variety in generating high-quality images. Moreover, a 2021 study (13) highlighted the impressive results of diffusion models all while also improving the diversity of the generated images. One of the diffusion techniques as discussed in (10), DiffuseMix, is known to preserve label integrity by avoiding unrealistic image generation which also offers robustness against adversarial attacks. The researchers who proposed DiffuseMix (14), where original image is combined with generated counterpart while preserving the basic image semantics, also showed that DiffuseMix technique demonstrated better performance, on multiple tasks such as general classification, fine-grained classification, data security, fine-tuning, and adversarial robustness, than existing image augmentation approaches.

A survey study (4) that uncovers various approaches in use to handle data imbalance issues showed that even though there are some algorithmic approaches proposed to address this issue, there is no universal approach that would solve data imbalance issue in all application domains. Therefore, it would require deep research to formulate such separate algorithms and strategies for separate domains. Another study (11) also suggested that different data augmentation methods can have different effects depending on the type of data and the task being performed. In the case of medical imaging, specific augmentation techniques are likely needed to create realistic data samples and help deep neural networks learn more effectively. Deep learning-based image data generation techniques are capable of generating more realistic images and broader diversity in data (11).

3 Methodology

3.1 Dataset

The data for this project on is collected from the Huggingface website (1). The dataset is on Chest X-ray from the National Institutes of Health (NIH). It contains 112,120 frontal-view X-ray images of 30,805 unique patients and has 14 different disease labels. For this project we will work on Hernia disease. Out of 112,12 X-ray images, only 227 images are categorized as hernia positive while 60,361 images are not classified as diseased. This is a highly imbalanced dataset where out of the total data, 43% of images are from non-diseased patients and only 0.16% of images are from hernia patients. The dataset contains data of Posteroanterior (PA) and Anteroposterior (AP) view. For this project we consider only PA view chest x-ray images. The labels were generated from a NLP model with accuracy more than 90%. That indicates label annotations are almost 90% accurate.

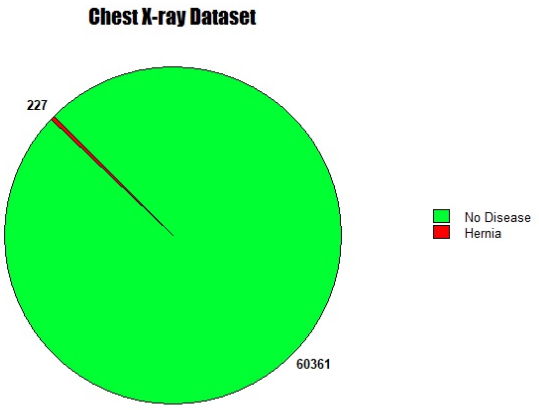


Figure 1: Chest X-ray dataset.

3.2 Proposed technique

The imbalanced dataset we have chosen contains 60,361 samples of non-diseased X-ray images and only 227 samples of hernia X-ray images. Therefore, to transform this highly imbalanced dataset

into a relatively balanced dataset, we will first create pairs of these images. The non-diseased X-ray images will be paired with other non-diseased X-ray images as well as the hernia X-ray images will be paired with other hernia X-ray images. These pairs will then be classified as being similar as each pair contains images from the same class i.e. non-diseased and hernia. Subsequently, the non-diseased X-ray images will be paired up with hernia X-ray images and these pairs will be classified as being not similar. Through this process, the new transformed dataset will not be as skewed as the original dataset.

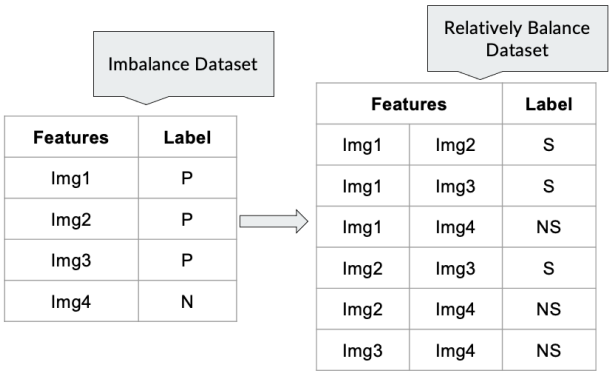


Figure 2: Dataset tranformation.

3.2.1 Model training

The newly transformed dataset will have pairs of images as its feature and similarity as target label. We will then train a CNN model as it is widely popular in machine learning to work with image dataset. The CNN model will learn to predict whether a pair of images are similar or not.

3.2.2 Inference

Once the model is trained, we will use it to classify the unseen instances of non-diseased and hernia X-ray images. Each unseen X-ray image will be paired up with a certain number of known non-diseased X-ray as well as known hernia X-ray. The trained CNN model will be asked to classify these pairs as similar or not similar. We will analyze the results obtained from the model and make an inference of whether the given unseen instance is non-diseased or hernia X-ray image. If the model classifies the given unseen instance paired up with non-diseased X-ray images as similar the greater number of times than its pairs with hernia X-ray images, the unseen instance will be classified as non-diseased and vice versa.

The exact strategy for pairing up the X-ray images to create the transformed dataset as well as for pairing up the unseen X-ray image with known non-diseased and hernia image will be explored and defined in the next couple of weeks of project work.

3.3 Performance and evaluation

Once the model is trained and tested, various model performance metrics like accuracy, precision, recall, F1-score, confusion matrix, ROC curve, AUC etc. will be calculated. These metrics will be used to evaluate the performance of the model and the effectiveness of the transformed dataset in addressing imbalanced dataset. The proposed method will be compared with the baseline model performance on imbalanced dataset and existing approaches for handling imbalance dataset.

4 Implementation and Experiments

4.1 Baseline model training

In this study, data centric approach is followed where instead of finding the best model, best dataset for a particular model is produced. Here, the classification model LeNet is used which is a Convolution

Test Sample
X

Features		Predicted Label
X	P	S
X	P	S
X	P	NS
X	N	NS
X	N	NS
X	N	S

Decision
P

Figure 3: Prediction inference.

Neural Network that has two convolution layers each followed by a pooling layer. Figure 4 shows the architecture of LeNet. For our experimentation, to reduce computational burden, 1000 normal labeled images and 192 hernia labeled images taken from the dataset. Images were reshaped into 128×128 in order to further reduce computational complexity. 70% images from each class is used for model training and 10% for validation and remaining 20% for testing the model.

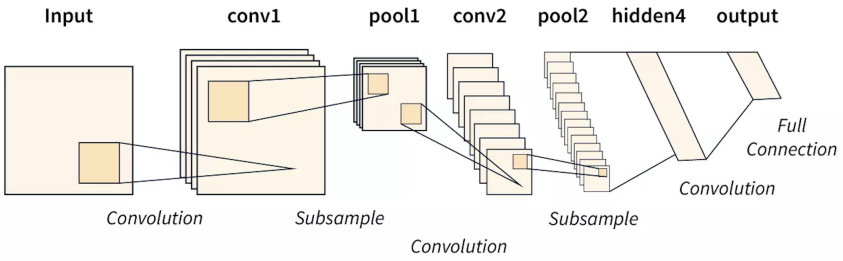


Figure 4: LeNet model architecture.

Performance of the classifier model on the imbalanced dataset is showed in confusion matrix and ROC curve in figure 5. It was observed that the the base model produced an accuracy of 81%, precision of 39%, recall of 34%, F1 score of 36.6%, and area under the curve (AUC) in ROC of 0.66.

4.2 Proposed technique

Training data was transformed, where each sample from the negative class (normal) is randomly paired up with a sample from the positive class (hernia) and labeled as different (1). For both positive

	Actual	
Predicted	Normal	Hernia
Normal	181	25
Hernia	20	13

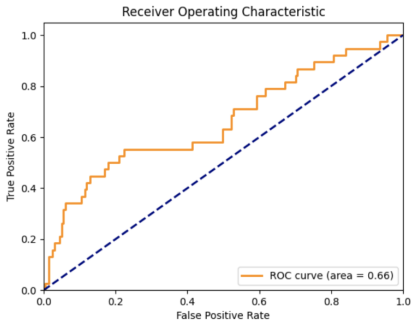


Figure 5: Confusion matrix and ROC for imbalanced data.

and negative classes, each instance is paired up with randomly selected instance from the same class and labeled as same (0). The validation dataset was transformed in a similar fashion. The classifier is trained on this transformed dataset and evaluated on the test set. Some samples from both classes were preserved from the training dataset as known positive and negative samples to make pairs with test samples while evaluating performance of the method. In figure 6, confusion matrix and ROC curve for the proposed technique is presented.

	Actual	
Predicted	Normal	Hernia
Normal	156	18
Hernia	44	21

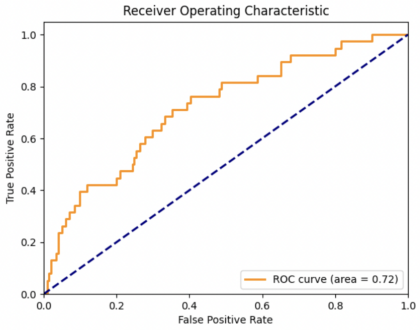


Figure 6: Confusion matrix and ROC curve for proposed method.

In this approach, even though the accuracy dropped to 74% from 81%, recall improved to 54% which resulted slightly higher F1 score of 40%. This was 3.4% increase from the baseline F1 score of 36.6%. In spite of improvement in performance with this approach, it was not a vast and significant improvement, which can also be seen from AUC which only increased by 0.6. Moreover, even though the transformed dataset is relatively less imbalanced than the original one, total information in the dataset is still imbalanced and new information is not being added. That posses significant threat for the model being over-fitted to pairing patterns. Therefore, the proposed approach alone is insufficient to address the issue of imbalanced data, necessitating the exploration of data augmentation methods as well.

4.3 Data augmentation techniques

Data augmentation is a widely adopted strategy for handling imbalanced dataset. Data augmentation techniques can be grouped into two broad categories: basic image data augmentation and advanced data augmentation (10). Fundamental techniques such as image manipulation and image erasing are considered basic image data augmentation, while the advanced image augmentation encompasses complex techniques like image mixing, auto augmentation, feature augmentation, neural style transfer, and generative methods. In this study, a few techniques like masked augmentation, interpolation, and diffusion models will be explored.

4.3.1 Masked augmentation

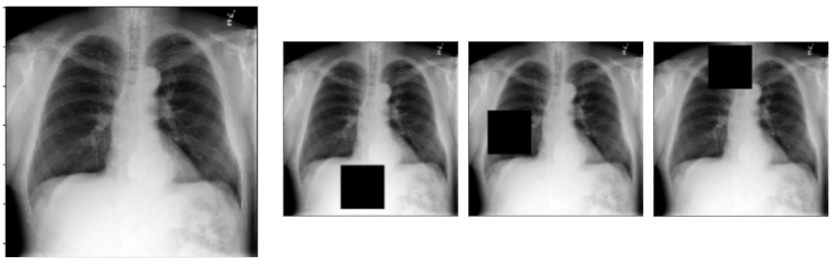


Figure 7: Augmentation using masking.

In masked augmentation, some parts of the input data are intentionally hidden or 'masked' to create modified versions (or augmented daset) of the original data. For chest x-ray images, a randomly selected patch of shape (16×16) was erased from the original image and used for augmenting the dataset. For each image in the hernia class, two masked images were generated. Data augmentation

was only performed on the training set. Figure 7 shows original image and images after masking. Performance of this augmentation technique is shown figure 8. It was observed that the performance of the model using masked augmentation technique is very close to the baseline performance and did not show significant improvement.

	Actual	
Predicted	Normal	Hernia
Normal	167	22
Hernia	34	16

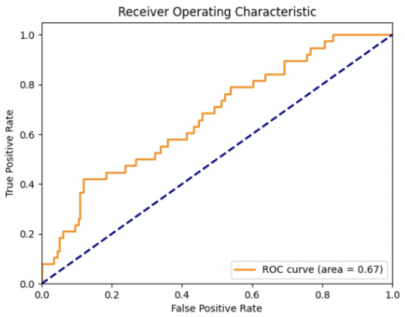


Figure 8: Confusion matrix and ROC curve for masking method.

4.3.2 Augmentation using interpolation

Interpolation is used for image data augmentation where each pixel of the generated image is calculated as a weighted sum of corresponding pixel values of multiple images.

$$I_{new} = f \times I_1 + (1 - f) \times I_2$$

Here, $f = 0.9$ was used and two images I_1 and I_2 were selected randomly from the hernia class. The number of generated images were two times the number of original images.

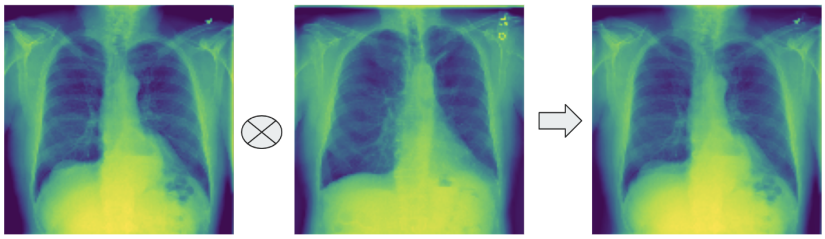


Figure 9: Augmentation using interpolation method.

Performance of the model using interpolation as data augmentation approach is shown in figure 9. It is clear that the classifier failed to distinguish hernia images which could be due to the feature dilation during performing weighted sum. This resulted in even worse performance than the baseline model.

	Actual	
Predicted	Normal	Hernia
Normal	200	36
Hernia	1	2

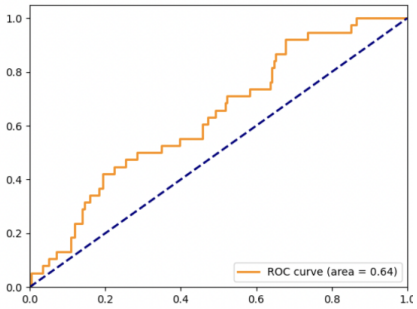


Figure 10: Confusion matrix and ROC curve for interpolation method.

4.3.3 Image generation using diffusion model

Diffusion models are a class of generative models that generate data by simulating a gradual noise-removal process, effectively learning how to reverse a diffusion process that degrades data into noise. The model starts with real images and applies a small amount of random noise obtained from a Gaussian distribution in a step by step manner. With each step data gets noisier, eventually becoming pure noise. This process is called forward process. Mathematically, the process can be represented as a sequence:

$$x_1, x_2, ...x_T$$

where, x_T is almost pure noise and x_0 is the original image. The forward process is described as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathcal{I})$$

where, $\alpha_t \in (0, 1)$ is a noise scaling factor at each timestep t. $\mathcal{N}(.; \mu, \sigma^2)$ denotes Gaussian distribution. $\mathcal{I}$ is the identity matrix. The goal of training is to approximate the reverse process by minimizing a loss function that measures how well the model recovers x_{t-1} from x_t. Loss function:

$$\mathcal{L} = \mathbb{E}_{x_0,t,\epsilon}[|\epsilon - \epsilon_\theta(x_t, t)|^2]$$

where, ϵ is the true noise added at timestep t, and $\epsilon_\theta(x_t, t)$ is the predicted noise from the model. Once the $\epsilon_\theta(x_t, t)$ is estimated, the original image x_0 can be estimated using:

$$\hat{x}_0 = \frac{1}{\sqrt{\hat{\alpha}_t}}(x_t - \sqrt{1-\hat{\alpha}_t}.\epsilon_\theta(x_t, t))$$

where, $\hat{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative product of noise scaling factors up to timestep t.

In reverse process or de-noising process the goal is to learn to gradually de-noising a pure noise to recover the original image.

An Unet architecture was implemented to predict the noise. It contains two convolution layer followed by two transposed convolution layers. Noise was added for $T = 10$, and starting and ending noise level was 0.0001 and 0.0005 respectively. Figure 11 shows original image, noisy image at timestep T, and reconstructed image which can be used as augmentation.

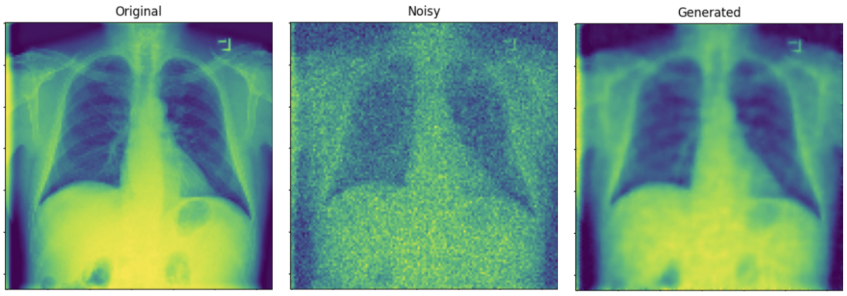


Figure 11: Image generation using diffusion model.

4.3.4 Stable diffusion

Pre-trained diffusion model Stable Diffusion v1.5 which is a text-to-image generative model based on latent diffusion framework is fine-tuned to generate hernia x-ray images. Stable Diffusion model do not add noise directly on the pixel space, instead it operates on a lower-dimensional latent space. This approach reduces computational costs and improves efficiency. In the core of the architecture there is a UNet model that predicts noise at each step of the reverse diffusion process. It operates on the latent features and learns to predict how to de-noise and refine the image progressively. The model uses text-encoder Constrastive Language Image Pretraining model that utilizes cross-attention mechanism. In this study, text-encoder part of the model was kept unmodified and only the UNet model's weight was updated. Architecture of the Stable Diffusion model is shown in figure 12. The model takes an original image and encodes it into latent space then passes through forward diffusion process. Then in reverse diffusion de-noising was performed conditioned with encoded prompt.

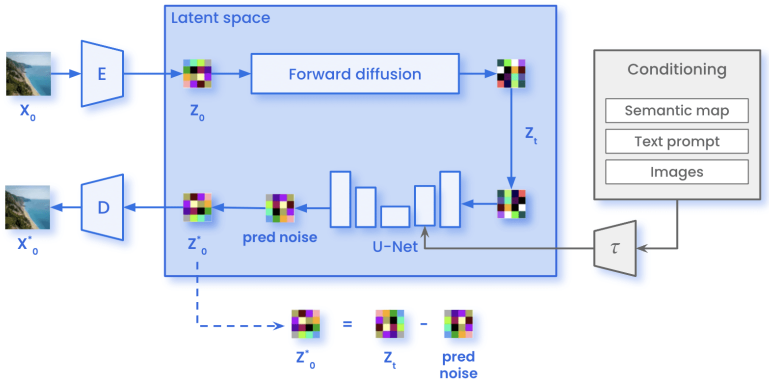


Figure 12: Stable diffusion model architecture (17).

The pre-trained model was collected from Hugging Face (16) and fine tuned for 10 epochs. After fine tuning the model was used to generate synthetic hernia images. As input an original hernia image and a text prompt ("chest xray image of hernia patient") was given. Figure 13 shows a generated image from the fine tuned model. Using synthetic images generated by the fine-tuned stable diffusion model hernia sample class was augmented. This augmented dataset was used to train the classification model. The classifier showed 36% recall and 29% overall F1 score. Confusion matrix and ROC curve of the classifier is presented in figure 14.

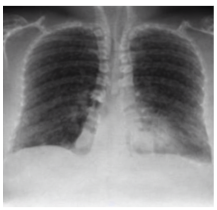


Figure 13: Generated image from stable diffusion model.

	Actual	
Predicted	Normal	Hernia
Normal	157	24
Hernia	43	14

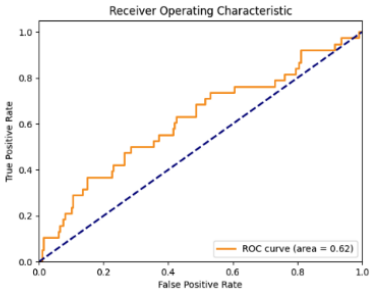


Figure 14: Classifier performance on stable diffusion model augmented dataset.

4.3.5 DiffuseMix

DiffuseMix is a label-preserving data augmentation technique with diffusion models (14). First diverse images are generated using a stable diffusion model via bespoke conditional prompts. The diffusion model used in DiffuseMix is InstructPix2Pix. In DiffuseMix, both original and generated images are used to create hybrid images for image data augmentation. Then, a portion of natural original image is concatenated with its generated counterpart to obtain hybrid image which preserves key semantics and avoids label ambiguities. A portion of fractal images are then blended into these hybrid images to improve the overall structure complexity of augmented image. The finally obtained image is then used as an augmented image. The architecture of DiffuseMix is shown below.

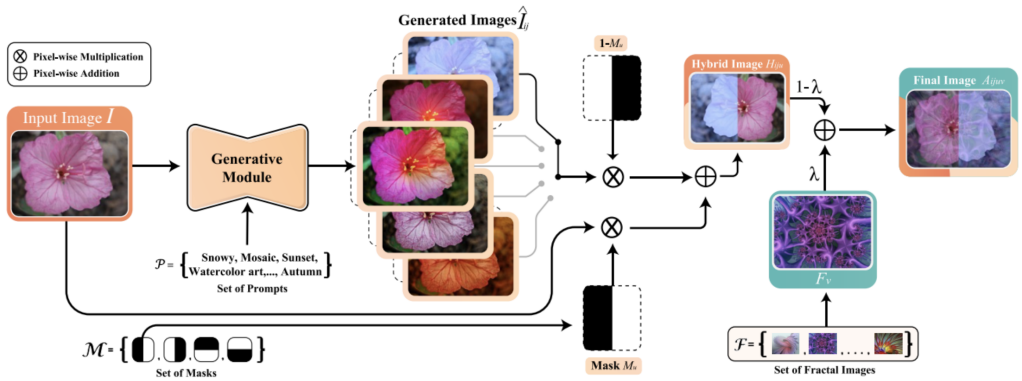


Figure 15: Architecture of DiffuseMix proposed in (14)

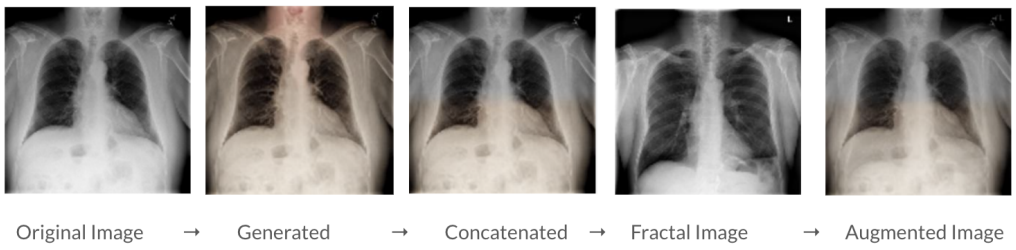


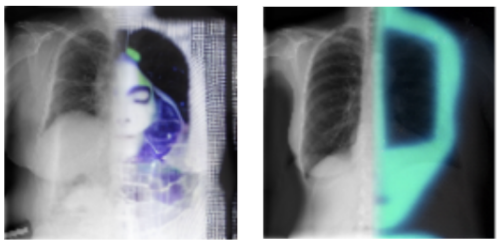
Figure 16: Augmentation using DiffuseMix

As shown in the architectural diagram figure 15, in our study, each input image from the dataset, went through image generation using prompts. Some of the prompts that were used on multiple rounds of trials of image generation were 'slightly bright', 'slightly dim', 'jitter', 'minor gaussian noise', 'minor subtle shadows.' Then, generated image and input image are divided into two equal parts through horizontal or vertical splitting, and one half of original image and the other half of generated image was used to get hybrid image. A set of hernia x-ray images itself were used as fractal images. For DiffuseMix, 20% of fractal images is blended into hybrid image to get final image. An overall process of image generation used in our study using DiffuseMix is shown in figure 16.

The performance of the classification model using DiffuseMix image data augmentation technique was evaluated. Even though the accuracy of the model was around 84%, it was attributed entirely because of the true negatives, so the image augmentation using did not produce much meaning to our study. It was seen that the classifier failed to detect hernia image at all. Upon further exploration of the augmented dataset, some extremely unrealistic images were seen to be generated, as shown in figure 17, which might have affected the model training and its ability to classify hernia images.

	Actual	
Predicted	Normal	Hernia
Normal	206	39
Hernia	0	0

Confusion Matrix



Unrealistic images generated by DiffuseMix

Figure 17: Performance of classification model using DiffuseMix augmentation

5 Project Summary

5.1 Conclusion

In this study, we implemented and evaluated performances for the baseline LeNet model with imbalanced dataset, proposed paired technique, some augmentation techniques like masked augmentation and interpolation, as well as more advanced generative augmentation techniques like diffusion model including some pre-trained models like Stable Diffusion v1.5 and DiffuseMix.

From our implementation, it was observed that, the classification model performed fairly poorly on the imbalanced dataset as expected. While implementing paired-images approach as proposed in our methodology, we observed a slight improvement in classification where recall and F1 score improved by nearly 20% and 3.4% respectively.

The image augmentation techniques like masking or interpolation performed either similar to the baseline performance with imbalanced dataset or even worse than the baseline. The performance could be attributed towards the nature of augmentation methodologies. Masking could have masked the portions of x-ray images that represent hernia anomalies while interpolation could have induced feature dilation while performing weighted sum.

We also studied about and used diffusion models to generate synthetic images of hernia class and evaluated the performance based generated images' augmented dataset. We implemented a simple diffusion model for image generation with UNet architecture which performed similar to the baseline model with no significant improvement in correct predictions.

Similarly for pre-trained diffusion models like Stable Diffusion v1.5 and DiffuseMix, we did not observe improvement in performance of the classification model. These pre-trained image generation models are not specifically trained to work for x-ray images, so their fine-tuning was extremely difficult, and due to time and resource constraints, the generative models could not be highly optimized. Even though no significant improvement was observed for the classifier performance using diffusion model in this study, it would not be rational just yet to state that these pre-trained models are not suitable for hernia x-ray image data augmentation. Further tuning the diffusion model with deeper understanding of the pre-trained models as well as biomedical features and semantics of hernia x-rays could be beneficial in generating more realistic and label-preserving augmented x-ray images.

5.2 Limitations and future work

We found out that the labels in our dataset was obtained from a natural language processing (NLP) model. Therefore, we believe that there is a possibility of wrongly labeled hernia images in our dataset. Moreover, for our generative images data augmentation implementation, we did not find a readily available pre-trained model specialized in x-ray image generation. This limited our ability to exploit the full potential of generative models. The pre-trained models are trained on simple images dataset like ImageNet, TinyImageNet, CIFAR datasets and did not necessarily have enough capabilities to effectively generate x-ray images preserving their peculiar semantics and labels. In addition to this, fine-tuning of pre-trained models take substantial amount of data and time, both of which we are limited to.

For future work in this study, we would like to understand the peculiarities of hernia and their x-ray images. This understanding would help to preserve the major semantics of hernia in x-rays during augmentation. For pre-trained diffusion models like Stable Diffusion v1.5 and DiffuseMix, further fine-tuning could be done. Careful adaptation to and change in conditional prompts for generative model and change in fractal images set could produce different result than what we have observed so far. Furthermore, using knowledge distillation techniques to train a smaller model to mimic a large pre-trained vision model, then fine-tuning it might also produce different results.

5.3 Teamwork

For this project, we have both coordinated well and worked together throughout the semester. We have set aside dedicated group project time on every Friday afternoons, and we use this time to discuss about the project and next directions, and to work together.

We collected over twenty papers on data imbalance and data augmentation. Both of us started with literature review separately (not all of them are referenced in the bibliography section as some of them turned out to be irrelevant for this proposed project). We would brief each other's understandings during the weekly project meeting and brainstorm ideas for project proposal. Both Saroj and Maruf worked on proposal presentation slides.

Once the project was proposed, Saroj did more of the literature review and developed the written proposal report while Maruf started setting up the LeNet model for our study. Once the model was setup, Maruf trained the model on imbalanced dataset and implemented the proposed approach using paired-up strategy. Saroj also implemented the pair-up approach separately by making the pairs of images from same class and different classes balanced and trained the model in this balanced pairs. Since no difference in performance was observed by implementing different pair-up strategies, we moved forward with original pair-up strategy.

As it was evident from our implementation, as well as from Dr. Yang's suggestion, that we would need to move forward with augmentation techniques, we collected more papers on data augmentation and individually did literature reviews. Maruf studied and implemented masking approach. On the other hand, Saroj studied and implemented interpolation. The performances of these approaches pushed us to further explore more advanced data augmentation technique. Both of us studied about generative models, especially diffusion models. For the midterm report, Saroj updated our project paper's abstract and literature review sections and Maruf implemented the diffusion model architecture from scratch for generating augmented images. Maruf implemented the diffusion model following the algorithm provided in the original paper and trained and tuned to generate high quality hernia images.

Both of us explored many pre-trained large vision models and decided to implement and experiment with two models. Saroj primarily worked on DiffuseMix and implemented DiffuseMix for our study while Maruf worked on implementing Stable Diffusion v1.5.

While we took responsibilities for separate sections and tasks, we did most of our work during weekly project meetings and thus shadowed each other's work and implementation for the most part. A summary of major work distribution is given below:

Saroj:

- Literature review and project proposal report.
- Implementation of proposed paired-images approach creating balanced pairs.
- Implementation of augmentation using interpolation method.
- Implementation of image augmentation pipeline using DiffuseMix.
- Updates to project midterm and final report.

Maruf:

- Literature review and project proposal report.
- Pre-process and prepare the dataset.
- Implement LeNet architecture.
- Training baseline classification model using imbalanced dataset.
- Implementation of proposed paired-images approach.
- Implementation of augmentation using masking method.
- Implementation and training diffusion model from scratch.
- Implementation of augmentation pipeline using pre-trained Stable Diffusion v1.5.
- Updates to project midterm and final report.

References

[1] https://huggingface.co/datasets/alkzar90/NIH-Chest-X-ray-dataset

[2] Ayyannan, M. (2024, April). Accuracy Enhancement of Machine Learning Model by Handling Imbalance Data. In 2024 International Conference on Expert Clouds and Applications (ICOECA) (pp. 593-599). IEEE.

[3] Chung, B. (2024, February). Addressing Data Imbalance in Plant Disease Recognition through Contrastive Learning. In 2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC) (pp. 1-6). IEEE.

[4] Deepa, N., & Sumathi, R. (2022, December). A survey on state of art approaches in handling imbalance, positive and unlabelled data. In 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS) (pp. 1-6). IEEE.

[5] Ugarković, A., & Oreški, D. (2022, October). Supervised and Unsupervised Machine Learning Approaches on Class Imbalanced Data. In 2022 International Conference on Smart Systems and Technologies (SST) (pp. 159-162). IEEE.

[6] Mahani, A., & Ali, A. R. B. (2019). Classification problem in imbalanced datasets. Recent Trends in Computational Intelligence, 1.

[7] Wang, C. R., & Shao, X. H. (2020). An improving majority weighted minority oversampling technique for imbalanced classification problem. IEEE Access, 9, 5069-5082.

[8] Deng, Y., & Li, M. (2023). An Adaptive and Robust Method for Oriented Oversampling with Spatial Information for Imbalanced Noisy Datasets. IEEE Access.

[9] Ishikawa, T., Yakoh, T., & Urushihara, H. (2022). An NLP-inspired data augmentation method for adverse event prediction using an imbalanced healthcare dataset. IEEE Access, 10, 81166-81176.

[10] Kumar, T., Brennan, R., Mileo, A., & Bendechache, M. (2024). Image data augmentation approaches: A comprehensive survey and future directions. IEEE Access.

[11] Rayavarapu, S. M., Prasanthi, T. S., Gottapu, S. R., & Singam, A. (2024, August). A Comprehensive Overview on Data Augmentation Techniques for Medical Images. In 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1324-1329). IEEE.

[12] Song, W., Jiang, Y., Fang, Y., Cao, X., Wu, P., Xing, H., & Wu, X. (2023, September). Medical Image Generation based on Latent Diffusion Models. In 2023 International Conference on Artificial Intelligence Innovation (ICAII) (pp. 89-93). IEEE.

[13] Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34, 8780-8794.

[14] Islam, K., Zaheer, M. Z., Mahmood, A., & Nandakumar, K. (2024). DiffuseMix: Label-Preserving Data Augmentation with Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 27621-27630).

[15] Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., & Merhof, D. (2023). Foundational models in medical imaging: A comprehensive survey and future vision. arXiv preprint arXiv:2310.18689.

[16] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

[17] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).