City University of London

MSc in Data Science

Project Report

2019 -2020

Data Augmentation for Multi-Classification in Medical Imaging

Supervised by: Giacomo Tarroni

Student: Aqsa Mahmood

Date of submission: 15/11/2020

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: **Aqsa Mahmood**

Abstract

The aim of this research was to discover the effectiveness of data augmentation in the development of a robust and precise computer aided diagnosis system to detect knee disorders. For this purpose, two data augmentation techniques particularly traditional data augmentation techniques, and GAN based data augmentation were chosen to synthesis new data that potentially helps to increase variance in the training dataset. The convolutional neural network based architecture, MRNet, was used to perform classification. The data augmentation and classification were performed on a publicly available MRNet dataset containing knee MRI scans. For traditional data augmentation, affine and pixel level transformations were applied on 2D images in knee MRI scans. Moreover, Wasserstein GAN was utilized to generate knee MRI scans. The classification model was trained with non augmented data, geometrically augmented data, WGAN based augmented data, and both geometric and WGAN augmented dataset. The results obtained by this study suggest that GAN based data augmentation assists to enhance the performance of the diagnosis system.

Keyword: Data augmentation, Computer aided diagnosis system, Magnetic resonance imaging (MRI), deep learning, Generative adversarial networks (GAN).

Contents

1 Chapter 1: Introduction: (10%) (1,539)

1.1 Background

The knee is an easily injured joint of the body during sports activities. Nearly 2.5 million injuries are experienced by athletes annually. The knee joint is composed of bones, meniscus, ligaments, and tendons. Anterior cruciate ligament (ACL) tear, meniscus tears, and other abnormalities such as sprains and strains, swelling, arthritis are common knee injuries. The knee injuries span from sprains to complete damage in ligament tissues. Physical therapy and braces are nonsurgical treatments for knee injuries. In most cases of ACL tear injuries, arthroscopic surgery is required to recover knee function. ACL injuries are mostly accompanied by damage in other ligament tissues, mainly meniscus tears (The knee: Anatomy, injuries, treatment, and rehabilitation, 2020).

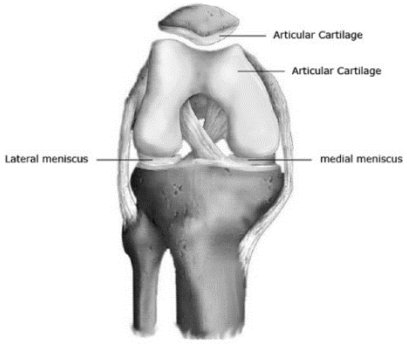


Figure 1: Anatomical structure of knee (Chung and Chung, 2020)

Physical examination of knee structure is the primary approach to diagnose knee injuries. In order to confirm the diagnosis, an imaging test is needed. X-ray and MRI scans are used to detect injuries. X-ray images do not indicate injuries in ligaments, but they are widely utilized to discover fractures in the bone. MRI scans give improved images that help diagnose injuries in soft tissues of ligaments, nerves, and tendons. Further, high diagnostic performance is achieved by the non-invasive medical imaging technique MRI. Due to these reasons, the anatomical structure of the knee is widely studied through MRI scans. Besides, MRI is expensive than other medical imaging techniques such as CT scans, X-rays.

MRI is a versatile and widely used medical imaging technique in a clinical setting for diagnosis. In order to generate a comprehensive image of a body part, MRI uses the natural magnetic properties of the body. When the body is placed in an MRI scanner, all protons' axes fall into line. The uniform arrangement produces a magnetic vector that is oriented alongside the axis of the MRI scanner. The MRI scanners have magnetic fields of different strengths, generally among 0.5 and 1.5 teslas (Berger, 2020). Different MRI sequences are utilized to generate images of specific characteristics to facilitate bone structure and ligament tissues' assessment. An MRI sequence is a particular arrangement of pulse sequences and pulsed-field gradients to create an image of specific characteristics. T1 weighted, T2 weighted, and proton density (PD) weighted sequences are commonly used MRI sequences. In the T1 weighted image, high signal intensities are generated by bone marrow and fat, whereas ligaments, cartilage, fluid cause low MR signal intensities. On the other hand, T2 weighted image has high signal intensities produced by ligaments, cartilage, fluid, and low MR signal intensities caused by bone

marrow and fat. Further, the tissues with high proton density are enhanced in PD weighted image (Normal knee MRI, 2020).

For the purpose of visualizing the structure of body organs at different angles, MRI generates images in axial, sagittal, and coronal planes. The coronal plane views the body part from front to back, the sagittal plane from left to right, and the axial plane from top to bottom (MRI Plane Mathematics, 2020). Analyzing MRI images of body organs taken at different angles helps to make a more accurate diagnosis. Furthermore, MRI is a medical imaging technique that does not damage body tissues as it uses radiations in the radiofrequency range (Berger, 2020).

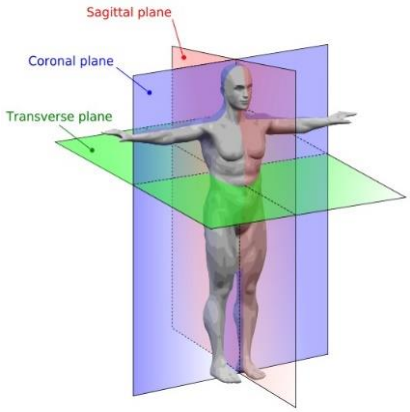


Figure 2: imaging planes (MRI Plane Mathematics, 2020)

1.2 Motivation and Beneficiaries of our study

The choice of the domain was motivated by the applications of data science in healthcare. Particularly, medical diagnosis is the area of interest of this project. The data science technologies such as deep learning have been successfully utilized to build a computer-aided diagnosis system that assisted medical experts in making a more accurate diagnosis and helps to reduce the diagnosis time significantly. An early diagnosis of disease reduces the progression of the disease and prevents the need for surgical procedures. This project is opted to build a robust computer-aided diagnosis system using deep learning techniques to diagnose knee disorders. The output of our study will assist the researchers in the clinical domain. Moreover, the potential output of this study could benefit radiologists in knee disorder diagnosis.

1.3 Objectives

This study seeks, with the assistance of MRNet dataset comprising knee MRI scans (Stanford Machine Learning Group, 2020), to answer the question.

Is data augmentation beneficial in developing a robust and accurate computer-aided diagnosis (CAD) system to diagnose knee injuries?

For this purpose, different data augmentation techniques were employed to explore their impact on medical image diagnosis. In order to get the answer to the research question, various key goals were set up.

- Perform data augmentation using traditional data augmentation methods.
- Build a generative adversarial network (GAN) model to generate knee MRIs.

- Develop a deep learning-based classifier to diagnose knee disorders and assess its performance with different types of augmented data.

1.4 Methods

The techniques used to accomplish the objectives are mentioned below.

1.4.1 Data Augmentation techniques

Data augmentation was performed using traditional data augmentation techniques and GAN based data augmentation.

1.4.1.1 Classical data augmentation

Classical data augmentation was employed by applying the geometric and pixel-level transformations such as horizontal flipping, rotation, translation brightness, and contrast on the images stacked in each MRI scan.

1.4.1.2 GAN based data Augmentation

For GAN based data augmentation, Wasserstein GAN (WGAN) was trained. The best model architecture was selected by trying different architectures for generator and critic networks along with experimenting with different hyperparameter settings. The quality of generated images was assessed by observing images manually and critic loss curves. For quantitative performance measure, the classifier's performance was investigated with different ratios of generated and real data. In total, three WGAN models were trained to generate knee MRI scans for three types of knee injuries, individually. Nearly 3,00 MRI scans comprising almost 10,000 2D image slices for each class were added to enlarge the size of the dataset.

1.4.2 Classifying patients with knee disorders

Convolution neural network (CNN)-based architecture, particularly MRNet, was trained to perform classification. In the MRNet architecture, feature extraction was performed through transfer learning with the pre-trained weights of the AlexNet model. Three MRNet models were trained to detect the absence or presence of three types of knee disorders individually. The performance of the classification model was assessed with non augmented data, geometrically augmented data, WGAN based augmented data, and both geometric and WGAN based augmented data. Accuracy, the area under the receiver operating curve (AUC), specificity, sensitivity, precision, recall, and f score were used to investigate the classification model's performance.

1.5 Work plan

The project plan illustrated in Figure 3 differs a little from the plan displayed in the proposal (Appendix A). Initially, it was planned to perform GAN based data augmentation using two variants

of GAN, namely WGAN and Progressive, growing GAN (PGGAN). Besides, the training of the classification model was scheduled using transfer learning approach and training the model from scratch. The training of WGAN and classification model took significantly longer than expected. Because of longer training time and limited computational resources, the PGGAN based data augmentation and classification model's training from scratch were excluded from the project plan. In general, no substantial changes were made to the objective of the project.

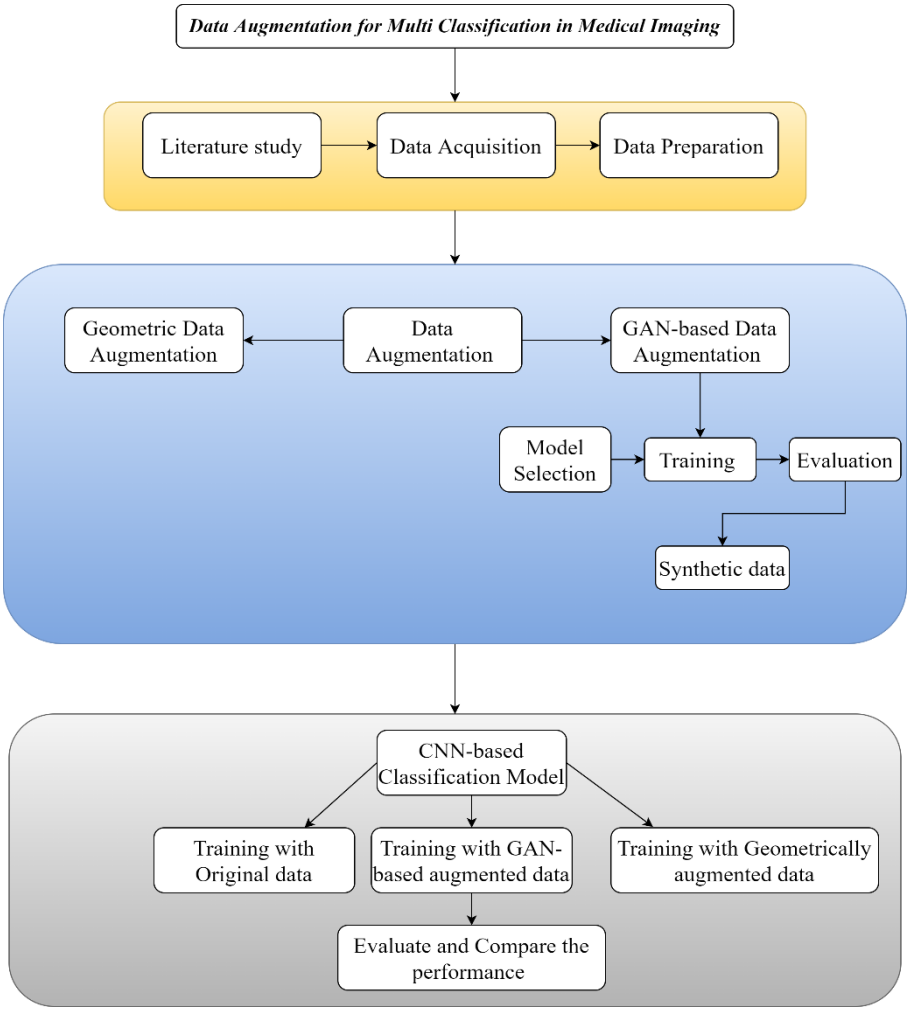


Figure 3: Work plan

The project's core purpose was to explore the diagnosis system's performance with different types of augmented data, potentially leading to the development of a robust and accurate computer-aided diagnosis system to detect disorders in knee MRI scans. In the initial stage, the previous research studies on this subject were explored to decide techniques to implement. Subsequently, a suitable

dataset for the project's purpose was chosen. Next, the dataset was pre-processed and divided into training and test set.

Two types of data augmentation techniques particularly traditional and Wasserstein based data augmentation techniques, were implemented. Further, the CNN-based classifier was utilized for classifying knee MRI scans. The classification was performed through transfer learning, considering computational resources and time.

The project was disrupted due to unpredicted personal circumstances for one month. Following this, a deadline extension was requested to accomplish all the procedures mentioned in the work plan.

1.6 Structure of the report

Section 2 gives a brief overview of image classification and data augmentation techniques, along with describes the research studies that have been completed on this subject. Besides, this section also illustrates the certain researches performed on the dataset used in our study.

Section 3 comprehensively describes the techniques operated in this project, as well as indicates the tools and software used to build deep learning models. The details about data acquisition and data analysis are also discussed. Moreover, the hyperparameter settings used to train deep learning models and evaluation metrics to evaluate the deep learning model performance are communicated.

Section 4 discusses the results of data augmentation and classification, both qualitatively and quantitatively. For qualitative analysis of results, performance plots are visualized, and evaluation metrics are explored for quantitative analysis.

Section 5 explains the finding with respect to the objectives of the project and summarises the knowledge acquired during the project.

Section 6 provides a brief overview of each phase of the project and suggests the techniques that can be implemented in the future to improve our research.

Section 7 gives the list of references for research papers reported in this project.

Section 8 provided all the necessary supporting documents which will assist the reader of the report.

2 Chapter 2: Critical Context (15%) (3,000)

2.1 Deep learning in medical image diagnosis

In image classification, labeled data is fed into a deep learning model, and it is asked to provide correct labels for unseen dataset based on features learned in the seen dataset (Lundervold and Lundervold, 2019). Due to the advancements in technology, an enormous amount of healthcare data is collected through means of medical devices and instruments, health-based internet of things (IoT), wearable sensors, and smartphones. The increase in the number of medical datasets has enabled the use of deep learning techniques in medicine. In academic research, deep learning algorithms have been widely applied in healthcare data analysis due to their potential to learn complex patterns in large datasets that can assist in disease diagnosis, prognosis, and treatment (Cao et al., 2018).

Convolutional neural networks (CNN) have been widely used for the image classification task. CNN has the potential to outperform classical machine learning methods due to their competency of learning data representations automatically instead of using handcrafted features. Feature extraction and classification are done as one problem in deep neural networks, which improve automatically during the training process (Lundervold and Lundervold, 2019). Besides, Deep CNN architectures performed astonishingly good in many other computer vision tasks such as image segmentation and object detection (Shorten and Khoshgoftaar, 2019).

Many deep and flexible CNN-based architectures, including AlexNet, VGG16, ResNet50, etc., have been introduced to perform feature extraction and supervised learning on the large-scale dataset (Gao, Jiang, Zhou and Chen, 2019). Classifying medical images using deep learning algorithms is a challenging task. As deep learning-based classifiers require a large number of labeled data for training and manually labelling medical images is complicated and necessitates professional experts. Moreover, collecting medical data is greatly aligned with privacy. Furthermore, the medical datasets mostly face the issue of class imbalance because the samples for the normal class are much higher than samples of pathological class (Cao et al., 2018). The deep learning model trained with the imbalance and limited size dataset is prone to overfitting and biased performance. In several studies, transfer learning has been applied to perform medical image classification due to limited clinical datasets. Transfer learning uses deep model architectures, which are trained on large-size datasets to perform feature extraction or classification on small-size datasets. For feature extraction using transfer learning, the weights of the last layers of the pre-trained model are preserved while the weights of the top layers are trained. On the other hand, the last layer's weights are trained while the weights of the top layers are held to perform classification using transfer learning. Furthermore, many other techniques, such as regularisation, weight balancing, and data augmentation, have been employed to

improve the classifier's generalization performance with small size and imbalance data (Shorten and Khoshgoftaar, 2019).

2.1.1 An overview of the research studies in medical image diagnosis

A summary of deep learning techniques that have been applied in recent studies for medical image diagnosis is listed below.

(Al-Dhabyani, Gomaa, Khaled and Fahmy, 2019) built a computer-aided diagnosis system by adopting deep learning techniques to classify breast masses based on ultrasound images. Two breast ultrasound image datasets were used in which the first dataset had 780 images of size 500×500 distributed among three classes, namely normal, malignant and benign, and the second dataset was composed of 163 examples of size 760×570 for two classes benign and malignant. Deep convolutional neural network based on AlexNet architecture, as shown in Figure 4 was built from scratch along with different pre-trained DCNN architecture VGG16, ResNet, Inception, and NASNet were employed to classify breast lesions. Due to the limited size of the dataset, data augmentation was performed to increase the dataset's size to enhance the performance of the classification model. Traditional and GAN based data augmentation techniques were applied to increase the size of the dataset. The geometric transformations which were applied on the breast ultrasound dataset were scaling, zooming, horizontal flipping, and brightness. Further, Wasserstein GAN, with a gradient penalty, was employed in order to mitigate the issue of unstable GAN training. The convolutional neural network-based generator and discriminator architectures were utilized. The GAN model was trained separately for each class, and 5000 generated images of each class were added to the dataset. The classification performance using CNN- AlexNet and pre-trained models VGG16, Inception, ResNet, and NASNet was assessed with non augmented data, geometrically augmented data, GAN based augmented data, and both geometrically and GAN based augmented data. Moreover, regularisation techniques such as dropout and normalization were considered to achieve the best performance. The training was performed for 60 epochs. In transfer learning, the pre-trained models VGG16, Inception, ResNet, and NASNet were fine-tuned by altering the last layer to classify breast ultrasound images into three categories normal, benign, and malignant. The pre-trained models were trained for ten epochs. The accuracy was used as a performance metric. The accuracy achieved with a large size dataset was higher than a small size dataset. Also, the performance of classifiers with GAN based augmented data was higher as opposed to geometrically augmented data. The highest accuracy, 99%, was achieved through transfer learning with (NASNet) using traditional and GAN-based augmented data.

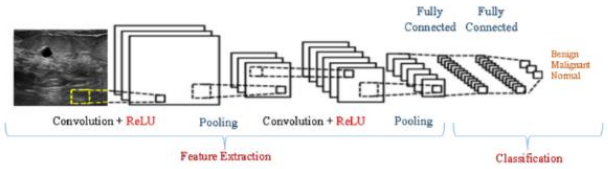


Figure 4: CNN-AlexNet architecture (Al-Dhabyani, Gomaa, Khaled and Fahmy, 2019)

(Rashid et al., 2019) employed GAN for both augmentation and classification of skin lesions. This study was performed upon International Skin Imaging Collaboration (ISIC) 2018 challenge dataset, which consists of images associated with seven categories of skin lesions. Data augmentation techniques were used to balance the distribution of the classes. The images generated by using GAN's generator network, real images, and geometrically augmented images were fed into a discriminator network. For geometrical augmentation, the transformations such as random cropping, Gaussian blurring, the addition of salt & pepper noise were applied. The discriminator network in the GAN model was a convolutional neural network that yielded the vector of size (number of target classes + 1). The discriminator network distinguished real data from fake data along with categorized real data into seven skin lesion categories. Figure 5 illustrates the GAN architecture used for data augmentation and classification. Moreover, DenseNet and ResNet50 were fine-tuned using only real and geometrical augmented data. The average accuracy was calculated to compare the GAN-based classifier and transfer learning performance with DenseNet and ResNet50. The GAN based classifier achieved the highest accuracy score of 86% compared to DenseNet (81%) and ResNet50(79%).

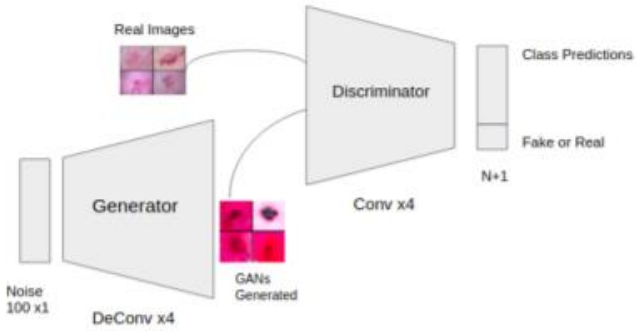


Figure 5: GAN for both classificiation and data augmentation (Rashid et al., 2019)

(Hon et al., 2017) investigated the effectiveness of training a convolutional neural network from scratch and transfer learning to classify Alzheimer's disease (AD) patients and Healthy Control (HC) patients. The MRI dataset was acquired from the OASIS website and contained 240 subjects; among them, only 200 were selected (100 AD and 100 HC). The images that provide the most relevant information were selected using the entropy-based method to train convolutional neural networks. The VGG16 model was trained from scratch along with pre-trained VGG16, and Inception V4 models were fine-tuned by training the weights of the output layer. In order to evaluate the model's performance, average accuracy from 5 fold cross-validation was examined. The VGG16 model, which was trained from scratch, achieved an average accuracy of 74.12%, while pre-trained models VGG16 and Inception V4 attained the average accuracy of 92.3% and 96.25%, respectively. The potential of transfer learning to achieve good diagnostic results with a smaller size dataset can be witnessed from this study.

2.2 Data augmentation in medical images

Data augmentation is one of the many techniques to alleviate the challenges related to medical image diagnosis. The data augmentation enlarges the size of the dataset by modifying or extracting information from the original data. The issues of viewpoint, lighting, occlusion, and background make image classification a difficult task. Due to these challenges, the data augmentation techniques are applied on small size datasets as well as large size datasets to increase diversity in the dataset that helps to enhance the robustness and generalization performance of the classification model. Furthermore, data augmentation has been considered to oversample minority class distribution in order to prevent the biased performance of the classification model. There are different types of data augmentation techniques such as geometric transformations, colour space augmentation, generative adversarial networks, neural style transfer, and-so-on, which have been utilized to increase the size and quality of the training data. Figure 6 depicts a categorization of image data augmentation, and colour lines indicate which sort of data augmentation method is used by the subsequent meta-learning system (Shorten and Khoshgoftaar, 2019).

Figure 6: A taxonomy system of the data augmentation (Shorten and Khoshgoftaar,

There are many techniques to perform data augmentation on medical dataset. This project mainly focuses on two data augmentation techniques: geometric or classical data augmentation and GAN based data augmentation.

2.2.1 Geometric data augmentation

Applying geometric transformations on images is described as geometric data augmentation. This data augmentation technique is easy to implement. In general, geometric data augmentation is a label preserving technique but not in some cases, for instance, digit classification task in which applying horizontal and vertical flips alters label 6 to 9. After applying geometric transformations on images, the transformed images need to be observed manually in order to make sure their labels have not been altered.

Positional biases in training data are best dealt with by applying geometric transformations. For instance, changing the colour of images helps to overcome the lighting issues experienced with test data. Different types of transformations such as Increasing or decreasing the brightness of images, rotation translation, random cropping, limiting pixel values to the fixed range, and noise injection can be applied to images to assist classifiers in learning more robust features. Combining different classical data augmentations results in large size of the dataset.

2.2.2 GAN based data augmentation

Generative adversarial network (GAN) is an unsupervised algorithm that was introduced by Ian Goodfellow in 2014. The objective of the GAN model is to generate new data that has a similar distribution to training data. GAN based generated images assist in increasing the size of the dataset and improving the diversity in the dataset, in addition, facilitate to enhance the performance of deep learning algorithms. The GAN model consists of two networks, generator and discriminator. The training of the GAN model functions through an adversarial mechanism. The generator network generates artificial images, and the discriminator network discriminates among real and fake images. During GAN training, the generator network learns to generate more realistic images to subside the discriminator's capability to distinguish between real and synthetic images. The GAN training is prone to instability and faces vanishing gradient and mode collapse, contributing to invariability in generated data (Goodfellow et al., 2014). The framework of GAN is shown in Figure 7.

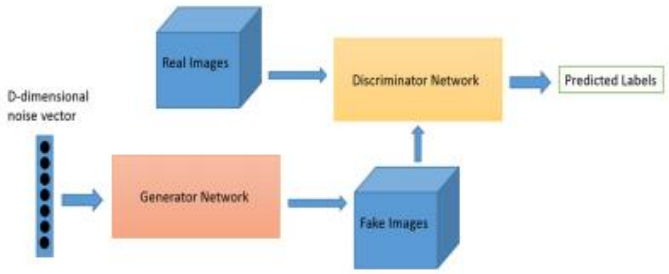


Figure 7: The Framework of GAN (Alqahtani et al., 2019)

Moreover, Jensen Shannon (JS) divergence is used by original GAN architecture to measure the association between real and fake distribution, which is not appropriate for calculating the association between mutually exclusive distributions. In order to make GAN training stable and improve the quality of generated samples, many variants of GAN have been introduced since 2014 by operating several modifications to original GAN architecture, for instance, various generator and discriminator architectures, loss functions, optimizers, and evaluation metrics to estimate the difference between real and fake data distribution (Lan et al., 2020). DCGAN, WGAN, CycleGAN, Progressive, growing GAN (PGGAN), and Super-resolution GAN (SRGAN) are few GAN variants.

In medical image data, the biases which discriminate training data from test data are complex as opposed to positional changes. It follows that applying the geometric transformation to increase data size and balance the class ratio is not a very effective approach in medical image analysis. Therefore,

deep learning algorithms, particularly the GAN model, has been considered a viable option in medical image synthesis to overcome the challenges of class imbalance and limited labeled training dataset (Shorten and Khoshgoftaar, 2019).In recent researches, GAN has attained so much attention in the task of data augmentation. Many studies have used GAN based data augmentation to improve medical image diagnosis (Lundervold and Lundervold, 2019).

2.2.3 Summary of the research studies performed data augmentation in medical imaging

Data augmentation methods are only useful if test and training datasets are from the same distribution. In other words, when training data has poor diversity with respect to test data, then data augmentation is not beneficial (Shorten and Khoshgoftaar, 2019). In some cases, data augmentation leads to poor diversity in the dataset that contributes to a decrease in classification performance. This section briefly outlines the studies related to both positive and negative influences of data augmentation on the performance of the classification model.

(Frid-Adar et al., 2018) utilized GAN based data augmentation in Liver lesion classification and observed significant improvement in classifier performance as compared to classification performance with conventional data augmentation. The classification model was increased from 78.5% to 85.7%. The GAN based data augmentation is not only applied in the classification task but also in object detection and segmentation.

(Changhee et al., 2018) used two GAN variants DCGAN and WGAN, to generate brain MRI images generated using different MRI sequences (T1, T2, and FLAIR). The dataset used for this study was the Brain Tumour Image Segmentation (BRATS) challenge dataset, which consisted of 270 MRI images of dimension 240×240×155 categorized into High-Grade Glioma and Low-Grade Glioma. Prior to GAN model training, the images were resized to 64×64 and 128×128 from 240×155, and 2D slices in the range of 80 and 149 were chosen among 240 slices in MRI scans. For the purpose of evaluating the GAN performance, an expert physician was invited to classify real and synthetic images among randomly chosen 50 real and generated images. The images generated by the WGAN model were more realistic as opposed to DCGAN. Also, the expert physician found it challenging to differentiate between real images and images generated by WGAN for all sequences, excluding FLAIR.

(Delplace, A., 2020) generated brain MRI images using three GAN variants DCGAN, ProGAN, and SRResNet. The performance of three GAN architecture was assessed with different loss functions (original Jensen Shannon divergence, LSGAN, WGAN, WGAN-GP, DRAGAN) and hyperparameters in order to examine the impact of hyperparameters on GAN training. The dataset used in this study was taken from the Open Access Series of Imaging Studies (OASIS) and contained

11328 brain MRI comprising 2D image slices of size 256×256. Image realism and diversity were calculated to evaluate the performance of GAN models. The diversity was estimated based on the total variation of generated images and the number of eigenvalues of the covariance matrix, which are greater than 1% of the total variation. The stable GAN training was achieved with WGAN GP and DRAGAN loss functions. It is observed from this study that adjusting the gradient norms helps to stabilize the GAN training.

(Kora, et al., 2020) used DCGAN to generate X-ray images in order to mitigate the issue of an imbalanced dataset. The X-ray image dataset was categorized into normal (1341) and Pneumonia (3875) cases. The normal X-ray images were augmented by using DCGAN. Fréchet Distance of Inception (FID) score was measured to evaluate the performance of the GAN model. FID score is the measurement of similarity between real and generated images. The lower value of the FID score indicates high GAN performance. The FID score of 1.289 was achieved by the DCGAN model.

(Wang et al., 2020) compared the performance of the CNN based classification model with data generated using different GAN variants, particularly WGAN- GP, PGGAN and pix2pix. The dataset used in this was obtained from the Shanghai Public Health Clinical Center and Zhongshan Hospital Affiliated to Fudan University which consists of 206 nodules regarding three pathologies, particularly adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IAC). Instead of passing 3D nodules into the classification model, the author opted to give 2D image slices in each nodule to the 2D classification model. Different common data augmentation techniques such as translation, rotation, flipping were employed on the 2D images. Next, WGAN, PGGAN and pix2pix GAN models were trained using the common augmented dataset to synthesise nodule images. The images generated by WGAN-GP and pix2pix GAN were of low quality and blurry while PGGAN contributed to the generation of high-quality and realistic samples. For each class, 10,000 images were generated by each GAN model. The area under the receiver operating curve (AUC) was used to compare the classification model's performance with different types of augmented dataset. The AUC achieved by classification model using non augmented dataset, common augmented dataset, WGAN GP based augmented dataset, pix2pix augmented data and PGGAN based augmented dataset is 60%, 78%, 73%, 75% and 83% respectively. In this study, the data augmentation based on WGAN GP and pix2pix GAN contributes to low performance than data augmented using traditional data augmentation techniques. It can be summarised from this study that GAN based augmentation is only useful when generated images are of high quality and realistic.

(Ganesan et al., 2018) assessed the effectiveness of traditional data augmentation techniques and GAN based augmentation for classifying abnormal chest X-ray images. The GAN based data augmentation was performed using PGGAN. A pre-trained VGG 16 model was used for classification. For traditional data augmentation, Gaussian smoothing, unsharp masking, and

minimum filtering were applied on the chest X-ray images. For each class, 6268 images were generated using PGGAN. Accuracy, F score, AUC and Matthew's correlation coefficient was utilised to investigate the performance of the classifier. The performance achieved with traditional data augmentation was slightly higher than the GAN based augmentation. This research demonstrates that GAN based data augmentation is not always superior to traditional data augmentation techniques.

(Hussain et al., 2018) conducted a study to find the augmentation techniques that portray the statistics of medical imaging to the highest degree. For this purpose, different geometrical data augmentation techniques were applied to Digital Database for Screening Mammography (DDSM) comprising 1650 mass cases and 1651 non-mass cases. The convolutional neural network-based architecture VGG 16 was utilized to classify mass and non-mass cases. The model was trained using 80% of original data and tested on 20 % dataset that was held out during training. The similarity between real and generated images was observed by calculating Mutual information. Different geometric transformations Gaussian noise, gaussian filtering, jittering, scale, powers, Gaussian blur, rotations and blur were applied to the training dataset. In total, 8 VGG16 models were trained for 8 different types of traditional data augmentation strategies. Accuracy was computed to assess the performance of the model. It was observed that the data augmentation techniques which gave higher mutual information between generated and real images contributed to an increase in the validation accuracy. Notably, flipping and Gaussian filters contributed to higher validation accuracy while noise and powers gave low validation accuracy. It can be summarised from this study that augmented images contribute to higher classification performance if they greatly portray the statistics of real images which indicates a strong association between generative learning and classification performance.

(Wu et al., 2018) used traditional and GAN based data augmentation to enhance the performance of ResNet 50 classifier to detect malignant and non-malignant cases in the Digital Database for Screening Mammography. Conditional infilling GAN architecture was utilised to synthesis malignant and non-malignant images. The Area under the receiver operating curve was utilised to assess the performance of the classification model. The traditional data augmentation techniques such as Random rotation, horizontal flipping and rescaling were applied on the dataset. The ResNet 50 model's performance achieved with no augmented data set, traditional based augmented dataset and GAN based augmented dataset is 88.2%, 88.7%, 89.6%, respectively. From the result of this study, it can be analysed that GAN based data augmentation only increased the performance by 0.014% compared to performance with non augmented dataset. This study witnessed insignificant improvement in the performance of the classification using GAN based data augmentation.

2.3 Researches practiced by academics on the MRNet dataset

In this section, the previous work that has been accomplished on the MRnet dataset is briefly described.

(Bien et al., 2018) proposed deep learning algorithm MRNet to diagnose knee abnormalities to be specific general abnormalities, ACL tear and meniscus tear in the MRNet dataset comprising MRI scans of the knee in three planes. The MRNet model is a convolutional neural network-based architecture. The convolutional layers of the pre-train AlexNet model were utilized for feature extraction in MRNet architecture. After that, the extracted features were passed through the average pooling layer, max-pooling layer, and fully connected layers to get a predicted value which indicates the presence of knee disorder. Furthermore, three types of geometric transformations such as random horizontal flip by probability 50%, random shift between -25 and 25 pixels, and rotation between -25 and 25 degrees were applied on the dataset. In order to alleviate the imbalanced class distribution, the loss of each exam was scaled inversely proportional to the prevalence of corresponding class distribution in the dataset. The MRNet architecture was trained for each class (abnormal, ACL tear, meniscus tear) and plane (axial, sagittal, coronal) individually, yielding in 9 MRNet models. Each MRNet model was trained for 50 epochs. The output obtained from three planes (axial, sagittal, and coronal) was combined using Logistic regression in order to get a single probability value for each task (abnormal, ACL, meniscus). The area under the receiver operating curve (AUC) was used as a performance metric along with the specificity, sensitivity, and accuracy of 9 medical experts were assessed. The model achieved AUC of 0.937, 0.965, and 0.847 in detecting abnormality, ACL tear, and meniscus tear, respectively, on the internal validation set. The model's performance was also assessed on an external validation dataset containing MRI scans of the knee in the sagittal plane labeled for ACL injuries. The AUC achieved by the MRNet model in detecting ACL injuries on 183 exams of external validation set was 0.824. Moreover, With model assistance, the radiologists attained higher sensitivity and higher specificity in detecting meniscus and ACL tear on the internal validation dataset.

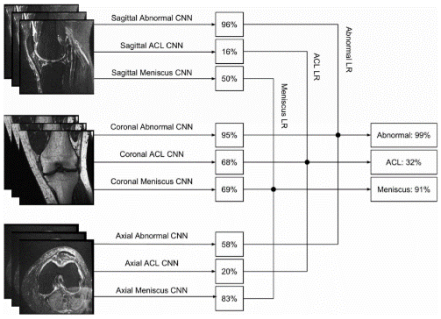


Figure 8: Combining results obtained from MRNet models using logistic regression

(Irmakci et al., 2020) compared and evaluate the performance of the MRNet model using features extracted from GoogLeNet, AlexNet, and ResNet18 architectures. The AUC, specificity, sensitivity, and accuracy were used as performance metrics to evaluate models' performance. The highest AUC for the abnormal class was 0.91% using GoogLeNet and 95% for ACL using ResNet 18, and 81% for meniscus tear using both ALexNet and ResNet 18. Overall, Resnet18 surpasses the AlexNet and

GoogLeNet in all performance parameters AUC (0.8579), sensitivity (0.7936), and accuracy (0.8167) except specificity.

This project is the extension of the existing work that has been done on this dataset and aims to apply different types of geometric and GAN based data augmentation techniques on the MRNet dataset in order to analyze the performance of the MRNet model with different types of augmented data. Due to limited time and computational resources, the MRNet model was trained using only axial view MRI scans. To the best of our knowledge, GAN based data augmentation has not been performed on this dataset. Moreover, untried geometric transformations will be applied to MRNet dataset in this research. The performance of MRNet architecture will be assessed using without augmented data, geometrically augmented data, GAN based augmented data, and both geometric and GAN based augmented data.

3 Chapter 3: Methods (20%)

3.1 Tools and Software

In this project, the Python 3.7.6 programming language was utilized as the language of choice due to its feature-rich libraries within the realms of the machine and deep learning. The deep learning library Pytorch version 1.2.0 was used, alongside Cuda toolkit 10.0, in order to build and train neural network models on GPU's. The library scikit-learn was also utilized mainly for calculating performance metrics and plotting ROC curves and confusion matrices. Tesnsorboard was also used primarily to visualize and track performance plots during the classification and GAN models training. Furthermore, the image manipulation library PIL was incorporated, as well as Matplotlib, Numpy, and OpenCV. In addition, Docker files were created alongside other set-up files, such that machine learning jobs could be submitted and trained on the google cloud platform via the 'ai-platform'.Note that the training of models on the cloud was not employed due to complexity and time constraints; thus, models were trained on a Desktop with an AMD Ryzen 9 3900 12 Core Processor, 16 GB of RAM, and NVIDIA GeForce RTX 2070 8GB graphics card. Spyder is the Integrated Development Environment utilized to code and run the neural network models.

3.2 Data Acquisition

The dataset used in this research is the MRNet dataset, which is taken from the web page of the Stanford machine learning group. The dataset consists of knee MRI scans which were operated in three planes of space axial, sagittal, and coronal with weightings T1, T2 with fat saturation, proton density (PD), PD with fat saturation. The knee MRI exams were taken at Stanford University Medical Centre between January 1, 2001, and December 31, 2012. Each MRI scan is comprised of a series of 2D slices. The number of 2D slices in each series varies between 17 and 61 over MRI scans. Overall, 775 MRI scans are taken in an MRI scanner of a 3.0-tesla magnetic field, and all other exams are performed using a magnetic field of strength 1.5 teslas. Figure 9 illustrates the MRI scans in axial, sagittal, and coronal planes. In this research, only axial plan views were utilized to reduce computation cost.

Table 1: knee MRI scans weightings across three planes

Plane	Weightings
axial	PD weighted with fat saturation
Coronal	T1, T2 with fat saturation
Sagittal	PD, T2 with fat saturation

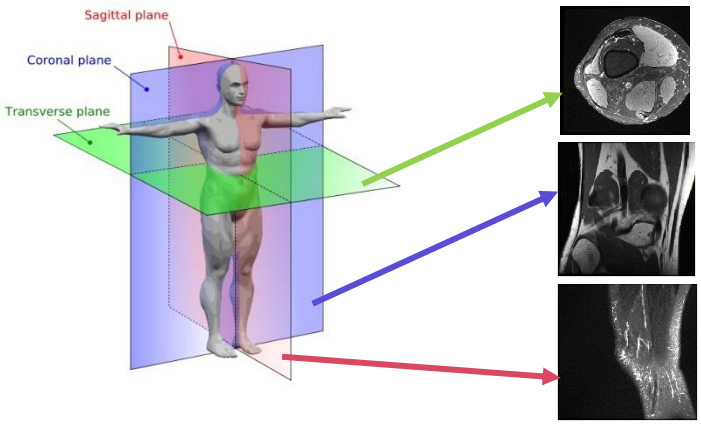


Figure 9: Knee MRI in three planes

3.3 Data Analysis

The knee MRI dataset consists of 1,250 MRI exams in which 1,104 exams are abnormal, together with 319 ACL tears and 508 meniscus tears. The labels of the knee MRI dataset are attained through manual extraction from medical reports. The dataset is divided into training and test sets. The training dataset consists of 1130 MRI exams, and 120 exams are present in the test set. Random sampling was applied to make sure that the validation dataset has a minimum of 50 positive exams for each knee disorder (general abnormalities, meniscus tear, and ACL tear).

The training dataset is categorized into 5 asynchronous categories: abnormal, abnormal with ACL tears, abnormal with meniscus tears, abnormal with both ACL and meniscus tears, and no abnormalities comprising 433, 83, 272, 125, and 217 exams, respectively. Following this, the validation dataset has 20, 21, 23, 31, and 25 cases for abnormal, abnormal with meniscus tears, abnormal with ACL tears, abnormal with both ACL and meniscus tears, and no knee disorder categories, respectively. Figure 10 illustrates the distribution of data concerning five individual categories in the training and validation datasets. It has been examined that abnormality is the class with the highest number of samples. Moreover, the number of cases with either meniscal tear or ACL tear must have abnormalities, and some have all knee disorders, including abnormality, meniscus, and ACL tears. The training and validation cases have three binary labels indicating the presence or absence of abnormality, ACL, and meniscus tear, respectively. The tasks of classifying abnormal and normal exams, ACL tear and normal exams, Meniscus tear, and normal exams were considered as a separate binary classification problem. In the future, It would be interesting to classify the Knee MRI dataset into five separate classes Normal, abnormal, abnormal with a meniscus tear, abnormal with an ACL tear, and abnormal with both meniscus and ACL tear. Furthermore, the knee MRI dataset has an imbalanced class distribution that needs to be addressed to ensure that the classification model is not biased towards the majority class.

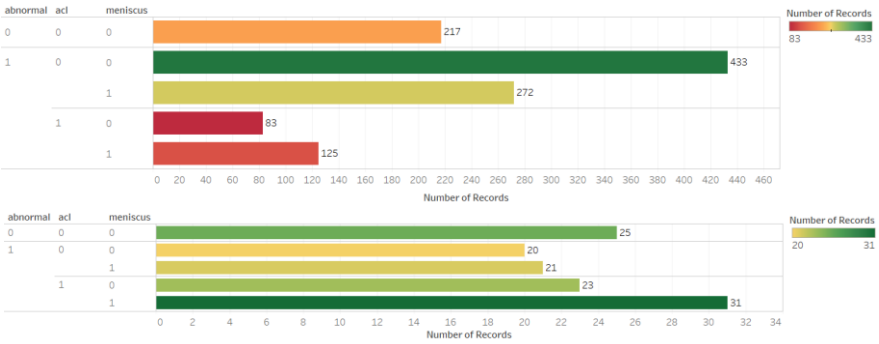


Figure 10: Distribution of data in training and validation dataset respectively

Prior to building deep learning models, pre-processing was performed. The knee MRI dataset is saved in NumPy format and has been standardized using Histogram based Intensity standardization. The pixel values of images range between 0 and 255. The size of the MRI scans is (3, s, 256, 256), where s is the number of 2D image slices. In the pre-processing step, the pixel values were changed to float32, and pixel intensities of image slices were scaled in the range [0, 1] to make computation more efficient. Figure 11 displays the knee MRI images after normalization.

Figure 11: knee MRI image slices after normalizing pixels in range [0,1]

3.4 Data Augmentation techniques

Different types of data augmentations were performed on the training dataset containing 1130 exams to increase the diversity and enlarge the training data size. This section describes the Data augmentation techniques that are applied to the training data.

3.4.1 Traditional Data Augmentation Techniques

For traditional data augmentation, two types of transformations, such as affine transformations and pixel-level transformations, were applied on 2D image slices in knee MRI series.

3.4.1.1 Affine transformations

Applying transformations such as rotation, cropping, zooming, translation, and flips are affine approaches to augment training data. The images generated by affine transformations do not facilitate to increase diversity in the training dataset and only contribute to a slight improvement in the performance of deep learning algorithms. Moreover, affine transformations are easy to implement and have been extensively applied in research studies. The affine transformations which are applied in this research are briefly described below.

3.4.1.2 Rotation

Rotating image by angle α over-centre pixels is referred to as rotation. The 2D slices in the knee MRI series were rotated between -20 and 20 degrees.

3.4.1.3 Translate

The translation is applied in order to shift the image position in a selected direction (up, down, left, and right) whilst preserving the spatial dimension of the image. Applying translation by a certain number of pixels in a chosen direction helps the model to learn spatially invariant features. The images in the knee MRI series are translated by [0.1,0.1].

3.4.1.4 Horizontal flips

Horizontal and vertical flips produce mirror reflection of an image along the X and Y axis, respectively. The 2D image slices in knee MRI scans were flipped horizontally by a probability of 50%.

3.4.1.5 Pixel level transformation

In pixel-level transformation, the geometric shape of images is not altered, but pixel intensities are transformed. The pixel-level transformations are helpful in mitigating the influence of images taken using different scanners. Injecting gaussian noise, blurring, modifying the brightness of images are few pixel-level operations (Nalepa et al., 2019). Brightness and contrast were applied to the images in the knee MRI series.

3.4.1.6 Brightness and contrast

Applying the brightness and contrast changes the pixel values instead of pixel images. The overall darkness or lightness of the image is referred to as brightness, and contrast is the degree to which light and dark colour differ in an image. In order to mitigate the influence of different scanner settings, brightness and contrast are modified. The brightness and contrast of the 2D knee images are altered by 0.1 and 0.3 factor, respectively.

Pytorch library was used to implement geometric data augmentation, which applies transformations on the batch of images randomly during the training of the model.

3.4.2 GAN Model for data augmentation

Due to the concurrent occurrences of classes abnormal, meniscus and ACL tears in the training dataset, the GAN based data augmentation was performed to generate cases having abnormal with an ACL tear, abnormal with a meniscus tear, and no abnormalities. All the cases having labels (abnormal:1, ACL tear:1, meniscus:0), (abnormal:1, ACL tear:0, meniscus tear: 1) and (abnormal:0, ACL tear:0, meniscus:0) in the training dataset were filtered to assist WGAN based data augmentation. The intuition behind choosing only these three cases was that all the meniscus and ACL tear cases must have an abnormality, so generating images for both ACL and meniscus cases

will automatically increase the size of abnormal cases. Further, the cases with no abnormalities were generated to avoid the considerable difference in each class's data distribution.

For the development of GAN architecture to generate realistic knee MRI scans, one of GAN's variants, Wasserstein GAN, was selected due to its ability to generate high-quality images and insufficiency in unstable GAN training. The original WGAN architecture proposed in (Arjovsky et al. 2017) can be directed to sub-standard behaviour due to weight clipping restriction on the critic network. In order to alleviate this issue, (Gulrajani et al. 2017) proposed WGAN with a gradient penalty (WGAN-GP), which includes gradient penalty term in critic loss function and is capable of producing high-quality results without any hyperparameter tuning. The Wasserstein GAN with gradient penalty, which is a modified version of the original WGAN model, was utilized in this research.

3.4.2.1 Wasserstein GAN

Wasserstein GAN (WGAN) was proposed in (Arjovsky et al. 2017). The motivation behind WGAN was to find ways to measure the degree of association between fake and real data distribution. The distance metric is used as a cost function. The cost function in the WGAN network is formulated on earth's mover or Wasserstein-1 distance. The original GAN architecture uses JS divergence as a cost function, while WGAN uses the earth's mover distance to measure the difference between real and fake data distributions. The earth mover's distance is described as

$$EM(P,Q) = \ inf_{\gamma \in \prod(P,Q)} E_{(u,v)\in\gamma}(\|u - v\|)$$

The earth mover distance requires minimum effort to change one distribution to another. $\prod(P,Q)$ is the collection of all joint distributions which have marginal P or Q and $\gamma(P,Q)$ is known as a transport plan that determines the approach to reallocate the earth from point u to v in order to convert the distribution P to distribution Q. The transport plan $\gamma(P,Q)$ is optimized, which makes EM distance an optimization problem. Moreover, the Wasserstein or earth's mover distance is continuous and differentiable.

In original GAN architecture, the discriminator is a classifier. In contrast, the discriminator appears as a critic in WGAN and returns a scalar value that indicates the realness of the generated images. The higher critic score suggests the synthesis of more realistic images. In the WGAN training, the critic is trained more than the generator. The longer critic training leads to reliable gradients in WGAN, which lessens the chances of mode collapse. Furthermore, the quality of generated images correlates with the loss of GAN during training in WGAN. With different generator architectures, WGAN is much more robust than the original GAN architecture (Arjovsky et al. 2017).

In WGAN, the critic function must satisfy the 1-Lipshitz property throughout the training process. In other words, the critic function f must fulfill $|f(x_1) - f(x_2)| \leq |x_1 - x_2|$.In order to ensure the 1-Lipshitz property of the critic, (Arjovsky et al. 2017) suggested weight clipping. Further, (Gulrajani et al. 2017) indicated that weight clipping could cause vanishing gradients and limit the learning of critic networks that contribute to learning simple function instead of complex function. In addition, the interaction between weight constraint and the cost function makes the optimization process of WGAN difficult. Due to the absence of accurate tuning of the clipping threshold, the problem of vanishing or exploding gradient can be experienced.

For the purpose of mitigating the challenges of weight clipping (Gulrajani et al. 2017) proposed a WGAN-gradient penalty (WGAN-GP) that penalizes the gradient norms of the critic regarding its input. Weight clipping is replaced by the gradient penalty to necessitate the 1-Lipshitz property in the WGAN-GP (Gulrajani et al. 2017).

3.4.2.2 WGAN-GP using different generator and critic Architectures

WGAN-GP model was trained with two architectures of generator and critic networks as follows:

- Convolutional network-based generator and critic
- Residual network-based generator and convolutional network-based critic

Initially, critic and generator architectures were experimented with to generate images of resolution 256×256, but memory errors were experienced during WGAN training due to limited computational resources. The only way to resolve memory issues was by reducing the batch size to 15 during training, which requires a significant amount of time to train. In order to make these experiments computationally efficient, the images downsampled to 64×64 resolution. Different generator and critic architecture were tried to generate images of resolution 64×64. The architecture, which contributed to the generation of high quality, and realistic 64×64 images, was selected to generate images of resolution 256×256.

3.4.2.2.1 Generator and Critic architecture for 64×64 images

3.4.2.2.1.1 Convolutional neural network-based generator

The generator was established on four 2D transposed convolutional layers, which are also referred to as deconvolutional layers. The kernel of size 4×4 size kernel and stride of 2 was chosen for transposed convolutional layers. Batch Normalization was applied to all the deconvolutional layers of the generator network except the output layer. The batch normalization standardizes the input of layers that helps to accelerate and stabilize the GAN training. Moreover, applying batch normalization on convolutional layers helps to prevent vanishing gradients (Gulrajani et al. 2017). Furthermore, activation functions are used in convolutional neural networks to intensify non-linearity, allowing the neural network to learn complex data relationships. In deep learning studies, the ReLU activation function is widely used and considered as an efficient activation function. The ReLU function returns 0 if the input is less than 0; otherwise, it returns the provided input. Mathematically, it can be represented as $f(x) = \max(0, x)$. LeakyReLu is a variant of the ReLU function, which has a slope for negative values rather than having all zeros. The LeakyReLU is more balanced as opposed to the ReLU function (A Practical Guide to ReLU, 2020). In this study, LeakyReLU with a 0.2 slope value was utilized in all the generator layers except the output layer, which used a hyperbolic tangent function for activation. See Figure 12.

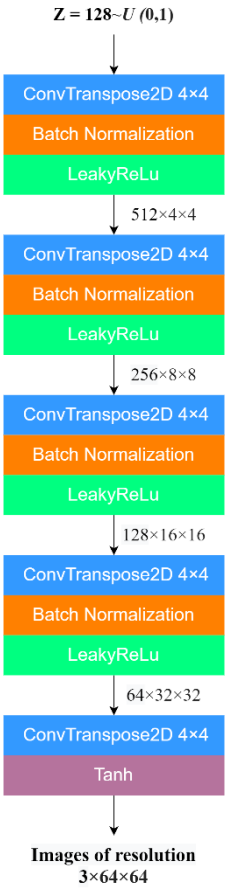


Figure 12: CNN based generator to generate images of resolution 64×64

3.4.2.2.1.2 Residual network-based generator

The generator architecture (Figure 13) is based on 101 layer ResNet, stimulated by (Gulrajani et al. 2017). The generator network's input layer is a fully connected layer whose output passes through four residual blocks, followed by two convolutional layers of filter size 3×3. The nearest neighbourhood upsampling is performed before the convolutional layers in residual blocks. Moreover, batch normalization and LeakyReLU with slope 0.2 are utilized in each convolutional layer of residual blocks. The generator network's output layer is a convolutional layer of kernel size 3×3 with a tan hyperbolic function for activation that gives the output in the range [-1,1].

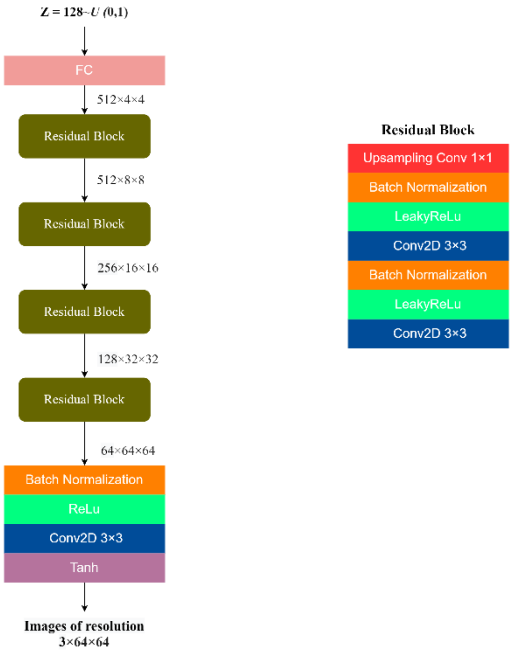


Figure 13: ResNet based generator architectureto generate images of 64×64

3.4.2.2.1.3 convolutional neural network-based critic

The critic network (Figure 14) is composed of four convolutional layers, which help to stabilize GAN training. The convolutional layers in the critic network use a 4×4 size kernel and stride of 2. Generally, batch normalization is applied to discriminator networks, but, in the case of WGAN-GP, layer normalization is suggested for the critic network by (Gulrajani et al. 2017). In batch normalization, the batch of inputs is mapped to the batch of outputs; however, the critic's objective function in WGAN GP penalizes the gradients of critic for each input individually. So, layer normalization, which maps single input to a single output, is recommended for the critic network. In the critic network, all the convolution layers, excluding the output layer, used layer normalization and LeakyReLu with slope 0.2 for activation.

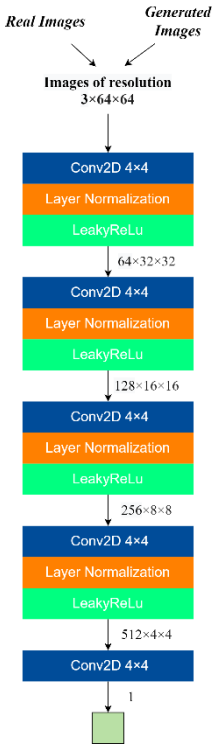
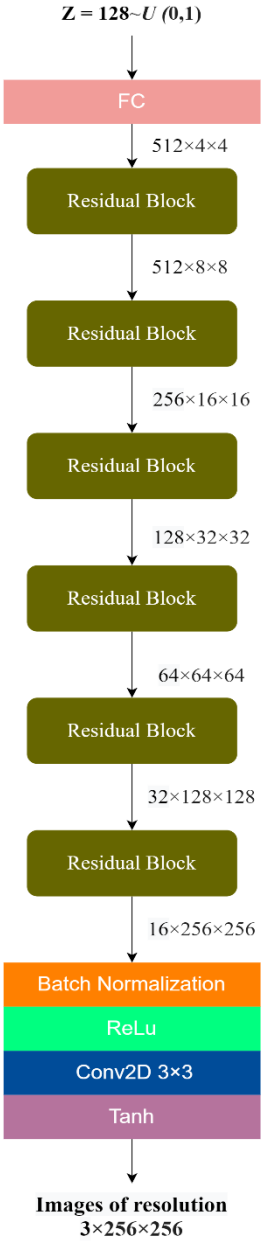


Figure 14: Critic architecture for 64×64 images

3.4.2.2.2 Generator and Critic architectures for 256×256 images

The high quality and realistic images of resolution 64×64 were generated by using ResNet based generator and CNN based critic network, which is explained in section 4. The ResNet based generator and CNN based critic, which were used to generate high-quality images of resolution 64×64, were modified by adding two residual block and convolutional layers, respectively, to generate 256 ×256 images. The ResNet-based generator and CNN-based critic architectures, which were used to generate 256×256 images, are illustrated in Figures 15 and 16, respectively.



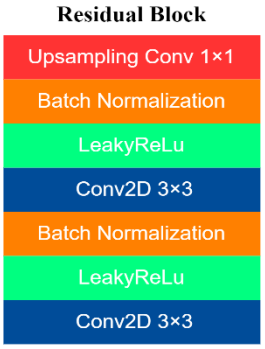


Figure 15: The Generator architecture to generate images of resolution 3×256×256

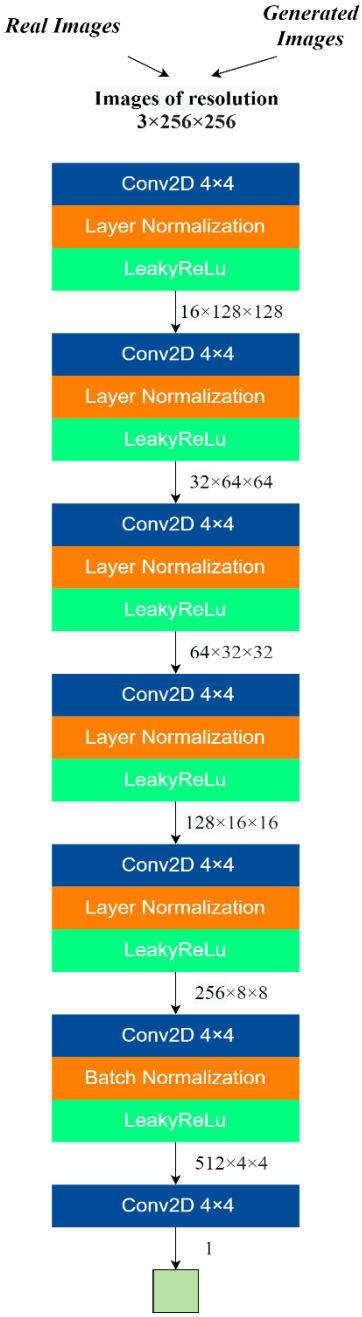


Figure 15: The critic architecture for 256 ×256

3.4.2.3 WGAN with different Hyperparameter settings

Before WGAN training, the weights of generator and critic networks were initialized using zero centred Gaussian distribution with 0 mean value and 0.02 standard deviation value to prevent exploding or vanishing gradients during training, which contributes to longer training time to reach convergence. Adam algorithm with learning rate 0.00005 was chosen for optimization, which computes the gradient's moving averages and the squared gradients. The decay rates of moving averages are controlled by parameters beta1 and beta2, respectively, with values of 0.5 and 0.999. The critic network was trained five times more than the generator network. The WGAN model is trained for 100,000 iterations. The value of the penalty coefficient was set to 10, which succeeded in WGAN-GP training (Gulrajani et al. 2017). A random vector of size 128 sampled from a uniform distribution was passed through the generator network.

Initially, the 2D images in the MRI series were resized from 256×256 dimensions to 64×64 dimensions to experiment with different generators and critic architectures. The WGAN model with two types of generator and the critic architectures was trained using the aforementioned hyperparameter setting to find the architecture that generates more realistic images. The hyperparameter set, which contributed to the synthesis of more realistic images, was carefully chosen. The number of slices within the MRI scans was considered as batch size varied between 17 and 61.

After finding the optimal generator and critic architectures, the WGAN model was trained to generate images of size 256×256×3. Different hyperparameter settings were tried to generate 256×256 images. The batch size was set to 15 due to memory errors. Different values for the learning rate [1e-4, 1e-5,5e-5] were tried.Further, the performance of WGAN was compared using two objective functions, particularly WGAN-GP and WGAN divergence. The objective functions WGAN-GP and WGAN divergence are briefly described below.

Parameters	Values
Image size	256×256
Critic iteration	5
Critic loss function	WGAN GP and WGAN div
Optimizer	Adam
Learning rate	[0.00001,0.0001, 0.00005]
(b1, b2)	(0.5,0.999)
Noise vector	128

WGAN GP objective function:

$$\boldsymbol{Wasserstein\ distance} = E_{x\sim P_{real}}[f(x)] - E_{x'\sim P_{generated}}[f(x')]$$

$$\boldsymbol{Gradient\ penalty} = k\ E_{x'\sim P_y}[(\|\nabla f(x^{\sim})\|_2 - 1)^2]$$

$$critic\ \ loss = \boldsymbol{Wasserstein\ distance} + \boldsymbol{Gradient\ penalty}$$

$$generator\ loss = -E_{x'\sim P_{generated}}[f(x')]$$

The Wasserstein distance estimates the difference between the critic score on real images and critic score on the generated images. The gradient penalty term is added in Wasserstein distance to fulfill the 1-Lipshitz constraint. If the gradient norms are no more than 1 then the 1-Lipshitz restriction is satisfied and vice versa. The gradient penalty is operated as the square difference from norm 1.

WGAN divergence objective function:

$$critic\ loss = \boldsymbol{Wasserstein\ distance} + k\ E_{x'\sim P_u}[\|\nabla f(x^{\sim})\|^P]$$

It is challenging to fulfill the restriction of k-Lipshitz. In order to ease this restriction, WGAN divergence objective function is introduced. The WGAN divergence does not necessitate 1-Lipshitz restriction because $\|\nabla f\|$ is not bounded by parameter k (Wu et al. 2018).

3.4.2.4 Training WGAN with optimal hyperparameter setting

Three WGAN models with optimal hyperparameter values were trained individually to generate samples for three categories (normal, abnormal with an ACL tear, and abnormal with a meniscus tear). Next, the trained WGAN models were used to generate MRI scans by stacking the generated images. The number of images per scan was dictated by the number of slices in the original dataset. Approximately 300 MRI scans, comprising around 10,000 image slices in total, were generated for each aforementioned category.

3.5 Classification model

3.5.1 MRNet Architecture

The MRNet model, which is proposed in (Bien et al., 2018), was utilized to classify knee abnormalities. The input of size s×3×256×256, where s is the number of 2D image slices in the MRI series, was passed into the MRNet model. Building a convolutional neural network from scratch requires a longer training time; therefore, feature extraction is performed using the transfer learning approach in the MRNet model. AlexNet based feature extractor, which is pre-trained on ImageNet database composed of 1.2 million images to classify 1000 classes, is fine-tuned to recognize features in the MRI dataset. Each 2D image slice passes through AlexNet based feature extractor to get a tensor of dimension s×256×7×7 comprising features for each slice. In other words, a set of 256 feature

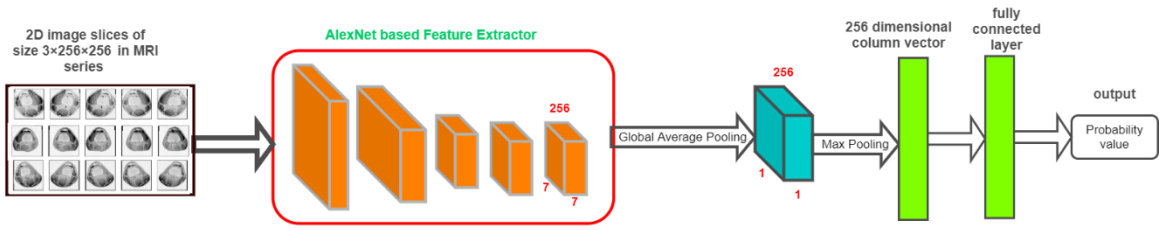


Figure 16: Classification model Architecture

maps of size 7×7 is attained from the last convolutional layer of AlexNet. Next, Global Average pooling is applied to transform s×256×7×7 to s×256 by taking the mean of each 7×7 square. Then, max-pooling is performed across slices to get the vector of dimension 256 from s×256 shaped tensor. The obtained vector of size 256 was passed through a fully connected layer to get the probability of having a corresponding knee disorder.

3.5.2 Choice of hyperparameters for classification model training

The training data was used to train the MRNet classification model. The classification loss was calculated using binary cross-entropy function. In order to minimize the loss value, the loss is backpropagated, and optimization is performed using Adam optimizer. For the purpose of alleviating class imbalance, the loss of each sample is scaled inversely proportional to the frequency of the sample's class in the dataset. The learning rate scheduler is used to fine-tune the learning rate during the classification model's training by reducing the learning rate according to the selected learning rate scheduler. The learning rate scheduler was set to reduce the learning rate by 0.3 if there is no improvement in the performance of the classifier for five epochs. Also, the weight decay is set to 0.01 to prevent the extreme enlargement of network weights. Further, weight decay is helpful to avoid overfitting. The hyperparameters values used to train the classification model are modelled from (Bien et al., 2018). Three MRNet models were trained individually using the aforementioned hyperparameter values for abnormal, ACL, and meniscus tears detection.

3.5.3 Investigating Classification performance with different types of augmented data

The classification model was trained with non augmented data, geometrically augmented data, WGAN based augmented data, and both geometrically and GAN based augmented data to analyse the impact of different data augmentation techniques on the performance of the classification model. Besides, the classification performance was assessed by adding the WGAN generated images for each category individually, which helped investigate the detailed effect of the WGAN based data augmentation on the classifier's performance.

3.6 Evaluation metrics

Evaluating the performance of deep learning models is a crucial phase. The metrics used to evaluate the performance of classification and Wasserstein GAN are briefly described below.

3.6.1 Metrics to assess the performance of the classification model

The classifier was trained using training data (1130), and its performance was assessed on test data (120). Accuracy, specificity, sensitivity, AUC, recall, precision, f score, and confusion matrix are considered to assess the performance of the classification model. Further, the training accuracy and loss curves are also plotted to examine the generalization performance of the classifier. The model

that has high performance on the training data and has poor performance on the test dataset is stated as an overfitting model, while the model that has poor performance on both training and test set is known as underfitted model. The classification metrics used in this research are briefly described below.

3.6.1.1 Accuracy

Accuracy is the measure of the percentage of correct predictions over the total number of cases.

$$accuracy = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ negative + false\ positive}$$

3.6.1.2 Specificity and Sensitivity

Sensitivity or recall estimates the ratio of the correctly predicted positive samples. On the other hand, the ratio of correctly classified negative samples is known as specificity.

$$Sensitivity = \frac{true\ positive}{true\ positive + false\ negative}$$

$$Specificity = \frac{true\ negative}{true\ negative + false\ positive}$$

3.6.1.3 Precision and f score

Precision is the ratio of correctly predicted positive samples over the total number of positive samples. The harmonic mean between precision and recall is specified as $f$score (M and M.N, 2015).

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

$$f\ score = \frac{2 \times precision \times recall}{precision + recall}$$

3.6.1.4 The area under the ROC curve (AUC)

AUC is considered a better metric than accuracy to assess the performance of the classifier. The AUC calculates the degree of separating positive samples from negative samples. The value of AUC varies between 0 and 1; besides, values near 1 depict the excellent performance of the classifier.

3.6.1.5 Receiver operating curve (ROC)

The receiver operating curve is plotted to visualize the performance of the classification model that determines the relationship between the false-positive rate and true positive rate. The false-positive rate is plotted along the x-axis and the true positive rate along the y-axis. The ROC curve is plotted at

different output thresholds (Receiver operating characteristic, 2020). In this project, the prediction threshold was set to 0.5.

3.6.1.6 Confusion matrix

A confusion matrix is a square matrix that gives the total number of true positive, false-positive, true negative, and false-negative.

True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)

3.6.2 Performance metric to investigate WGAN performance

The GAN performance was assessed qualitatively and quantitively. The manual inspection of generated images and the critic loss curve was performed to estimate the GAN performance qualitatively. A higher critic loss indicates higher quality images are generated. For the quantitative analysis of WGAN performance, the ratio of generated and real data was varied to create a concatenated dataset to train the classifier. The classifier's performance was observed on the test dataset. This strategy of quantitative measurement of the GAN performance was motivated by (Salehinejad, et al., 2018).

4 Chapter 4: Results (25%)

4.1 Images generated with traditional data augmentation

The images generated by applying traditional data augmentation techniques such as horizontal flipping, rotation, translation, brightness, and contrast are depicted in Figure 17.

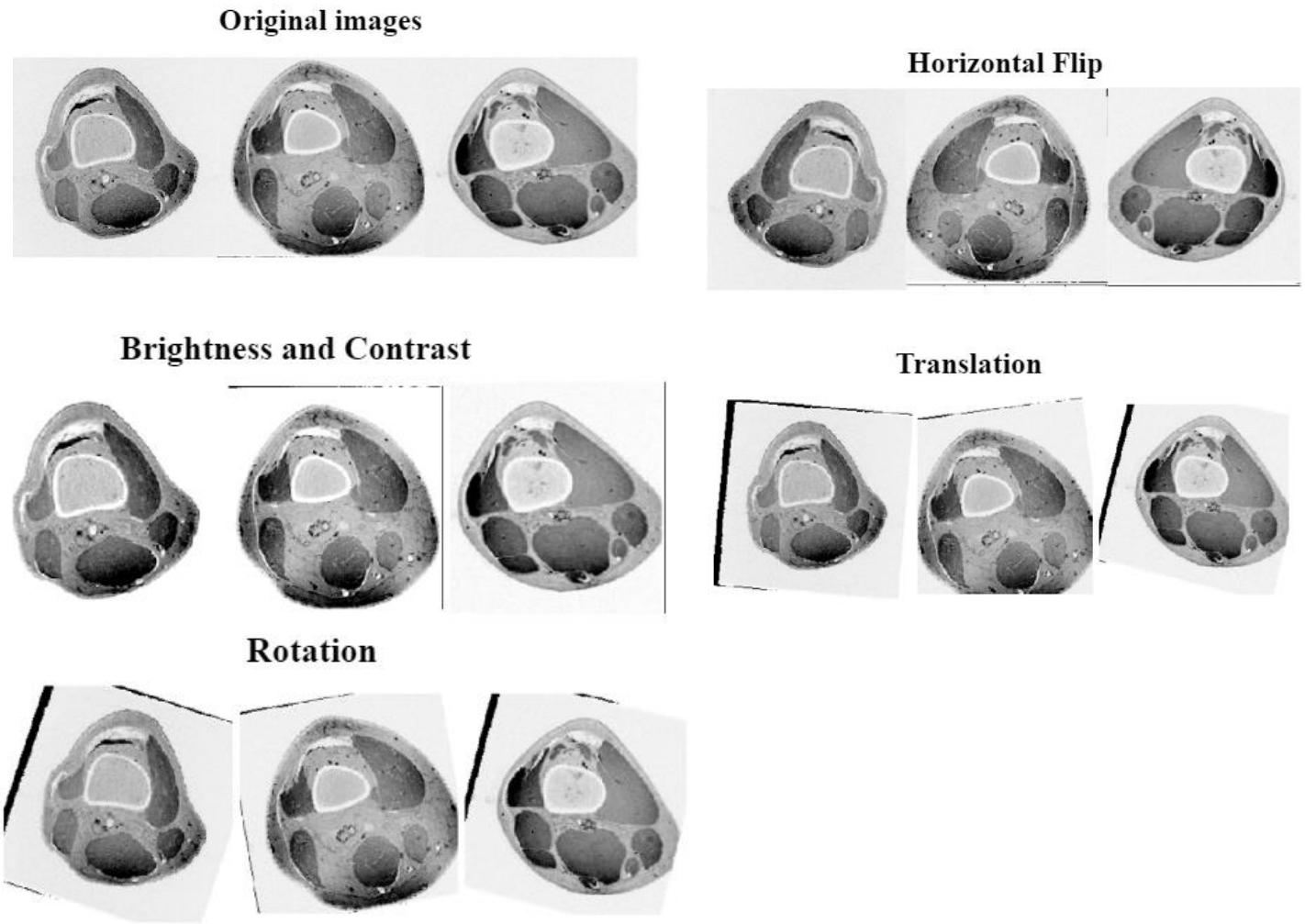


Figure 17: Geometrically augmented images

4.2 Data generated using WGAN based data augmentation

4.2.1 WGAN performance with different generator and critic architectures

First of all, the WGAN model was trained with CNN and ResNet based generator and critic networks to generate the images of size 64×64. The hyperparameter values used to train WGAN are mentioned in section 3. The main purpose of this experiment was to assess the quality of generated images with different generators and critic architectures. The experiment was performed on the images of size 64×64 to reduce training time and computation power.

4.2.1.1 Images generated using CNN based generator and critic networks

Figure 18 displays the real images of size 64×64. The generated images of dimension 64×64 after training WGAN for 100,000 iterations along with generator and critic loss curves are illustrated in Figure 19. The critic loss curve has more oscillations than generator loss. The reason behind high oscillations can be due to batch size that varies between 17 and 61. It seems the oscillations in the critic network becomes stable towards the last iterations. The generator loss is decreasing with the increasing number of iterations, while the critic loss is increasing with the increasing number of iterations. The increasing critic loss indicates that the critic becomes bad at discriminating between real and generated images, and the generator becomes good at generating more realist images.

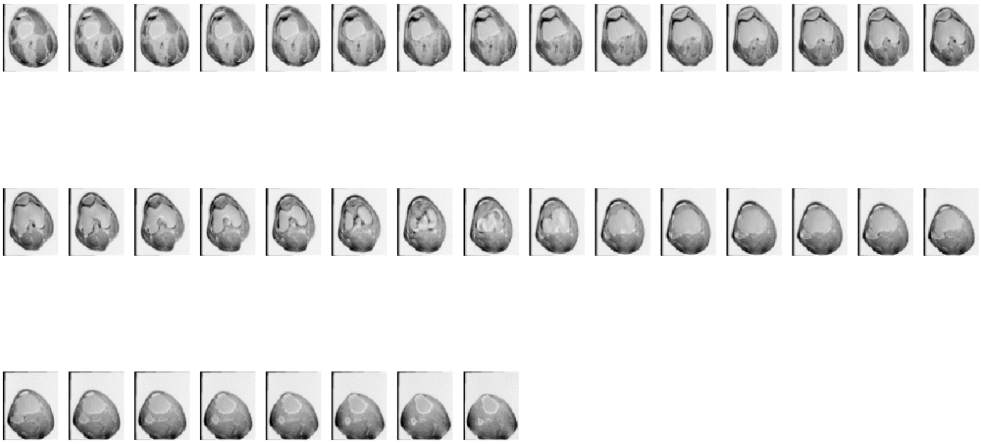
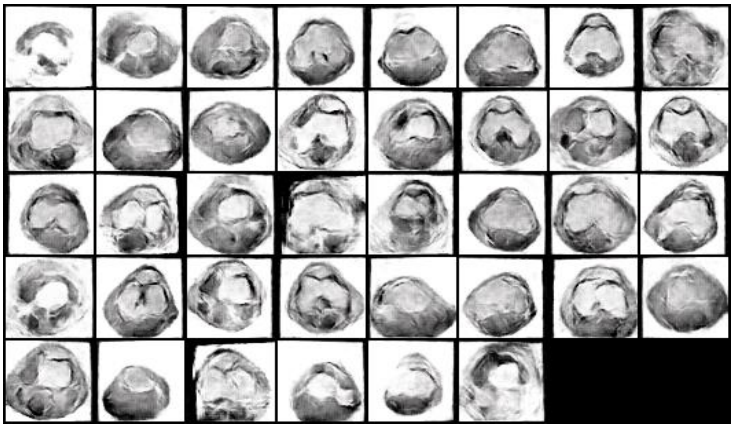


Figure 18: Real images of size 64×64



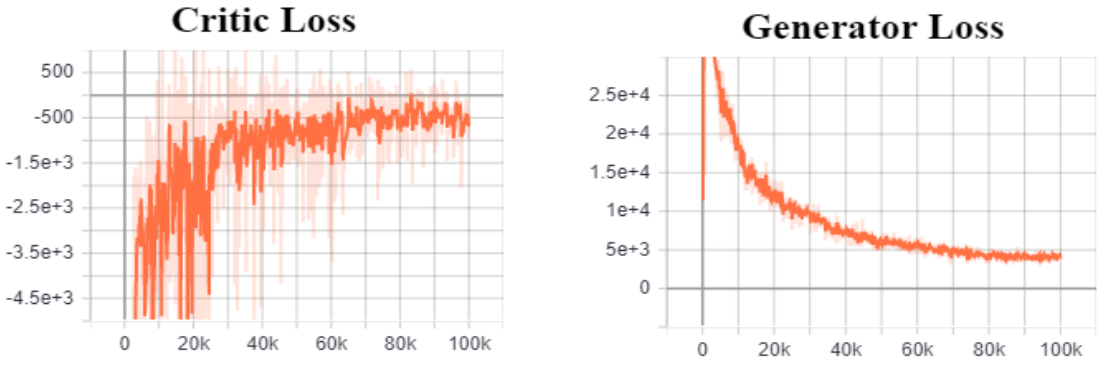


Figure 19: generated images along with generator and critic loss curves

4.2.1.2 Images generated using ResNet based generator and CNN based critic network

The image samples generated with the residual network-based generator and CNN based critic network with critic and generator loss curves are displayed in Figure 20. The critic and generator loss curves have high oscillations, but the quality of generated images is high, and generated images look more realistic. By observing loss curves, it can be analyzed that the WGAN model needs more training to stabilize the critic and generator loss curves. Further, the overall critic loss is increasing, and generator loss is fluctuating. Fluctuations in the generator loss curve indicate the learning of the generator network. Having different batch sizes at each iteration can be the reason for high oscillations in the loss curves.

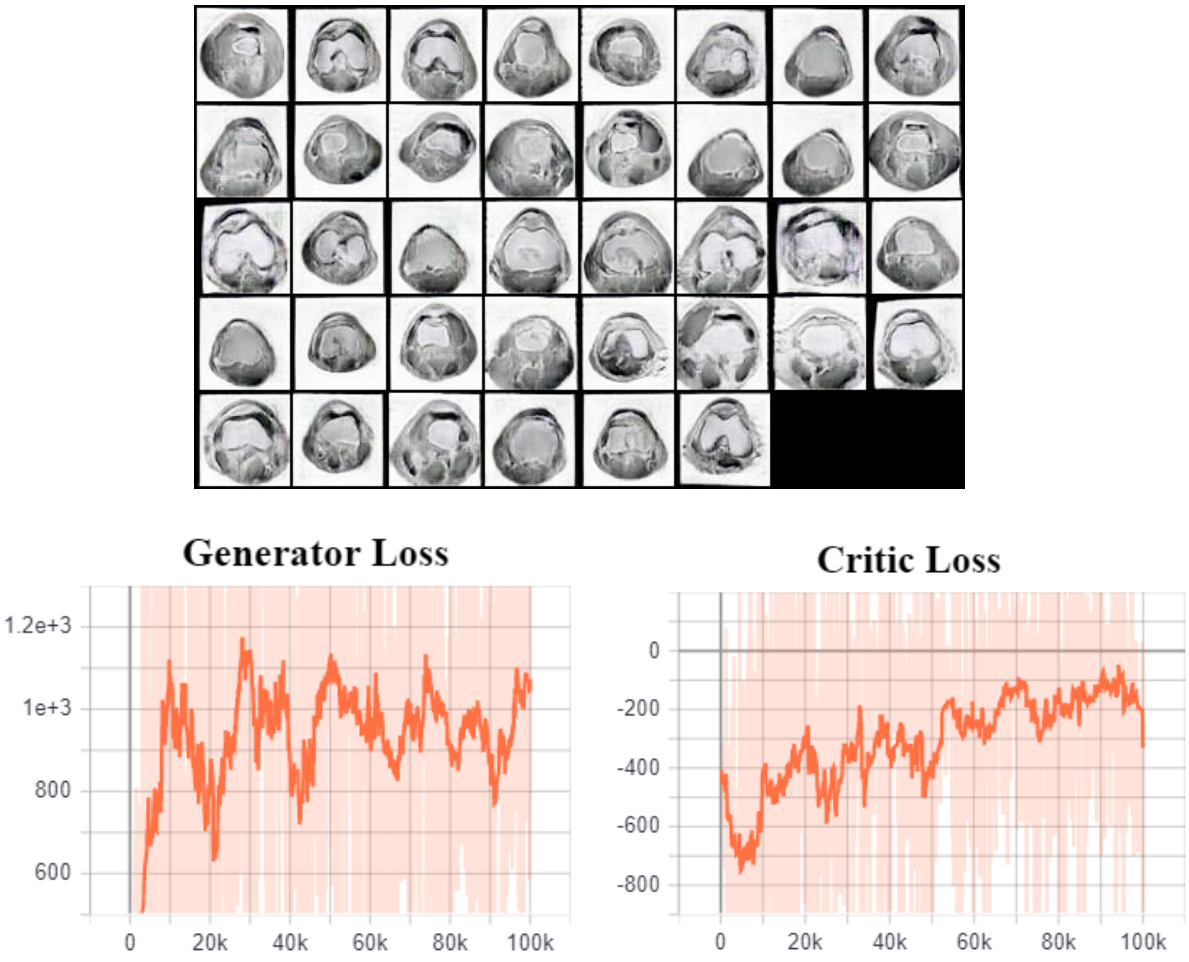


Figure 20: Images generated with CNN based C and Resnet based G along with loss curves

4.2.1.3 Comparing the WGAN performance with different generator and critic architecture

By comparing the performance of WGAN with the aforementioned generator and critic architectures, high-quality images were generated with ResNet based generator and CNN based critic networks. Although WGAN with CNN based generator and critic have more stable loss curves, the quality of generated images is not very high. On the other hand, the loss curves of WGAN using ResNet based generator and CNN based critic are not stable, which can be stabilized by training it for longer, but

generated images are more realistic. Based on the quality of generated images, the WGAN using ResNet based generator and CNN based critic was selected to generate 256×256 images. Moreover, the training time (1 day 2 hours 25 min and 8 sec) of the WGAN with ResNet-based generator and CNN-based critic was two times higher than the training time (14h 38 min 52 sec) WGAN with CNN based generator and discriminator. The longer training time of WGAN with ResNet based generator and CNN based critic is due to more convolutional layers in the generator and critic networks.

4.2.2 WGAN GP with different Hyperparameter values

The WGAN with ResNet based generator and CNN based critic was utilized to generate images of dimension 256×256 as this architecture succeeded in generating high-quality images of size 64×64. The values of learning rates and objective functions were altered while keeping the other hyperparameters the same, as mentioned in section 3. The results of a few experiments that were performed to find the optimal values of hyperparameters to generate realistic 256×256 images are illustrated in Figure 21.

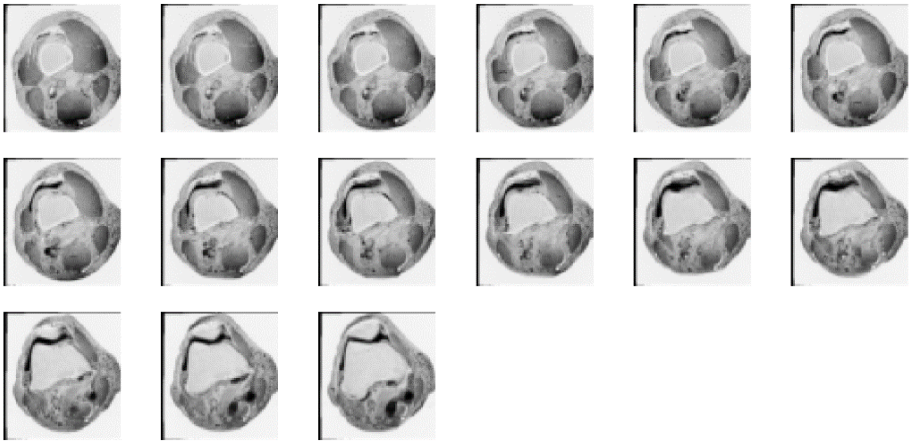


Figure 22: Real images of size 256×256

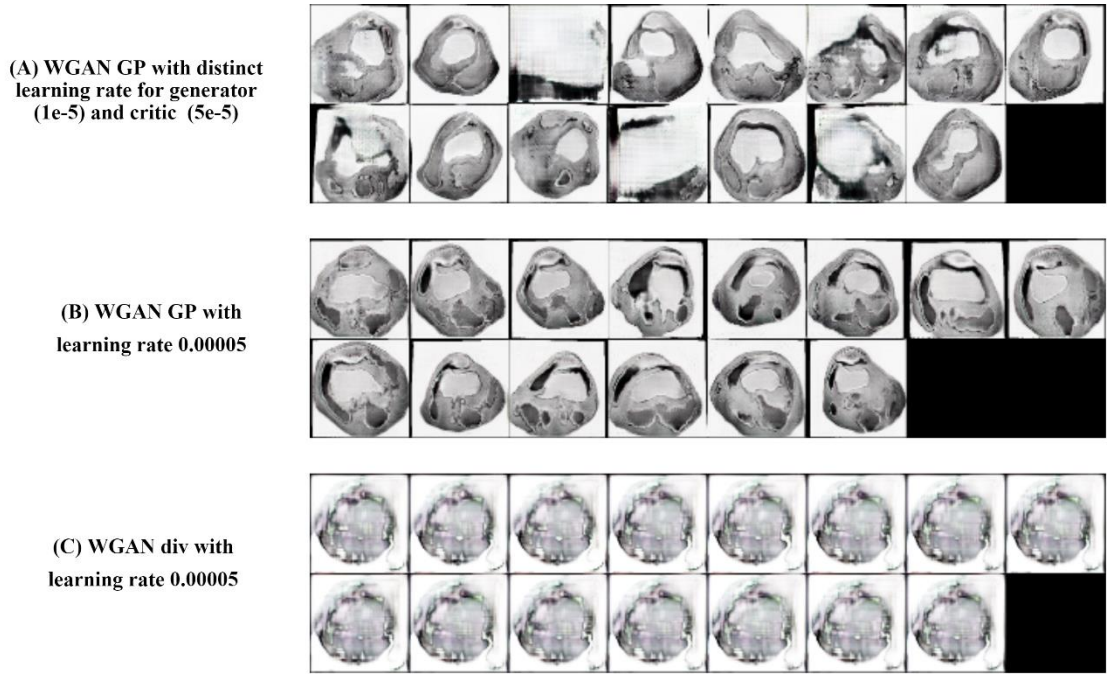


Figure 21: Images generated with different values of hyperparameters of WGAN

The WGAN model with different values of hyperparameters was trained for approximately 10,000 iterations. Firstly, the WGAN model was trained using different learning rates for generator (0.00001) and critic network (0.00005). The images generated after 10,000 iterations by WGAN with an individual learning rate for the generator and critic network are depicted in Figure 13. It can be observed that generated images are of low quality; even a few generated images have no pattern at all. The loss curves for WGAN training with different hyperparameters are illustrated in Figure 23. The critic and generator loss curves of WGAN training with distinct learning rate for critic and generator are fluctuating upward and downward, respectively. A more extended training and updating generator after each critic update rather than after 5 critic updates can potentially stabilize the loss curves.

Next, WGAN was trained with WGAN divergence objective function. The images generated after 12,000 iterations are shown in Figure 24. The generated images do not look near to realistic images. Moreover, the generator and critic loss curves have high oscillations. Training WGAN divergence for longer can potentially contribute to more realistic images with stable training curves. Similarly, the WGAN was trained with learning rates of 0.0001 for 10,000 iterations but resulted in poor quality generated images. The reason for oscillations in each experiment's loss curves can be due to a smaller batch size of 15 that was selected to resolve memory errors.

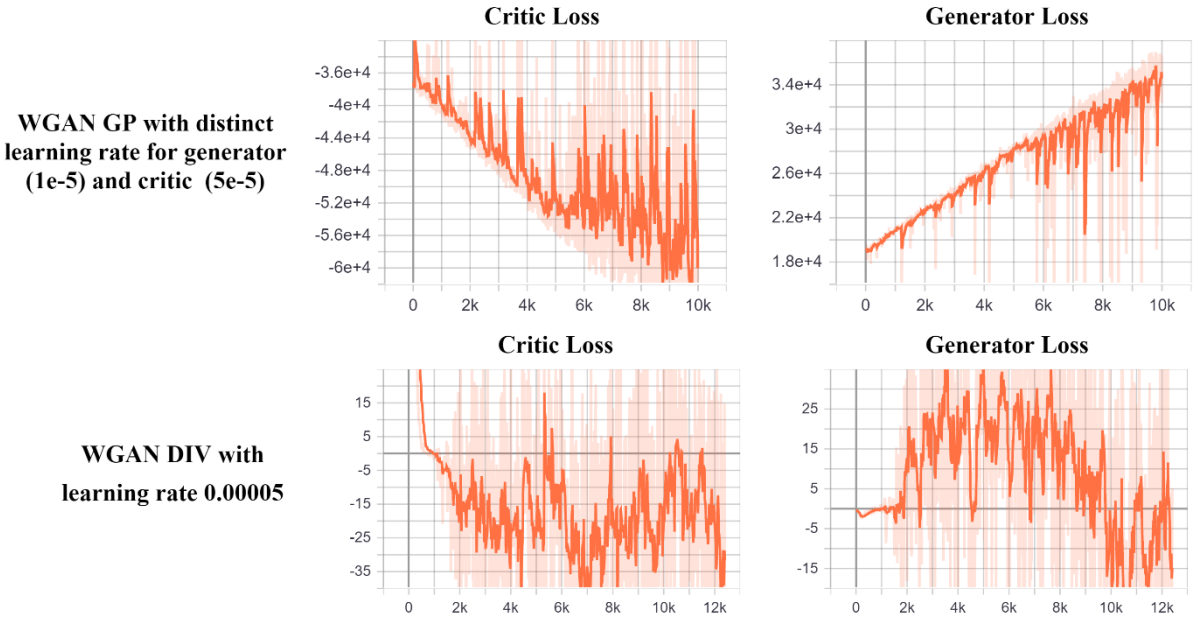


Figure 23: Loss Curves of WGAN with different hyperparameter setting

After that, WGAN with Gradient penalty was trained with a 0.00005 learning rate for 10,000 iterations. The images generated by WGAN GP with a learning rate of 0.00005 look more plausible than images generated by executing all the above experiments. The images generated after 10,000 iterations are displayed in Figure 23. All the experiments performed above can lead to high performance by fine-tuning the parameters and training them for longer. Moreover, the time taken to

train each experiment for 10,000 iterations is approximately 9 hours. Due to the time limit, we did not opt to train these experiments for longer or try to fine-tune hyperparameters values further.

Since the better quality images after 10,000 iterations were generated with WGAN GP and a learning rate of 0.00005; therefore, we selected this set-up to generate images for three categories(abnormal with an ACL tear, abnormal with a meniscus tear, no knee disorder). Table 2 illustrates the time taken by WGAN with different settings.

Table 2: Time taken to train WGAN with different Settings.

WGAN with different settings	Number of iterations	Training time
64×64 with ResNet based G and C	100,000	1 day 2 hours 25 min and 8 sec
WGAN GP with distinct learning rate for generator (1e-5) and critic (5e-5)	10,000	9 h 29 min 6s
WGAN div	12,000	14 hours 16 min 5s
WGAN GP with learning rate 0.00005	10,000	9 hour 5 min 26 s
64×64 with CNN based G and C	100,000	14h 38 min 52 sec

Table 3 depicts the WGAN model setting used to generate images of size 256×256 for three categories. Three WGAN models were trained separately for approximately 80,000 iterations to synthesize images of three categories.

Table 3: WGAN optimal hyperparameter setting

Parameters	Values
Generator (G)	ResNet based architecture
Image size	256×256
Noise vector	128
Critic(C)	CNN based architecture
Critic Loss	Wasserstein distance + Gradient penalty
Generator Loss	-[critic value on generated images]
Optimizer	Adam
Batch size	15
Critic iterations	5
Learning rates	0.00005
(b1,b2)	(0.5,0.999)

4.2.3 Data generated by WGAN for three categories

4.2.3.1 Abnormal with a meniscus tear class

The WGAN model was trained for 80,000 iterations to generate images for abnormal with a meniscus category. Figure 24 illustrates the generated images at different iterations with generator and loss curves. The images generated at the 1st iteration are of poor quality. With the increasing number of

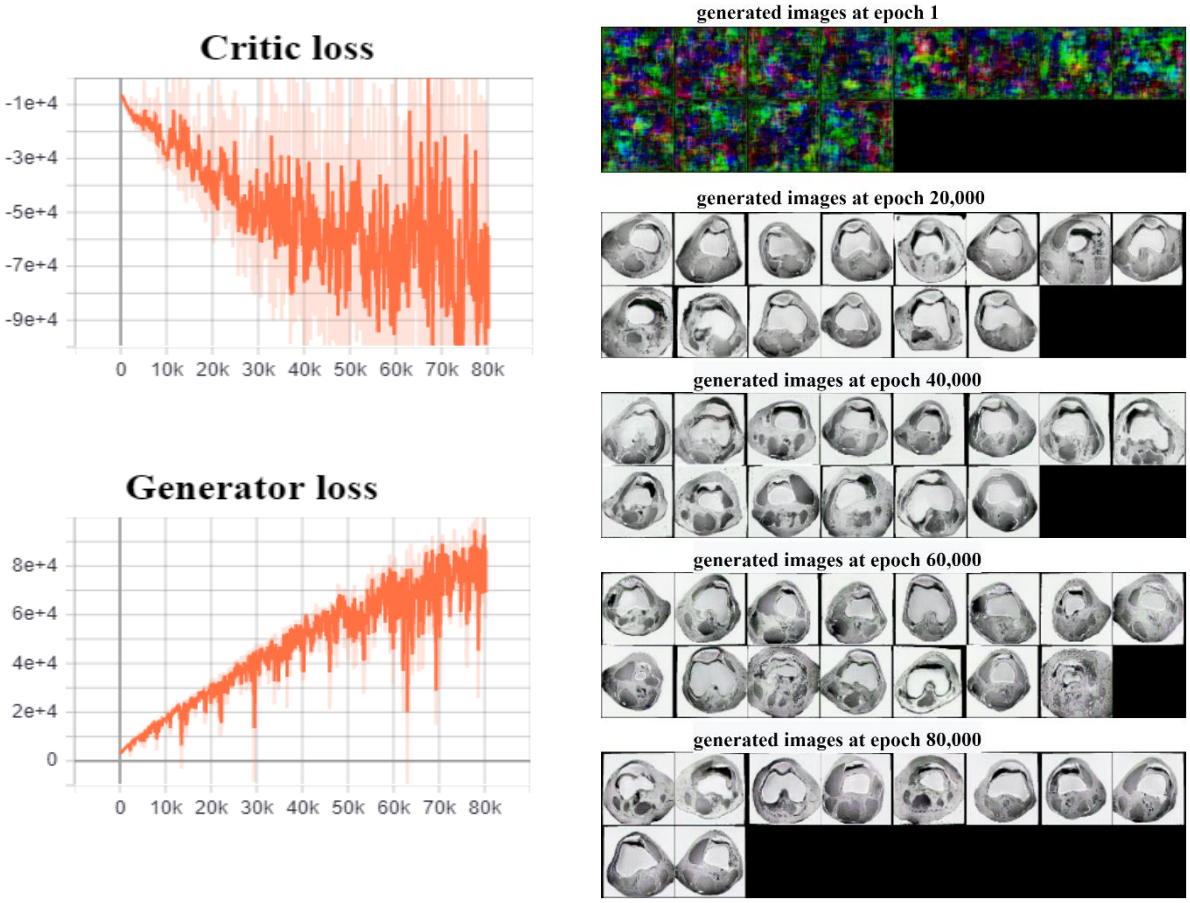


Figure 24: Images generated at different iterations along with generator and critic loss curves

iterations, the generator loss is fluctuating downward while critic loss fluctuates upward. With the increasing number of iterations, WGAN begins to generate more plausible images. The critic loss is notably high, while generator loss is considerably low between 60,000 and 80,000 iterations, which suggests images generated between 60,000 and 80,000 are high quality and realistic. The loss curves indicate that the model didn't reach convergence during 80,000 iterations, which implies training the WGAN model for longer iterations (possibly 200,000) is required. Further, the time taken to train WGAN for 80,000 iterations was approximately 3 days and 8 hours. The WGAN training for 200,000 iterations requires significant time and computational resources. Due to limited time and computation resources, further training of WGAN was not performed. The WGAN model trained for 80,000 was evaluated to generate around 300 MRI scans consisting of around 10,000 images in total for abnormal

with a meniscus class, which were labelled before adding to the MRNet dataset. The generated and real MRI scans are displayed in Figure 25, and 26.

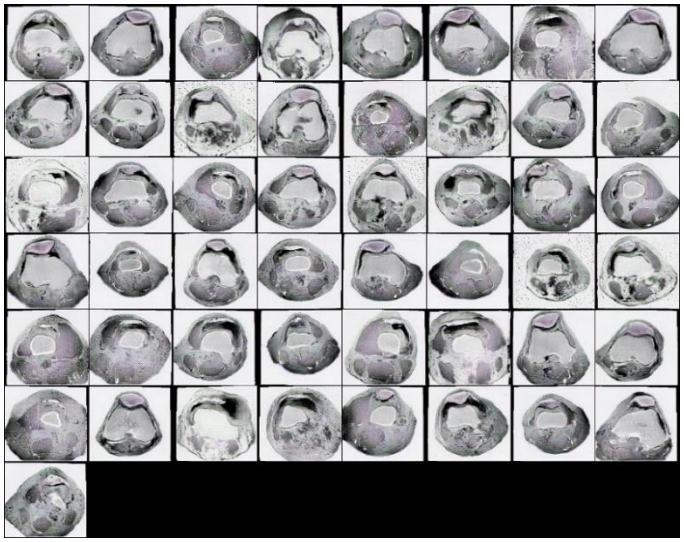


Figure 25: Images generated by WGAN for abnormal with a meniscus tear cases

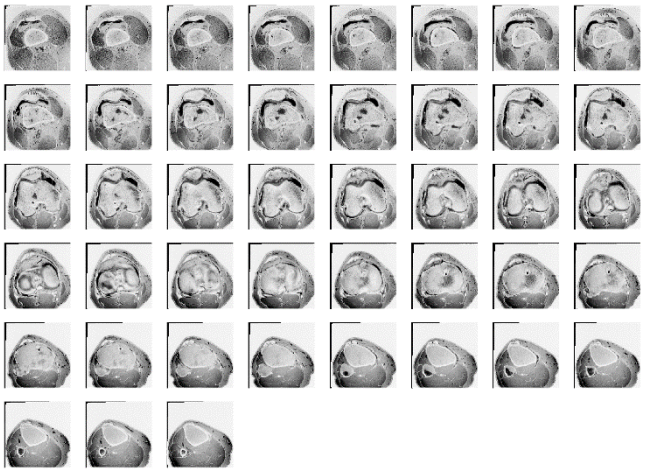


Figure 26: Real images of abnormal with meniscus tear category

4.2.3.2 Abnormal with ACL tears

Figure 27 illustrates the generated images at different iterations during the training of the WGAN model. The generated images are more plausible with the increasing number of iterations. The loss curves of critic and generator are shaky, as indicated in Figure 27. There has not been a significant drop in generator loss during the training. On the other hand, the critic loss is very high most of the time during training, which implies that critic could not discriminate between generated and real images. Also, the critic loss is more important than generator loss when assessing WGAN performance. The higher critic loss is the indication of the synthesis of more realistic images, as mentioned in (Arjovsky et al. 2017). However, the generated images look realistic but can be enhanced by training the WGAN model for longer iterations. The WGAN model was trained for about 90,000 iterations to generate abnormal with ACL tear images. The WGAN training took approximately 3 days and 5 hours. Approximately 300 MRI scans comprising about 10,000 were generated and labeled. After labelling, the generated MRI scans were added to the MRNet dataset.

The generated and real images of abnormal with ACL tear class are displayed in Figures 28 and 29, respectively. However, the loss curves are zig-zag curves as the loss is reported after every ten iterations. The smooth loss curves can be achieved by displaying the average loss over every 1000 iterations. Moreover, the number of MRI scans for abnormal with ACL tear category are 83, comprising 2, 833 2D image slices. Increasing the number of abnormal with ACL tear cases by applying geometric transformations and then using the enlarged dataset to train WGAN can help to achieve more stable training loss curves and highly realistic generated images.

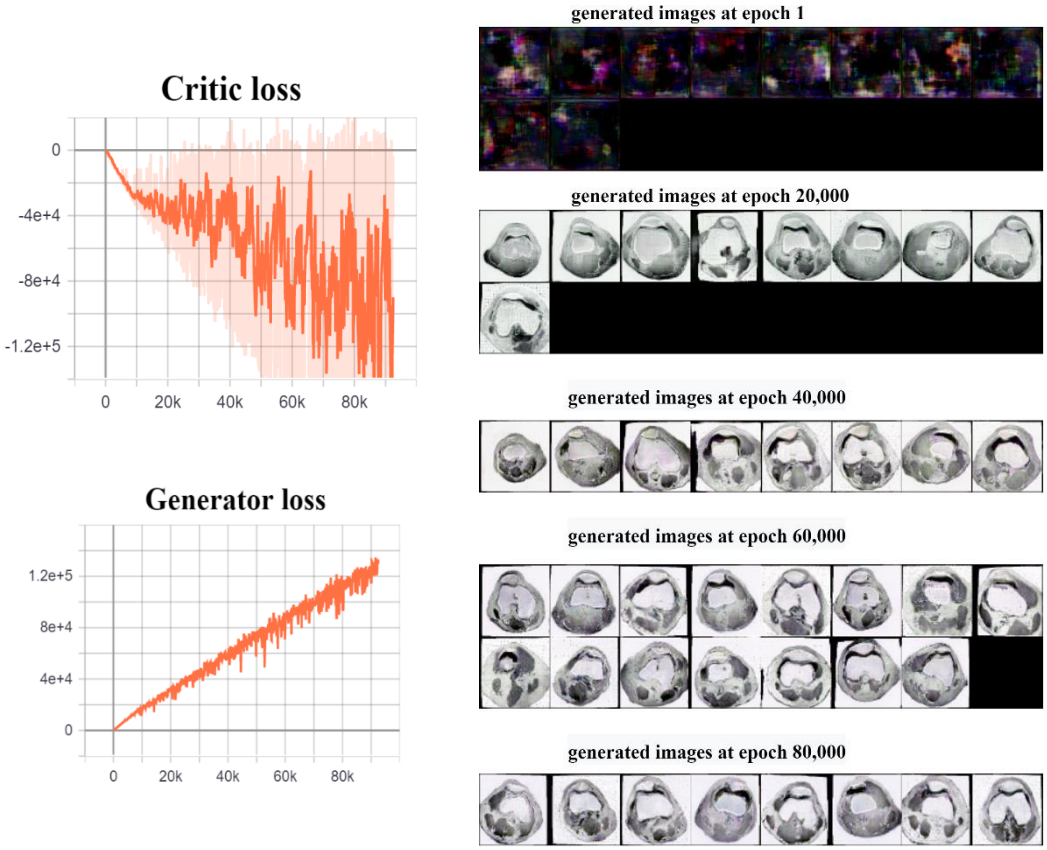


Figure 27: Image generated at different epochs plus critic and generator los curves

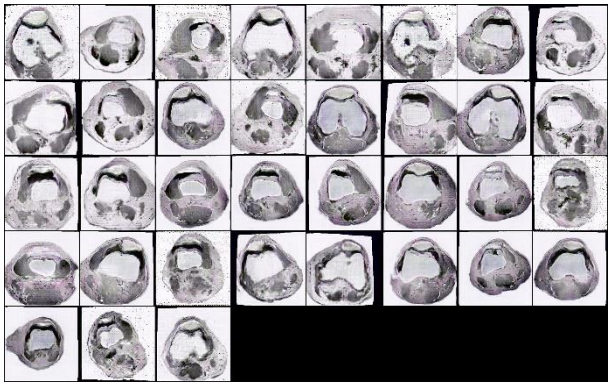


Figure 29: generated MRI series of abnormal with an ACL category on evalution of WGAN model

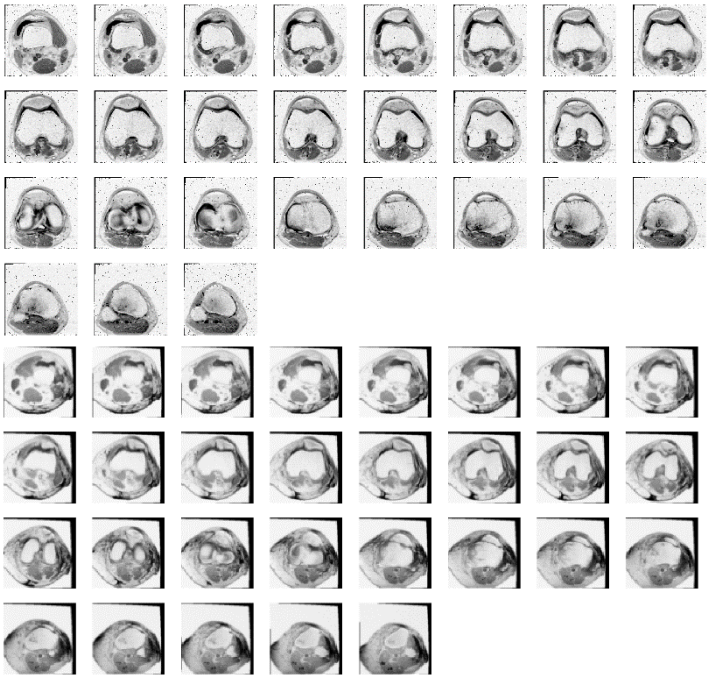


Figure 29 : Real Images of abnormal with ACL tear class

4.2.3.3 No abnormalities

The images generated at different steps during training can be examined in Figure 30. After training for 20,000 iterations, the WGAN is generating more credible images, and the quality of generated images increasing with the number of iterations. The loss curves of the generator and critic network are varying during the training, which suggests the WGAN model is learning. The critic loss is fluctuating upward during training, particularly between 70,000 and 80,000 iterations. Besides, the generator loss has seen significant drops between 60,000 and 80,000 iterations. From Figure 30, it can be analyzed that both generator and critic losses did not converge to one optimum point during training, but it does not denote that the model did not learn anything. The images generated after training for 80,000 iterations are of high quality and look realistic. Training the WGAN for longer would potentially assist the generator and critic losses to converge to the optimum point. After training for 80,000 iterations, the model is able to generate realistic images. Further training was not executed considering the time and computational resources. It took approximately 2 days and 14 hours to train WGAN for 80,000 iterations. After training, approximately 300 MRI scans comprising around 8,000 images were generated. The generated MRI scans were labeled and included in the

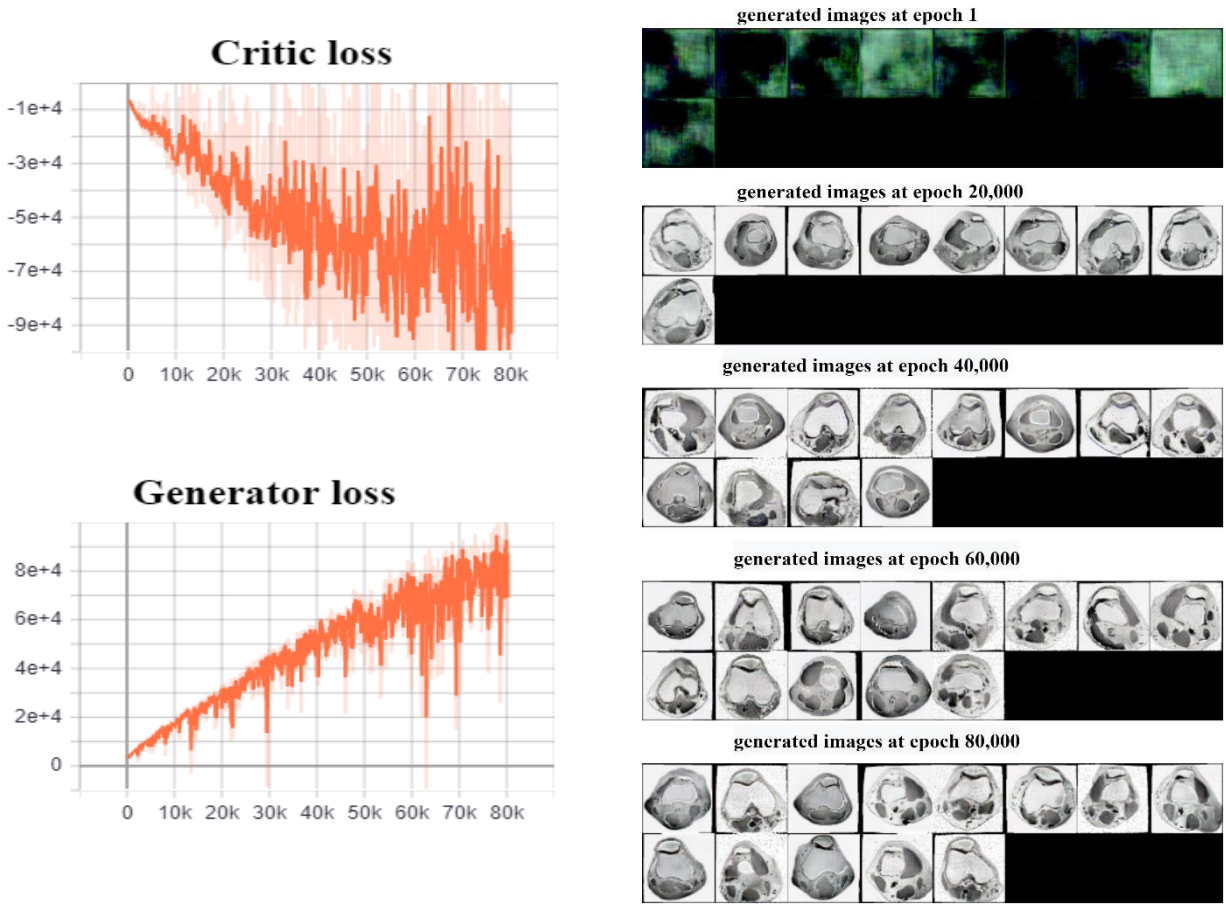


Figure 30: Images generated at different iterations along with generator and critic loss curves

MRNet dataset. The generated and real images of the category with no knee disorders are displayed in Figures 32 and 33, respectively.

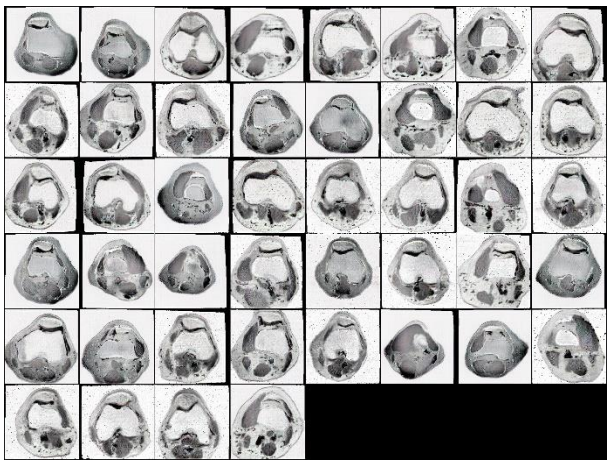


Figure 32: images generated by WGAN for cases with no abnormalities

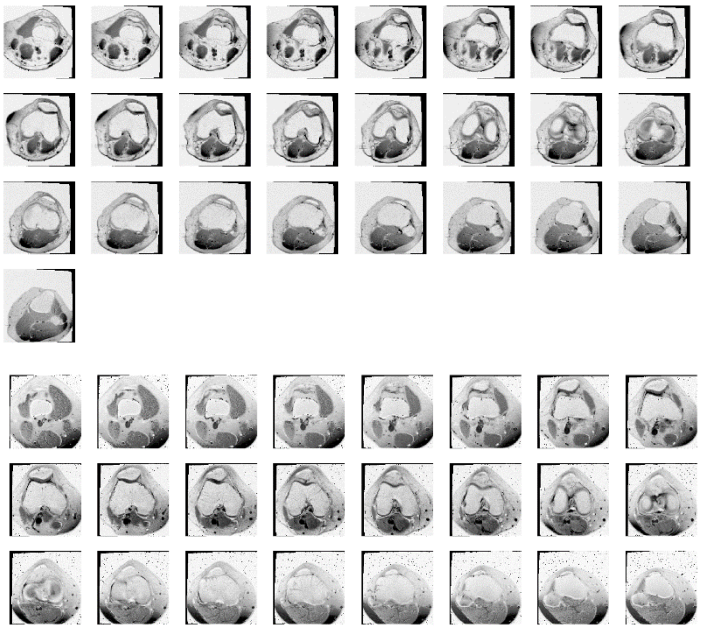


Figure 32: Real image of cases with no abnormalities

4.2.4 WGAN performance Analysis quantitatively

For the quantitative analysis, two classification models were trained with 300 and 100 original data samples from the training data along with 200 and 175 WGAN based augmented samples, respectively. The performance of the trained classifiers was tested on the test dataset (120). The values of AUC and accuracy were calculated to assess the performance of the classifier. Both classifiers have higher AUC and accuracy in abnormal detection. The classifier trained with the majority of WGAN augmented samples achieved poor performance in ACL detection. On the other hand, the classifier trained with the majority of original data samples has 77% capability to differentiate ACL and non ACL tear cases. It indicates that the GAN generated images do not have similar characteristics to the original data samples in ACL detection, but generated images have the same characteristics to training data in abnormality and meniscus detection. Overall it can be summarised that WGAN model did learn the characteristics of the original data, but there is space for improvement. The quality of WGAN generated images can potentially improve by experimenting with hyperparameter settings, and training for longer until the training is stabilized. Moreover, It is

specified in (Arjovsky et al. 2017) that training the critic for longer gives reliable gradients. Table 4 illustrates the classifier performance with different ratios of real and WGAN generated images.

Table 4:Classifier performance with different ratios of real and WGAN based augmented

Original Data Samples	WGAN based Augmented Data Samples	AUC			Accuracy		
		abnormal	ACL	Meniscus	abnormal	ACL	Meniscus
300	200	89%	77%	75%	85%	67%	65%
100	175	78%	63%	69%	81%	55%	64%

4.3 Classification model performance Analysis

After data augmentation, developing a classification model is the second stage of this project. First, the classification model was trained with original data samples. After that, the classifier was trained with geometrically augmented data, WGAN based augmented data, and both geometric and WGAN augmented data to examine its performance with different types of data augmentation techniques. Moreover, the classification model performance was assessed with WGAN generated images for each category individually to examine a comprehensive impact of WGAN data augmentation on the classifier's performance.

4.3.1 Classification performance without any data augmentation

First of all, the MRNet model was trained using original samples in the training dataset. Three MRNet models were trained for abnormal, ACL, and meniscus detection individually. The MRNet model with hyperparameters setting stated in section 3 of the report was trained for 50 epochs using training data (1130 cases) as there was no improvement in the validation performance after 50 epochs.

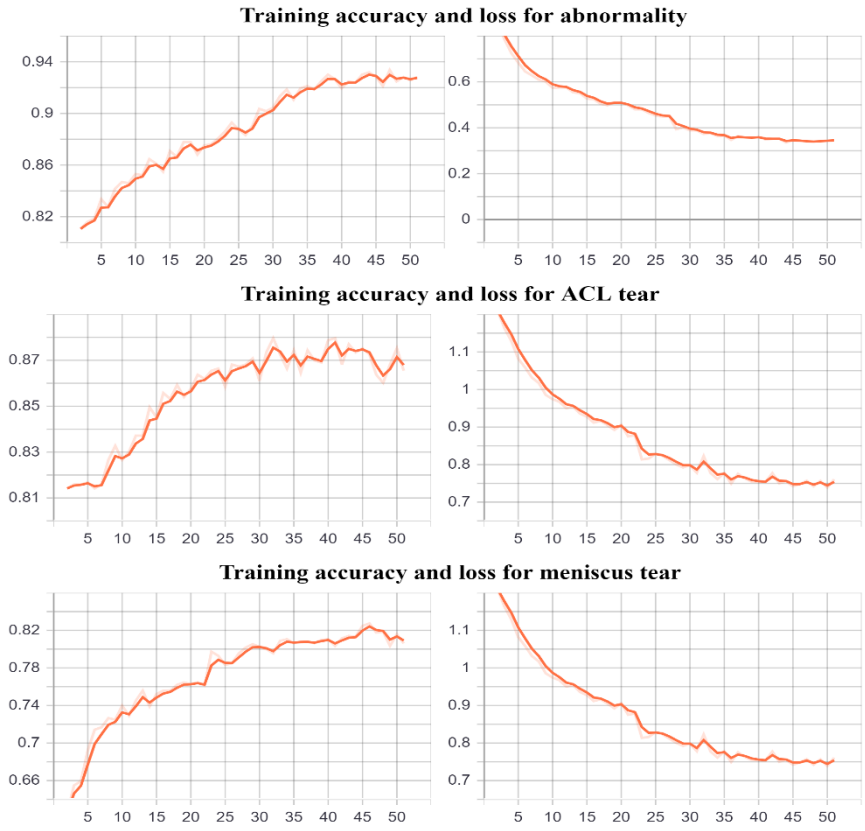


Figure 33: Training accuracy and loss curves for each class

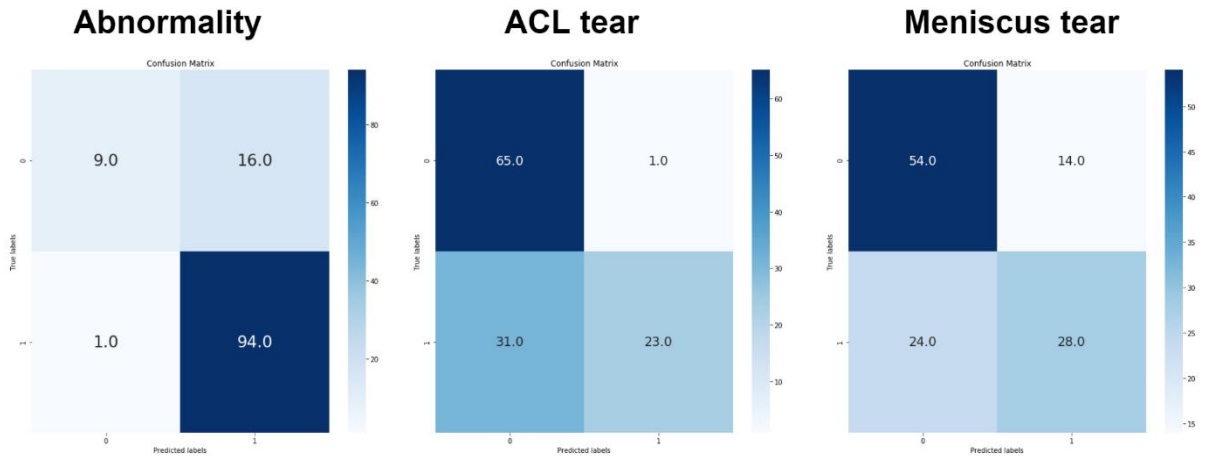


Figure 34: Confusion matrices on test dataset

Figure 33 illustrates the training accuracy and loss curves for each class. By looking at training and loss curves, it can be observed that the accuracy and loss curves are stabilized and have converged to optimum accuracy and loss value. The model's training accuracy in abnormal detection (93%) is the highest compared to ACL and meniscus tears detection. After training, the performance of the MRNet model was tested on the test data (120 cases). The performance metrics to be specific confusion matrix, ROC curve, area under the receiver operating curve (AUC), accuracy, precision, recall, and sensitivity were utilized to investigate the performance of the model.

The confusion matrices were plotted for abnormal, ACL, and meniscus tears detection on the test dataset, as shown in Figure 34. For abnormal class, the model made the maximum number of correct predictions for cases with the abnormality. For ACL and meniscus tears detection, the maximum number of predictions were made for cases with the absence of ACL and meniscus tears.

The MRNet models trained for abnormal, ACL, and meniscus classes have 93%, 87%, and 75% capability to discriminate abnormal from non-abnormal cases, ACL tear from non ACL tear cases, and meniscus tear from non-meniscus tear cases, respectively. Moreover, the MRNet model trained for abnormal class made 90% correct predictions for abnormal cases and 85 % correct predictions for non-abnormal cases. Following, the MRNet model for ACL tear made 67% correct predictions for ACL tear cases and 95% for non ACL tear cases. Next, the MRNet model trained for meniscus

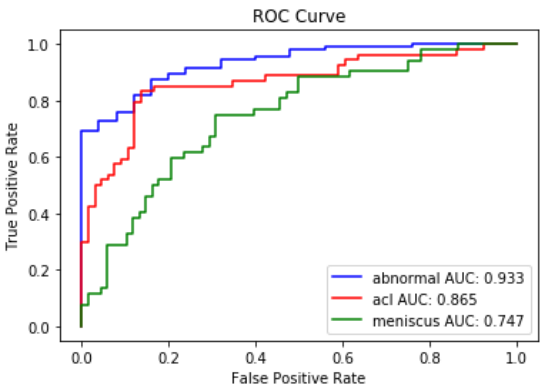


Figure 35: ROC curve on test dataset for abnormal, ACL tear and meniscus tear

class made 69% correct predictions for meniscus tear cases and 66% for non -meniscus tear cases. In terms of generalization performance, training and test accuracy are observed. Abnormal training and test accuracies are 93% and 86%, respectively. The training and test accuracies for ACL and meniscus tear detection are 87%and 73%, and 80 and 68%, respectively. There is no significant difference in the three models' training and test accuracy, indicating good generalization performance for each task. The ROC curve is plotted at a 0.5 output threshold and can be visualized in Figure 35.

4.3.2 Classification performance with geometric based data augmentation

In order to increase diversity in the training dataset, the affine and pixel-level transformations mentioned in section 3 were applied randomly on 2D image slices in the MRI series. After that, randomly transformed knee MRI data was used to train the MRNet model. The training was performed for 100 iterations that took approximately 19 hours. Figure 36 illustrates the training and loss curves for abnormality, ACL tear, and meniscus tear. The training and loss curves for each class are stable and converged to the optimum value of loss and accuracy. The highest training accuracy for abnormality, ACL tear, and meniscus tear is 87%, 86%, and 76%, respectively. With the increasing number of epochs, the accuracy is increasing while the loss value is decreasing. After training, the performance of the classification model was evaluated on the test dataset (120 MRI scans). Confusion matrices and ROC curves (see Figure 37 and Figure 38) were plotted to visualize the performance of three MRNet models trained for abnormality, ACL, and meniscus tear detection. In abnormality detection, the highest number of accurate predictions are made for cases with abnormality by the MRNet model. On the other hand, the MRNet model correctly classified patients with the absence of ACL and meniscus tear than patients with the presence of ACL and meniscus tear. Further, the AUC value in the ROC curve indicates that the trained models have 92%, 89%, and

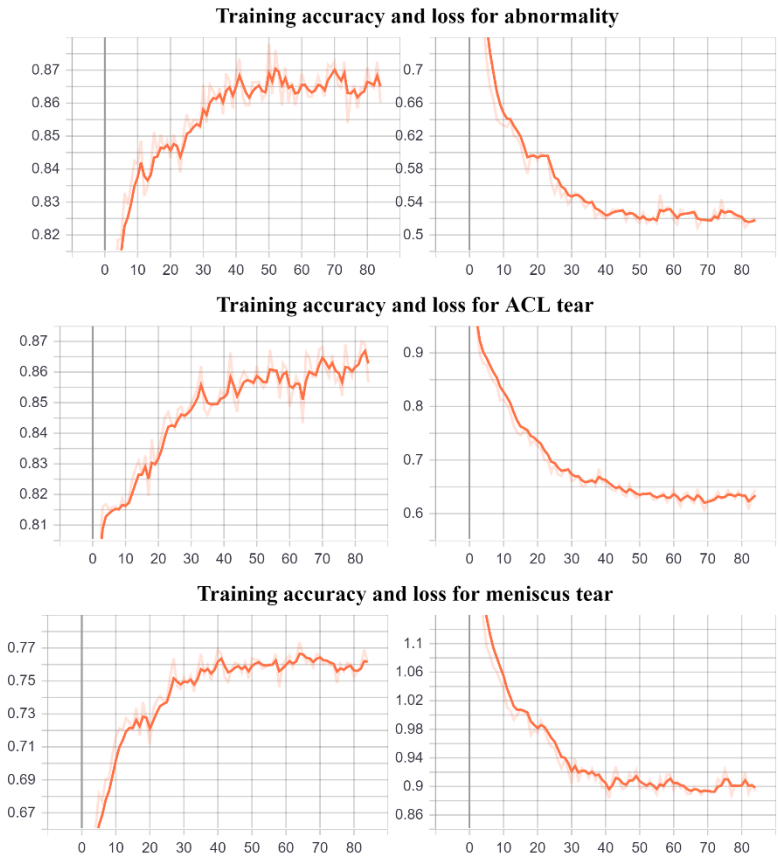


Figure 36: Training and loss curves for abnormality, ACL tear and meniscus

77% competency to discriminate abnormal from non-abnormal cases, ACL tear from non ACL tear cases, and meniscus tear from non-meniscus tear cases, respectively. Moreover, the MRNet model for

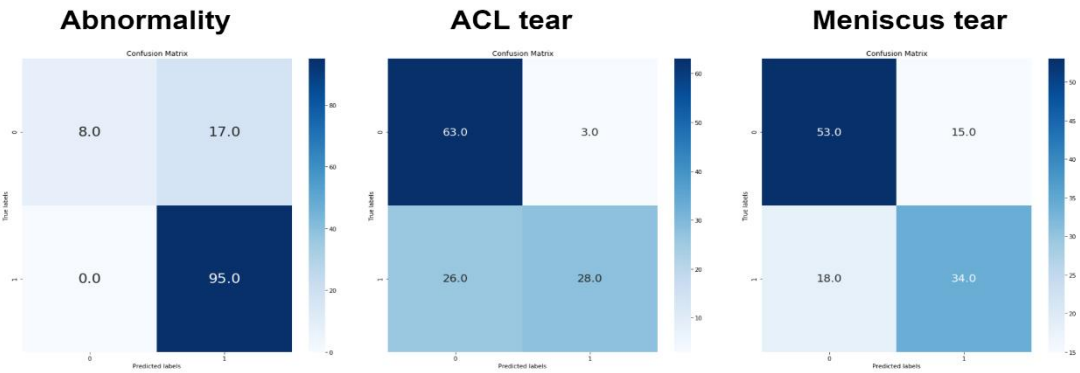


Figure 37: Confusion matrix on test set for abnormality, ACL tear and meniscus tear

abnormality, ACL tear, and meniscus tear made 100% correct predictions for abnormal cases and 85% for non-abnormal cases, 95.8% for ACL tear cases, and 67% for non-ACL tear cases, and 66% for meniscus tear cases and 69% for non-meniscus tear cases, respectively. The highest weighted F score is 83% for abnormality.

The training and test accuracy for abnormality, ACL tear, and meniscus tear were compared to investigate the model's generalization performance. The test accuracies for abnormality, ACL tear, and meniscus tear are 86%, 76%, and 73%, respectively. The difference between training and test accuracy is 1%, 10%, and 3%, which is very low. Overall, applying geometric transformations helped to improve the model's generalization performance.

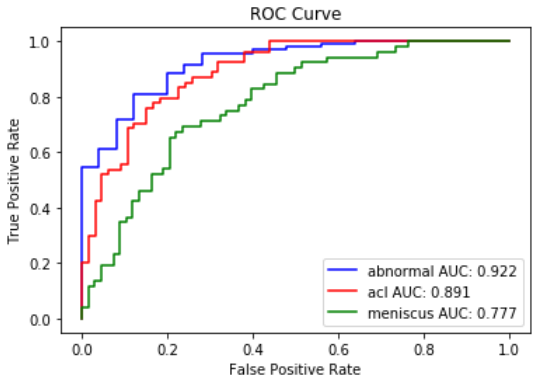


Figure 38: ROC curve on test dataset

4.3.3 Classification performance with WGAN based augmented data

The data generated using WGAN was added to the MRNet dataset. Approximately 300 MRI scans for abnormality, ACL tear, and meniscus tear were included, which increased the number of cases to 2031 from 1130 in the dataset. After that, the MRNet model was trained with WGAN based augmented dataset. The model training was performed for 90 epochs, which took one day and 12 hours to complete. Figure 39 illustrates the training accuracy and loss curves for abnormality, ACL tear, and meniscus tear detection. The loss and accuracy curves are steady and converged to optimal values. The loss is decreasing while the accuracy is increasing with the increasing number of epochs. The highest training accuracy of 96% is achieved for ACL tear detection.

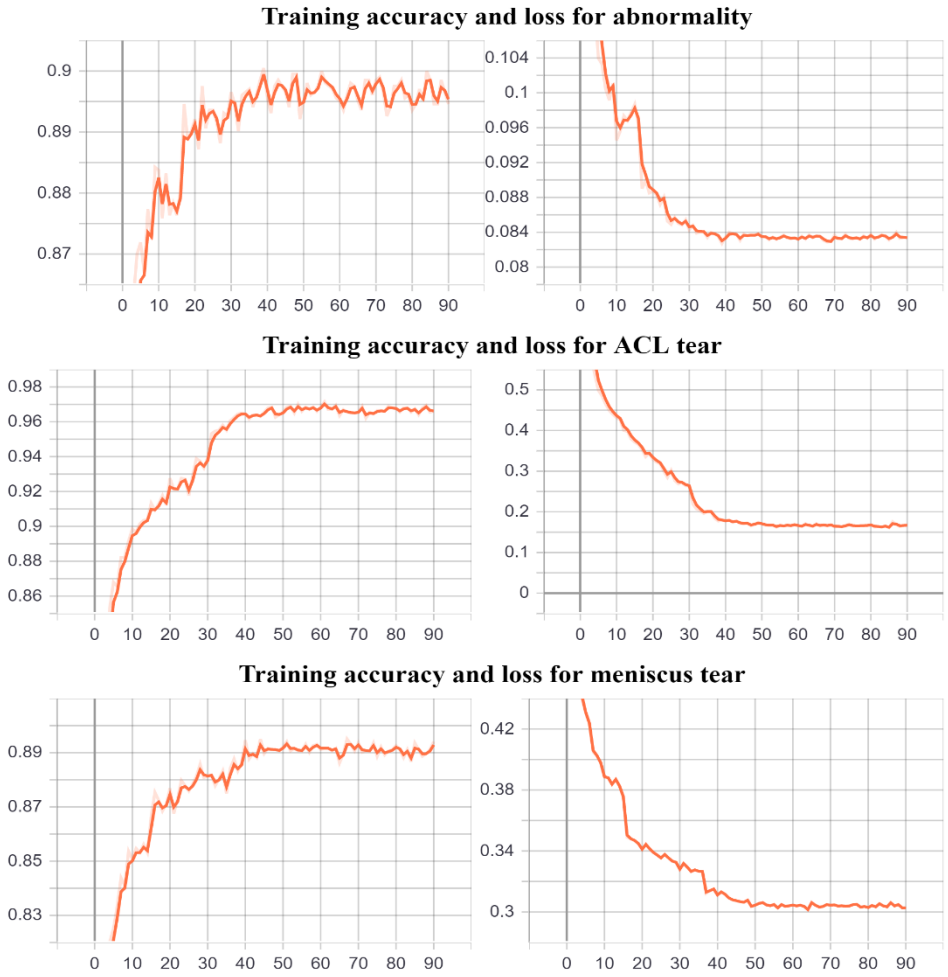


Figure 39: Training and loss curves for abnormality, ACL tear and meniscus tear

After training, the model's performance was assessed on test data. Confusion Matrix and ROC curve (see Figure 40, Figure 41) on test data were plotted for abnormality, ACL tear, and meniscus tear. The ROC curve was plotted with the output threshold at 0.5. The values of AUC in the ROC curve indicate the model's ability to discriminate abnormal from non-abnormal cases, ACL tear from non ACL tear cases, and meniscus tear from non-meniscus tear cases, which is 88%, 88%, 78%, respectively. Moreover, the specificity, sensitivity, accuracy weighted precision, and F score were computed on the test dataset. The model achieved the highest specificity of 91% for abnormality detection and sensitivity of 81% for ACL tear detection. The lowest specificity of 58% was observed for meniscus tear detection. The confusion matrix gives the number of true negative, false-negative, false-positive, and true positive predictions, respectively. The maximum number of true

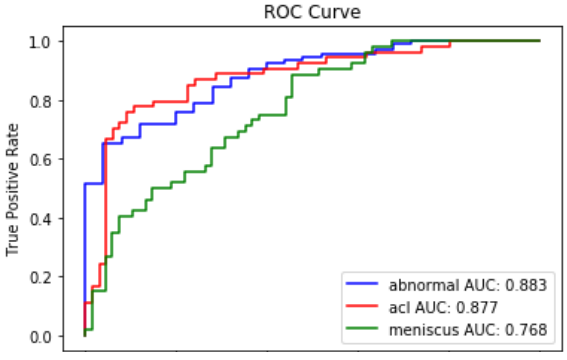


Figure 40: ROC curve

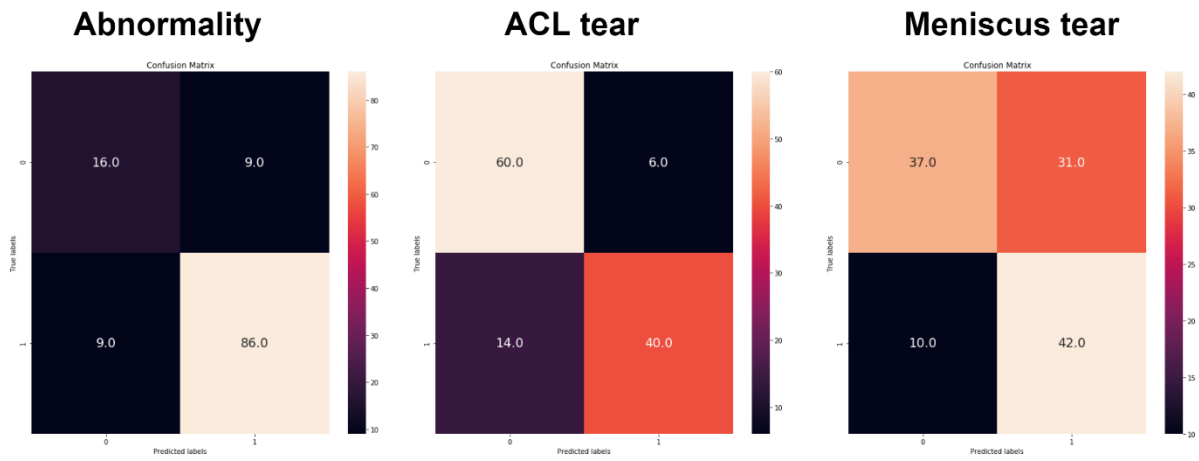


Figure 41: Confusion Matrices

positive predictions were made for abnormality detection, and the true negative predictions were top in ACL tear detection. The weighted precision and f score have the highest ratio of 85% for abnormality detection.

Regarding generalizability, the model's accuracy on the training and the test data was observed. The test accuracy is stated in table. There is no considerable difference in the training and test accuracies of abnormality, ACL, and meniscus tear detection, indicating good generalization performance. Overall, the model trained with WGAN based augmented dataset has the highest classification performance for abnormality detection as compared to ACL and meniscus tear detection based on the classification metrics.

4.3.4 classification performance with both geometric and WGAN based data augmentation

Subsequently, the classification performance was investigated with data augmented using both geometric and WGAN data augmentation techniques. The training was performed for 100 epochs that accomplished in 1 day and 19 hours. Figure 42 illustrates the training and loss curves for abnormality, ACL, and meniscus tears. The loss and training curves indicate the steady learning of three classification models.

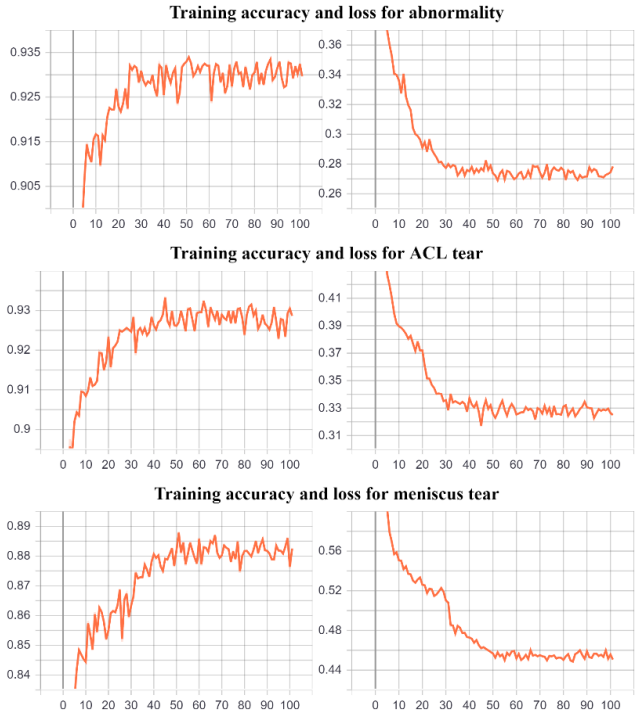


Figure 42: Training and loss curves for abnormality, ACL and meniscus tears detection

The classification model's performance on the test dataset was assessed using the classification evaluation metrics mentioned in section 3. The classification model trained for abnormality, ACL, and meniscus tear detection has a specificity of 90%, 73%, and 74%, respectively. Further, the sensitivity for abnormality, ACL tear, and meniscus tear is 85%, 84%, and 70%. In order to investigate the relationship between specificity and sensitivity, the ROC curve is plotted, which is illustrated in Figure 43. The AUC values in the ROC curve indicates that the abnormality detection model has the highest discriminative power. In order to get an overview of the number of true positive, false positive, true negative, and false negative predictions for each category, the confusion matrix was analyzed. Figure 44 illustrates the confusion matrix for each binary task. For abnormality detection, the model made correctly predicted 94 cases out of 95. For ACL and meniscus tear detection, the top predictions made were true negative. Next, training and test accuracies were used to investigate the generalization performance of the models. By

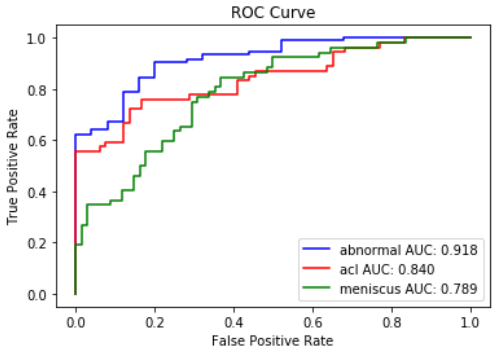


Figure 43: ROC curve on the test dataset

observing the training accuracy plots and test accuracies, which are stated in Table 5, no considerable variations was examined.

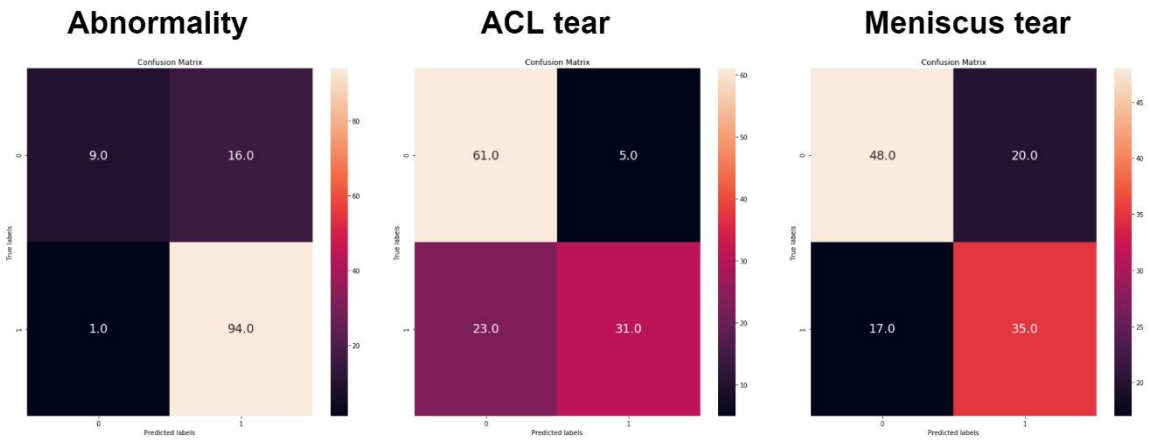


Figure 44: Confusion matrix for abnormal, ACL and meniscus tears detection

4.3.5 Classification performance with augmented images for normal category

For a comprehensive analysis of the effectiveness of WGAN based augmentation on the performance of the classifier, the classification model was trained with original samples along with images generated for the normal category. About 300 MRI scans generated by WGAN for the normal category were added to the original dataset. The model was trained for 50 epochs. Next, the trained classification model was tested on the test dataset.

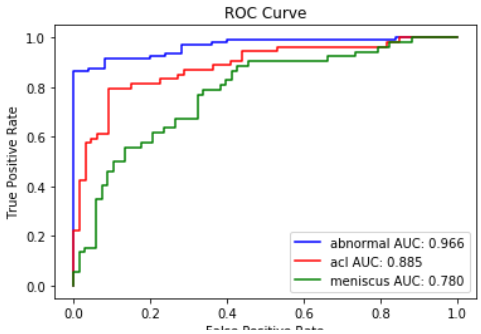


Figure 46: ROC curve on the test dataset

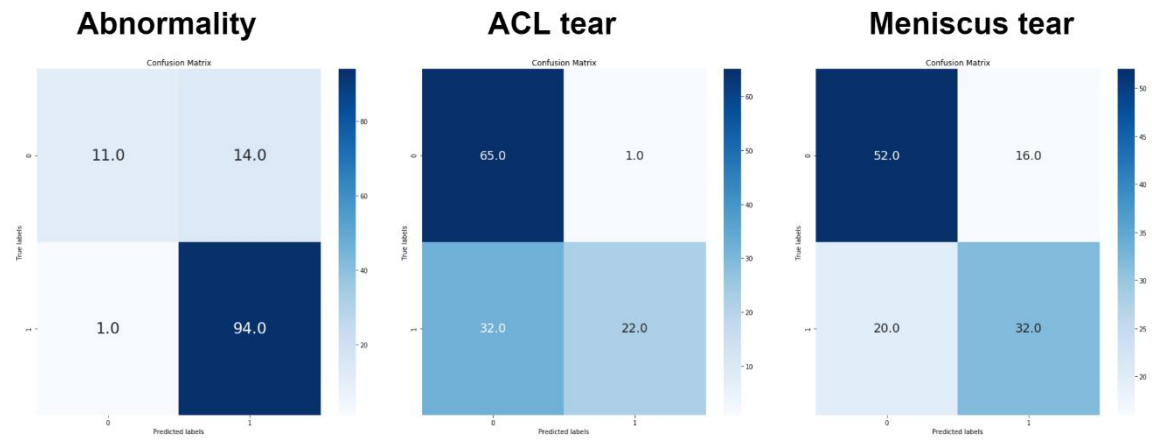


Figure 45: Confusion matrices for abnormality, ACL and meniscus tears

The performance of the classifier was measure using the evaluation metric mentioned in section 3. The model achieved 88%, 73%, and 70% accuracy in abnormal, ACL, and meniscus tears detection. The ROC curve plotted at 0.5 output threshold is depicted in Figure 45. The value of AUC in the ROC curve demonstrates the classifier's potential to discriminate between the cases with the presence or absence of abnormality, ACL, and meniscus tear, which is 96.6%, 88.5%, and 78.0%, respectively.

Confusion matrices, illustrated in Figure 46, show that the cases with abnormality and absence of ACL tear are very accurately classified than all other cases. The model trained with original samples and normal WGAN augmented images can correctly classify 91.6% abnormal and 95.8% ACL cases. The percentage of weighted precision(88%) and F score (86%) is higher for abnormality detection. Overall, the highest classification performance is achieved for abnormality detection.

4.3.6 Classification performance with augmented images for abnormal with meniscus tear category

Finally, the classification model was trained with original samples plus WGAN augmented images for abnormal with meniscus tears category. The ROC curve and confusion matrices, illustrated in Figure 47 and Figure 48, were plotted on the test dataset to visualize the model's performance.

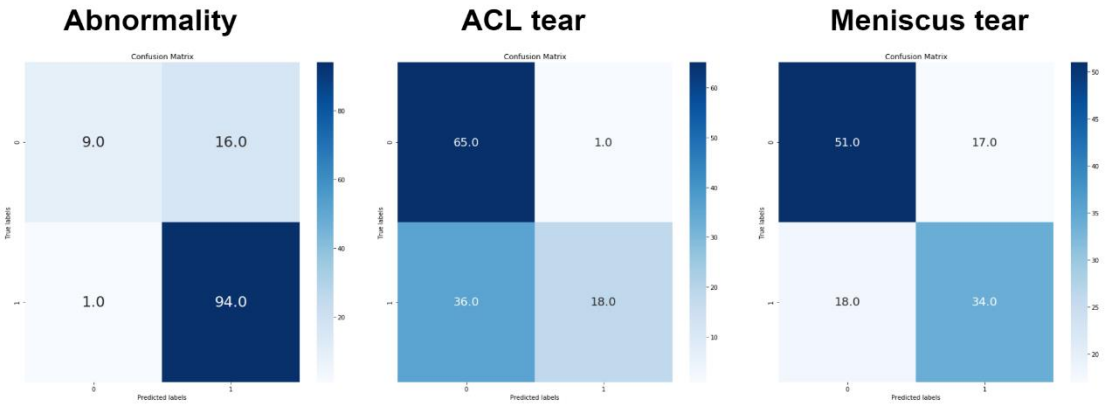


Figure 47: Confusion matrices on the test data

The values for AUC in Figure 48 demonstrate the discriminative power of classification models trained for abnormal, ACL, and meniscus tears detection. The highest specificity and sensitivity were achieved for an ACL tear and abnormal detection. The values of Sensitivity, accuracy, precision, and F score are stated in Table 5.

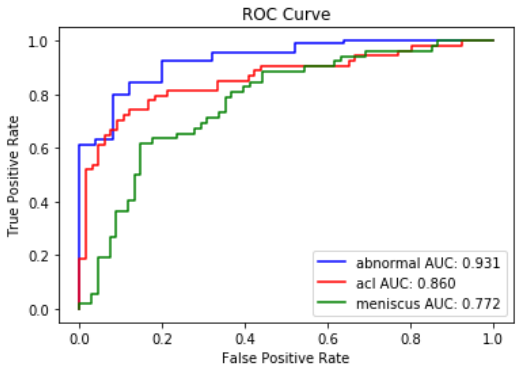


Figure 48: ROC curve on the test dataset

The results achieved by testing the classifier with different types of augmented data are illustrated in Table 5.The last column in Table 5 estimates the variations in accuracy metric by using 95% Wilson score confidence interval.

Table 5: Classification model performance with different types of augmented dataset

Classes	Dataset	AUC	Accuracy	Specificity	Sensitivity	Weighted Precision	Weighted F score	Accuracy with 95% CI
Abnormal	Original samples	93.3%	86%	85.5%	90%	86%	83%	(0.79, 0.91)
ACL		86.5%	73%	95.8%,	67.7%	80%	71%	(0.64, 0.80)
Meniscus		74.7%	68%	66.6%,	69.23%	68%	68%	(0.59, 0.76)
Abnormal	Original with geometrically augmented data	92.2%	86%	84.8%	100%	88%	83%	(0.79, 0.91)
ACL		89.1%	76%	68%	95.8	80%	74%	(0.68,0.83)
Meniscus		77.7%	73%	69%	66%	72%	72%	(0.64,0.80)
Abnormal	Original data with augmented images for normal category	96.6%	88%	87%	91.6%	88%	86%	(0.81, 0.93)
ACL		88.5%	73%	67%	95.8%	80%	70%	(0.64, 0.80)
Meniscus		78%	70%	69%	66.6%	70%	70%	(0.63, 0.78)
Abnormal	Original data with augmented images for abnormal with meniscus tears	93%	86%	85%	90%	86	83	(0.79, 0.91)
ACL		86%	69	67%	95.8%	78	65	(0.60,0.77)
Meniscus		77%	71%	69%	66.6%	71%	71%	(0.62, 0.78)
Abnormal	Original data with WGAN based augmented data	88%	85%	91%	64%	85%	85%	(0.76, 0.90)
ACL		88%	83%	87%	81%	84%	83%	(0.75, 0.89)
Meniscus		78%	66%	58%	78%	70%	66%	(0.57,0.74)
Abnormal	Original data with both geometric and WGAN augmented data	92%	86%	85%	90%	86%	83%	(0.79, 0.91)
ACL		84%	77%	86%	73%	79%	76%	(0.69, 0.74)
Meniscus		79%	69%	64%	74%	69%	69%	(0.60, 0.77)

5 Chapter 5 Discussion (10%)

5.1 Objective 1: Perform data augmentation using traditional data augmentation techniques.

Applying traditional data augmentation was not a complicated task and had been achieved without facing any challenges. First of all, the research studies that performed traditional data augmentation techniques in the medical domain were considered to get the idea of the types of data transformations operated on medical images. Subsequently, few affine transformations, including horizontal flip, rotation, translation, and pixel-level transformations, particularly Gaussian blur, brightness, and contrast, were decided to activate on the images in the knee MRI. All the transformations were applied randomly on the training dataset with the intention to increase the variation.

5.2 Objective 2: Build a generative adversarial network to perform data augmentation.

The generative adversarial network's development to synthesize 2D image slices of resolution 256×256 in the knee MRI series was challenging. Besides, generating images of resolution 256×256 requires large computational resources. This study opted to develop WGAN-GP to generate images in the knee MRI scans. Due to computational resources, memory errors were faced in image synthesis of resolution 256×256. In order to resolve memory errors, reducing the batch size was beneficial but increased the training time of the WGAN model. CNN and ResNet based architectures were tried for the generator and critic networks to examine their impact on the quality of the generated images. For the purpose of making this experiment more efficient, the images were down sampled to 64×64 and fed into WGAN architecture to generate 64×64 images. The high-quality images were obtained with a ResNet based generator and CNN based critic network, which were later used to generated images of resolution 256×256. This trial indicated that deeper generator network architecture results in high quality generated images than smaller generator architectures. The ResNet based generator and CNN based critic network were utilized to generate images of size 256×256. Further, different values for hyperparameters of the WGAN model were tried to generate images of resolution 256×256. Due to limited time and computational resources, only certain values were tried for hyperparameters, and the values which contributed better performance in the first 20,000 iterations were selected. Three WGAN models were trained for approximately 80,000 iterations to generate images of three categories individually. Between 8,000 to 10,000 images were generated for each category. Overall, around 28,000 images from three classes were obtained. The generated 2D images were stacked based on the number of images in real MRI scans. A total of 1000 MRI scans were added to the dataset, which doubled the size of the original dataset.

5.3 Objective 3: Develop a deep learning-based classifier to diagnose knee disorders and Assess the performance of the classifier with different types of augmented data.

For this purpose, the MRNet model, which is proposed in (Bien et al., 2018), was simulated to classify knee MRI scans. The knee MRI dataset is labelled for three types of knee disorders, namely abnormality, ACL tear, and meniscus tear. Each MRI scan has three binary labels regarding each knee disorder, which indicates the presence or absence of a certain knee disorder. Three MRNet models were trained to classify abnormal from non-abnormal cases, ACL tear from non ACL tear cases, and meniscus tear from non-meniscus tear cases, respectively. The hyperparameter values were replicated from (Bien et al., 2018). First, the classification model was trained with the original data. After assessing the classifier's performance with original samples, the classification model was trained using various types of augmented data.

The effectiveness of data augmentation on the classifier performance was assessed by training the classification model with different types of augmented datasets. For the classifier's performance analysis, the performance metrics, including accuracy, AUC, specificity, sensitivity, and F score, were computed. Table 5 shows the details of the performance achieved by the classifier with different types of augmented data. The following examinations are assembled from Table 5.

With respect to AUC, the performance of the classification model with different types of augmented datasets is compared. The AUC values achieved with original data samples are 93.3%, 86.5%, and 74.7% for abnormal, ACL, and meniscus detection, respectively. The traditional data augmentation techniques increased the AUC for ACL and meniscus tears detection from 86.5% to 89.1% and 74.7% to 77.7% while decreased the AUC for abnormal detection from 93.3% to 92.2%. The classification models trained with original samples along with WGAN augmented images for normal category attained an AUC of 96.6%, 88.5%, and 78% for abnormality, ACL tear, and meniscus tear detection, respectively. The performance was increased by 3.3%, 2%, and 3.3% for abnormality, ACL tear, and meniscus tear detection as compared to the performance achieved with the classification model trained with original samples. Next, the classification model was trained using original data, and WGAN generated abnormal with meniscus tear cases and achieved 93%, 86%, and 77% AUC in the detection of abnormality, ACL, and meniscus tear. The model achieved the same performance as the model trained with original data samples for abnormal and ACL tear detection, but for meniscus tear detection, the performance is increased by 3% due to WGAN augmented images of abnormal with meniscus category.

Subsequently, the training of classification was performed with WGAN based augmented data in which data generated for three categories was added to the original dataset. The model achieved an AUC of 88% for both abnormality and ACL tear detection and 78% for meniscus tear detection. As

compared to the AUC values with original data classifier's performance, the AUC for abnormal detection is dropped by 5% while the AUC for meniscus and ACL tears detection is increased by 1.5% and 4%, respectively. Finally, the performance of the classification model was assessed with both geometric and WGAN based augmented images. For abnormal and ACL tear detection, the AUC was decreased from 93.3% to 92%, 86.6% to 84%, while increased for meniscus detection from 74.7% to 79%, respectively, compared to the performance of the classifier trained with original samples.

Overall, all types of augmented data increased the classifier's performance in either one or two tasks, such as in abnormality and ACL tear detection or meniscus tear detection. The WGAN augmented data for the normal category increased the classifier's performance in all tasks (abnormal detection, ACL tear detection, and meniscus tear detection). It indicates that WGAN generated images of the normal category greatly portray the characteristics of the real data. The notable point is that; same WGAN architecture was utilized to generate images for other two categories, but images generated for those categories did not improve the classifier's performance to that extent. The reason behind no improvement in ACL tear and abnormality detection of classifier with WGAN generated cases for abnormal with meniscus tear can be that the generated dataset is not representative of real data as (Gupta et al., 2019) stated that the generated images containing poor biases with respect to real data can have negative impact on the performance of the classification model.

The results obtained from this objective replicate the results obtained by most of the studies in this subject as mentioned in section 2. Based on the studies and obtained results, the GAN based synthetic data helps to increase variation in the dataset and is more beneficial for limited dataset with low variations. However, there are some limitations of GAN based data augmentation which include longer training time and powerful computation resources. (Gupta et al., 2019) mentioned that the images generated by GAN are prone to biases. The main drawback is that biases in GAN generated images are hard to interpret. It would be useful to perform GAN based augmentation with a strategy that can detect and prevent the biases in the generated dataset.

5.4 Answering the research question

Is data augmentation beneficial in developing a robust and accurate computer-aided diagnosis (CAD) system to diagnose knee injuries?

The main goal of this project was to explore the significance of data augmentation on the performance of diagnosis system. All stages in the project research the answer. The results discussed above positively answer the research question. In other words, data augmentation had positive effect in the development of computer aided diagnosis system. Although, there is still room for improvement. The

results can be further improved by exploring different strategies to improve GAN based augmented images.

6 Evaluations, Reflections and Conclusions

6.1 Project goals

The main objective of this project was to investigate the effectiveness of data augmentation techniques on the performance of knee diagnosis system. For this purpose, two types of data augmentation techniques particularly traditional and GAN based data augmentations, were tried and their impact on the performance of deep learning classifier was assessed. Generating data using traditional data augmentation techniques was trivial and took less time to accomplish. On the other hand, generating images using GAN requires high computational cost and longer time to stabilize the training. Due to limited computational resources, it was challenging to train GAN to generate medical images of resolution 256×256. The memory issues were faced during GAN training. Besides, storing GAN generated images of size 256×256 requires high memory. A large amount of the time and effort was used to accomplish this objective. The WGAN model was trained to generate cases with no knee disorders, abnormal with ACL tears and Abnormal with meniscus tear. Finally, the classification model was trained with different types of augmented data and its performance was examined using different performance metrices. For a comprehensive analysis of WGAN data augmentation, the generated cases with respect to each category were added to dataset individually, which were then used to train classification model. The classification model trained with WGAN generated cases of normal category attained the highest AUC value in the abnormality, ACL tear and meniscus tear detection. In contrast, the WGAN generated cases for abnormal with meniscus category decreased the classifier's performance in the tasks of abnormality and ACL tear detection while increased the classifier performance for meniscus tear detection. Overall, the classification model trained with different types of augmented data had higher performance than classifier trained with original samples in either one or two tasks but WGAN augmented images for normal category outperformed baseline model performance in all three tasks.

6.2 Reflections and Future Work

In terms of future work, further investigation can be undertaken to understand the reasoning why classifier performance decreased when augmented images of meniscus and ACL category were utilised. One possible reason can be is that the WGAN models were not sufficiently trained in terms of their number iterations; most studies have trained WGAN models for more than 200,000 iteration (Gulrajani et al. 2017). Furthermore, the Fréchet Inception Distance (FID) score metric was implemented, but it was not utilised during training in order to prevent CUDA Out of memory Errors.

This metric could have allowed the generated images to undergo quality assessment against the actual images during training and thus allow for early bailout. Furthermore, it would be worthwhile to explore the techniques to detect and prevent biases in the GAN generated images. In addition, only one plane of data i.e. Axial was considered during this study, due to time constraints and long training times, and to get a holistic view on whether WGAN augmented images have an impact on classifier performance, all three planes Axial, Sagittal and Coronal should be considered. There are many studies in literature which use 3D models to generate 3D MRI scans such as (Kwon et al., 2019) trained 3D WGAN model to synthesise 3D brain MRI volumes using the smaller training dataset. So, it could be possible to investigate the approaches considered in such studies, implement a 3D based WGAN for each category of injury, and asses the classification performance. This could be done initially for one plane of data, and then progressed to the remaining planes. It is obvious, training numerous 3D WGAN models would require substantial compute power. Moreover, training other GAN variants such as PGGAN to generate knee MRI can be considered as most of the literature studies achieved good classification performance with PGGAN based data augmentation.

In order to consider all or some these improvements, the models can be trained on the cloud either using AWS or Google Cloud Platform (GCP) which provide significant compute power and hopefully lower training times. A start was made on running training jobs on the Google Cloud Platform using the ‘ai-platform’ utility, however due to lack of time this process could not be completed. A Docker file, scripts and other code fragments were created according to GCP standards, to consider training on the cloud. Moreover, if the cloud is to be utilised, then the code written in PyTorch can be modified easily to consider distributed learning i.e. to parallelise learning across several processes and clusters of machines. An alternative option to PyTorch would be to consider implementing the models in Tensorflow 2 because Google Cloud Platform is more complaint with Tensorflow and it also allows for many variants of distributed machine learning.

7 References

Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X. and Xie, Z., 2018. Deep Learning and Its Applications in Biomedicine. *Genomics, Proteomics & Bioinformatics*, 16(1), pp.17-32.

Shorten, C. & Khoshgoftaar, T.M. 2019, "A survey on Image Data Augmentation for Deep Learning", Journal of big data, vol. 6, no. 1, pp. 1-48

Lundervold, A.S. & Lundervold, A. 2019;2018;, "An overview of deep learning in medical imaging focusing on MRI", *Zeitschrift für medizinische Physik*, vol. 29, no. 2, pp. 102-127

Medicalnewstoday.com. 2020. *The Knee: Anatomy, Injuries, Treatment, And Rehabilitation*. [online] Available at: <https://www.medicalnewstoday.com/articles/299204#injury-prevention> [Accessed 29 September 2020].

Chung, D. and Chung, D., 2020. *Knee Joint Anatomy: Bones, Ligaments, Muscles, Tendons, Function*. [online] Healthpages.org. Available at: <https://www.healthpages.org/anatomy-function/knee-joint-structure-function-problems/> [Accessed 29 September 2020].

Kenhub. 2020. *Normal Knee MRI*. [online] Available at: <https://www.kenhub.com/en/library/anatomy/normal-knee-mri> [Accessed 30 September 2020].

Berger, A., 2020. How Does It Work?: Magnetic Resonance Imaging.

My-ms.org. 2020. MRI Plane Mathematics. [online] Available at: <https://my-ms.org/mri_planes.htm> [Accessed 1 October 2020].

Al-Dhabyani, W., Gomaa, M., Khaled, H. and Fahmy, A., 2019. Deep Learning Approaches for Data Augmentation and Classification of Breast Masses using Ultrasound Images. *International Journal of Advanced Computer Science and Applications*, 10(5).

Delplace, A., 2020. Synthetic Magnetic Resonance Images with Generative Adversarial Networks.

Hon, Marcia & Khan, Naimul. (2017). Towards Alzheimer's disease classification through transfer learning. 1166-1169. 10.1109/BIBM.2017.8217822.

Kora, Sagar. (2020). Evaluation of Deep Convolutional Generative Adversarial Networks for data augmentation of chest X-ray images.

Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G. & Nakayama, H. 2018, "GAN-based synthetic brain MR image generation", IEEE, , pp. 734.

Rashid, H., Tanveer, M.A. & Aqeel Khan, H. 2019, "Skin Lesion Classification Using GAN based Data Augmentation", *Conference proceedings (IEEE Engineering in Medicine and Biology Society. Conf.)*, vol. 2019, pp. 916

Gao, J., Jiang, Q., Zhou, B. and Chen, D., 2019. Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: An overview. *Mathematical Biosciences and Engineering*, 16(6), pp.6536-6561.

Bien, N., Rajpurkar, P., Ball, R., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B., Yeom, K., Shpanskaya, K., Halabi, S., Zucker, E., Fanton, G., Amanatullah, D., Beaulieu, C., Riley, G., Stewart, R., Blankenberg, F., Larson, D., Jones, R., Langlotz, C., Ng, A. and Lungren, M., 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Medicine*, 15(11), p.e1002699.

Irmakci, I., Anwar, S.M., Torigian, D.A. & Bagci, U. 2020, Deep Learning for Musculoskeletal Image Analysis.

Nalepa, J., Marcinkiewicz, M. and Kawulok, M., 2019. Data Augmentation for Brain-Tumor Segmentation: A Review. *Frontiers in Computational Neuroscience*, 13.

Medium. 2020. *A Practical Guide To Relu*. [online] Available at: <https://medium.com/@danqing/a-practical-guide-to-relu-b83ca804f1f7> [Accessed 1 November 2020].

Arjovsky, M., Chintala, S. & Bottou, L. 2017, "Wasserstein GAN",.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. Bengio, Y. 2014, "Generative Adversarial Networks", .

Radford, A., Metz, L. & Chintala, S. 2015, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", .

Wu, J., Huang, Z., Thoma, J., Acharya, D. & Van Gool, L. 2018, "Wasserstein Divergence for GANs" in Springer International Publishing, Cham, pp. 673-688.

En.wikipedia.org. 2020. *Receiver Operating Characteristic*. [online] Available at: <https://en.wikipedia.org/wiki/Receiver_operating_characteristic> [Accessed 7 November 2020].

M, H. and M.N, S., 2015. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), pp.01-11.

Stanfordmlgroup.github.io. 2020. *Stanford Machine Learning Group*. [online] Available at: <https://stanfordmlgroup.github.io/> [Accessed 13 November 2020].

Gupta, A., Venkatesh, S., Chopra, S. & Ledig, C. 2019, "Generative Image Translation for Data Augmentation of Bone Lesion Pathology", .

Hussain, Z., Gimenez, F., Yi, D. & Rubin, D. 2017, "Differential Data Augmentation Techniques for Medical Imaging Classification Tasks", AMIA . Annual Symposium proceedings, vol. 2017, pp. 979-984.

Salehinejad, H., Valaee, S., Dowdell, T., Colak, E. & Barfett, J. 2018, "Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks", IEEE, , pp. 990.

Wu, E., Wu, K., Cox, D. & Lotter, W. 2018, "Conditional Infilling GANs for Data Augmentation in Mammogram Classification" in Springer International Publishing, Cham, pp. 98-106.

Wang, Y., Zhou, L., Wang, M., Shao, C., Shi, L., Yang, S., Zhang, Z., Feng, M., Shan, F. & Liu, L. 2020, "Combination of generative adversarial network and convolutional neural network for automatic subcentimeter pulmonary adenocarcinoma classification", Quantitative imaging in medicine and surgery, vol. 10, no. 6, pp. 1249-1264.

Ganesan, P., Rajaraman, S., Long, R., Ghoraani, B. & Antani, S. 2019, "Assessment of Data Augmentation Strategies Toward Performance Improvement of Abnormality Classification in Chest Radiographs", Conference proceedings (IEEE Engineering in Medicine and Biology Society. Conf.), vol. 2019, pp. 841.

Alqahtani, H., Kavakli-Thorne, M. & Kumar, G. 2019, "Applications of Generative Adversarial Networks (GANs): An Updated Review", Archives of computational methods in engineering, .

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. 2017, "Improved Training of Wasserstein GANs", .

Kwon, Gihyun & Han, Chihye & Kim, Dae-shik. (2019). Generation of 3D Brain MRI Using Auto-Encoding Generative Adversarial Networks.

8 Appendix -A: Project Proposal

Data Augmentation for Multi-classification in Medical Imaging

I. INTRODUCTION

Medical image classification has great importance in disease diagnosis and medical treatment. It aims to differentiate medical images based on certain standards, for instance medical pathologies. Interpretation of medical images is time consuming and likely to suffer from inter and intra review variability despite being performed by medical experts (Bien et al., 2018). In order to deal with these downsides, development of automated systems for diagnosis is the field of interest for researchers these days. This subject constitutes the perspective of our project.

Image classification is supervised learning in which multiple images are used as input and output consists of one single diagnostic variable. Remarkably, present research has displayed the capability of deep learning algorithms to unravel the image classification problem. Moreover, deep neural networks consist of multiple layers which enable them to model and interpret the complex relationships in medical images. In recent studies, convolutional neural networks are widely used for classification. Furthermore, Transfer learning plays an important role in this application. In transfer learning existing pre-trained networks are used either for feature extraction or fine-tuning them on medical data for classification. There are many pre-trained architectures of CNN such as Alexnet, VGG16, Resnet50 etc. Even though, Convolutional neural network architectures performed good for medical imaging datasets, but they alone cannot determine the good solution for data. In literature, it has seen that the researchers who achieved good results performed different pre-processing steps along with hyperparameter optimization of deep learning models (Litjens et al., 2017).

There are many challenges associated when applying deep learning models to medical images. Medical datasets are often small which means there is lack of data to train our deep learning models. In supervised learning, the medical image data need to be labelled which require the help of medical experts. Further, the model trained with limited data is prone to overfitting. Gathering medical data is a difficult and costly process that involves the co-operation of researchers and radiologists (Frid-Adar et al., 2018). Moreover, class imbalance is the most faced challenge in medical image classification. As, there are some medical conditions which are commonly faced in hospitals and cause a large amount of data related to them. But some medical conditions are very rare and have very limited amount of stored data. Additionally, deep learning models have low generalization performance with class imbalance dataset. Recent research has used data augmentation techniques as part of data pre-processing to tackle the issues of data scarcity and class imbalance in medical imaging (Wu et al., 2018).

From previous studies, we have seen the potential of deep learning models in image classification along with capability of data augmentation techniques to tackle the challenges associated with image classification task. Our work aims to further explore different data augmentation techniques in order to achieve more robust and accurate deep learning model which can improve diagnosis process. For this purpose, we are specifically interested in

applying geometric augmentation and generative adversarial network (GAN) based augmentation techniques particularly progressive growing generative adversarial network (PGGAN) (Karras et al., 2018) and Wasserstein generative adversarial network (WGAN) (Arjovsky et al., 2017) to knee MRI dataset and compare and critically evaluate the performance of deep Convolutional neural network along with pre-trained convolutional neural network for each type of augmented data. We aim to implement these techniques on knee MRI data set which has three classes (abnormal, anterior cruciate ligament (ACL) tears and meniscal tears) and their distribution is not balanced. Classifying multi classes is more challenging and complex as compared to binary classification task. Our research question follows that:

- Can we build a robust deep learning classifier with the help of data augmentation techniques which can accurately diagnose the abnormalities in Knee MRI?
- Compare and critically evaluate the performance of different deep convolutional neural network architectures with non-augmented data, geometrically augmented data, PGGAN and WGAN based augmented data?
- Is the highest performance achieved by constructing a convolutional neural network from scratch or using pre-trained deep convolutional neural network architecture in knee MRI data?

The motivation to choose this research is to explore the potential of deep learning algorithms in diagnosis of clinical pathologies based on medical images. Our work will have positive impact on human lives by making the process of knee disorder diagnosis more accurate and efficient. Moreover, it will allow radiologists and orthopaedic surgeons to identify the patients with critical conditions and treat them prior to low risk patients in order to narrow the progression of disease (Bien et al., 2018).

The novelty of our project is that data augmentation techniques based on generative adversarial network have not been applied on knee MRI data. Particularly, we will apply PGGAN and WGAN based data augmentation to assess their impact on performance of convolutional neural networks on knee MRI data. Also, we aim to build convolutional neural network from scratch on knee MRI data. Further, pre-trained DCNN based DenseNet-121 will be fine -tuned which have not been tried before on this particular dataset.

Medical imaging techniques such as MRI have been widely used for diagnosis of disease, injuries etc.in healthcare sector. The chosen dataset is collected between January 1, 2001 and December 31, 2012 at Stanford medical health centre. It consists of 1370 knee MRI exams in which 1,104 (80.06%) are abnormal exams along with 319 (23.3%) ACL and 508 (37.1%) meniscal tears. We are interested to develop a robust deep learning model that can classify pathological knee MRI scans into abnormal, anterior cruciate ligament (ACL) tears and meniscal tears. The nature of our problem is supervised and multi-classification due to presence of three labels. Based on data description, we can analyse that there is class imbalance in knee MRI dataset. Moreover, MRI scans are taken from different planes to help radiologists in decision making. For each exam, we have three MRI scans which are taken from three different planes: sagittal, coronal and axial. The knee MRI dataset is divided into training and validation sets by using stratified random sampling. Training set consists of 1130 records of 1088 patients and validation set has 120 exams of 111 patients (Bien et al., 2018).

II. CRITICAL CONTEXT

Data augmentation plays key role in image classification. It helps to address the scarcity of large labelled data to train deep learning model. Moreover, it is not only applicable on limited dataset while it can be implemented on large datasets due to its capability of making data more diverse. Thus, it helps deep learning models to achieve good generalization performance. A common approach of data augmentation is applying different transformation on data for instance horizontal flip, rotation, gaussian blur, translation etc but it does not make data much diverse. Furthermore, transformation of data strongly depends on understanding of data in order to preserve characteristics of data for instance a specific transformation might improve model performance for a dataset, but it might not give very effective results when applying on another dataset (Saad Ali et al., 2019).

To combat these limitations, Generative Adversarial Networks (GAN's) have been widely used for data augmentation in recent studies. GAN's have the capability to automatically learn data space and generate data samples that reserve labels. GAN consists of two networks namely generator and discriminator. Generator network generates images while discriminator network distinguishes the real and fake images generated by generator network. The input of generator network is random noise from some distribution and output is generated image. These generated images and real images use as input of discriminator network that gives the probability of input image being real or fake. During training process, generator becomes good at generating more realistic images while discriminator turns out to be unable to distinguish fake images from real images.

(Frid-Adar et al., 2018) applied data augmentation using GAN to increase data size for classification of liver lesions. Standard augmentation was implemented to enlarge the dataset which was used to train GAN to generate images. Further, GAN was used to generate images for each class separately. Then images generated by GAN were combined with geometrically augmented data and this data was used to train CNN in order to classify liver lesions. A significant improvement in performance was observed by using GAN based augmented data.

(Saad Ali et al., 2019) used GAN for data augmentation in order to tackle the issue of class imbalance in skin lesion classification. The data was augmented by incorporating self-attention mechanism to progressive growing GAN (PGGAN) in order to generate high quality images of resolution 256×256 pixels. For the purpose of training stability of PGAN, two-time update rule (TTUR) was implemented. The main intuition behind PGGAN is that they take low resolution images as input and add layers to generative network in order to increase image resolution progressively along with training stability of discriminator network rises, which lead to synthetic images up to resolution of 1024×1024 pixels. Combining self-attention mechanism with PGAN, assists discriminator to automatically learn to cut off unrelated structures in input image and highlight the

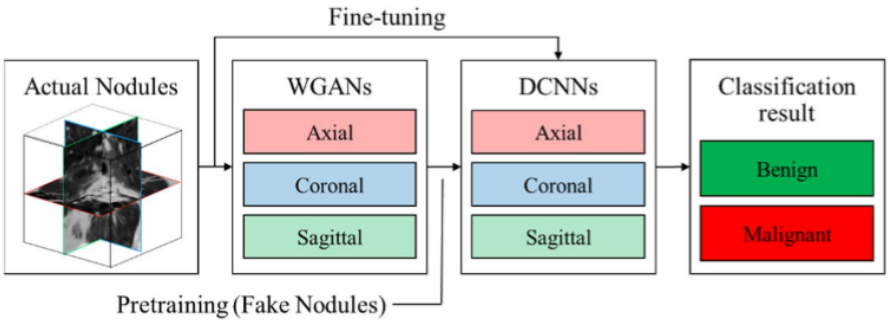


Figure 49: Proposed workflow in (Onishi et al., 2019)

features which are important for a particular task, as a result high quality images are generated. For stable training of GAN, TTUR was used in which learning rate of discriminator was set 5 times more than generator at the same time discriminator is kept with generator step ratio 1:1. Resnet18 was finetuned to classify different type of skin lesion. Data augmentation using PGGAN with self-attention mechanism +TTUR had substantially improved the performance of resnet18 to classify skin lesions compared to standard data augmentation.

Deep convolutional GAN which are used extensively for image synthesizing in medical imaging. They use deep convolutional neural network to define the architecture of generator and discriminator. The distance between two probability densities is expressed using the Jensen–Shannon divergence in training of conventional GAN. They are more likely to face the problems of vanishing gradient loss and mode collapse due to generation of similar images which lead to unstable training process. To resolve these issues, WGAN was invented. Wasserstein distance is used as loss function in WGAN that helps to avoid gradient vanishing problem as well as make training process more stable (Arjovsky et al., 2017).

(Onishi et al., 2019) utilized data augmentation based on WGAN to improve performance of deep convolutional neural network based on Alexnet to differentiate benign and malignant lung CT scans. For this purpose, deep convolutional neural network was pretrained using data generated by WGAN. After that, pre-trained deep convolutional neural network is fine tuned for original data samples to distinguish benign and malignant CT scans. Moreover, the performance of pretrained deep convolutional network is compared with convolutional neural network that is generated from scratch. It had noticed that pretrained model using WGAN based generated images gave significantly better performance as compared to convolutional neural network that is trained from scratch in classification of benign and malignant CT scans. This procedure was performed using one cross section of lung nodule. After that, (Onishi et al., 2019) further extended this work to perform lung nodule classification using different cross sections of lung nodules because observing different cross sections of lung nodules can lead to more accurate diagnostic. First, extraction of volume of interest (VOI) was performed. Then, three slices of image cross sections axial for normal, coronal and sagittal for lung nodule were extracted by means of VOI. The obtained image slices have a smaller number of images which are increased by using WGAN based data augmentation. Moreover, different slice angles regarding three slice images are used to generate multiplanar lung nodule data. WGAN were used to generate images sequentially for each cross section. For lung nodule classification, deep convolutional neural networks based on Alexnet was applied. At first stage, images generated from WGAN were used to pretrain deep convolutional neural networks for each MRI plane (axial, coronal, sagittal) separately and then these pretrained DCNN were fine-tuned by using original images and output obtained from these three DCNN was combined to distinguish benign and malignant images. The classification results significantly improved by using multiplanar images.

It has been analysed that training GAN separately to generate images for each label leads to better classification performance as compared to using one GAN to generate images for multiple classes. DCGAN, WGAN and PGGAN are unconditional data synthesis techniques in which random noise is used to generate images without any conditional information (Yi et al., 2019).

Finally, we consider the previous work that has been done on knee MRI dataset. (Bien et al., 2018) established MRNet in which three convolutional neural networks was trained for each MRI plan and their output was combined by applying logistic regression to classify knee pathologies. Due to limited dataset, a pretrained Alexnet was fine-tuned to initialize weights in earlier layers of CNN to extract useful features while later layers of convolutional neural network were trained and optimized. The area under the receiver operating curve (AUC) is achieved 93%, 96% and 84% for general abnormality, ACL tears and meniscal tears respectively. (Irmakci et al., 2020) used transfer learning to classify knee disorders. Pre-trained convolution neural network-based architectures Alexnet, GoogleNet, ResNet18 were fine tuned for each MRI plan separately and their output was combined using logistic regression for knee pathology classification. Our work aims to make use of transfer learning by applying convolutional neural network based DenseNet-121 for classifying knee pathologies.

III. APPROACHES: METHODS and TOOLS for DESIGN, ANALYSIS and EVALUATION

In this section, we demonstrate stages that we will follow in order to accomplish our project. Also, we will present and justify the techniques that we will employ at each step.

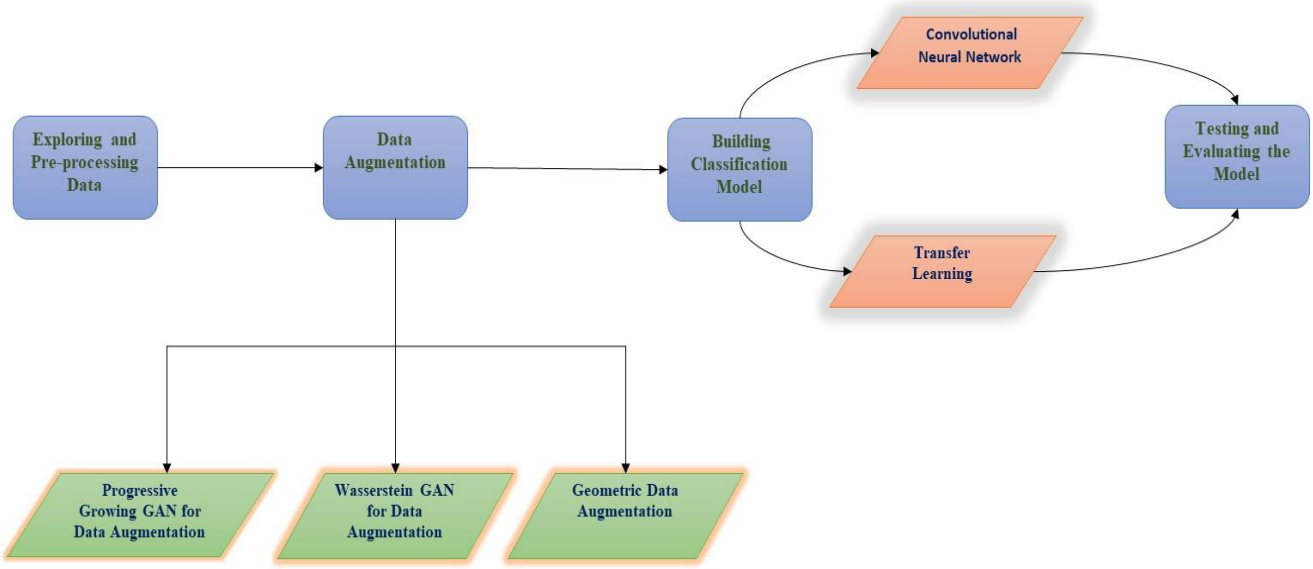


Figure 50: Workflow of the project

Prior to explaining each stage of our project, we will briefly mention the tools that we will use throughout our work. Our preferred choice of programming language is python as it is a high-level programming language and has widely been used to solve problems related to data science and machine learning. Further, it has useful packages which are pre-written programmes and helps to reduce development time. We aim to use NumPy, OpenCV, Matplotlib, Seaborn, and Pillow for image related processing and visualisations. Furthermore, we will also employ the general machine learning library scikit-learn as it contains efficient tools for machine learning and statistical modelling, which include classification, regression, clustering and dimensionality reduction. For the deep learning framework, we have a few choices to consider PyTorch, Keras and TensorFlow. Keras, is a higher-level framework

capable of running on top of TensorFlow, CNTK and Theano. It has gained immense interest for its ease of use and syntactic simplicity, facilitating fast development. In terms of performance it falls short and it does not provide lower-level manipulation and fine-grained control of the neural network architecture and other deep learning aspects compared to its competitors. TensorFlow is a great library and is much better for production ready models, scalability and great performance; however, the learning curve is steep. On the other hand, PyTorch is lower level API that provides users with the flexibility to create custom layers and look more in depth under the hood to perform optimisation tasks. Given that it's slightly more verbose than Keras, still it is easier to learn with on-par performance in comparison with TensorFlow, great for rapid prototyping and is the preferred framework of choice amongst industry professionals and academic researchers. To conclude, PyTorch has been opted for the framework as choice to perform deep learning model building. Given that we are working with GANs and they are notorious to train with moderate to longer training times, thus it's viable to consider using the cloud environment that would allow us to run the code on high specification instances that provide multiple GPUs , significant compute power and a sizeable RAM. Google Cloud Platform (GCP) is our preferred platform of choice due to its advances in machine and deep learning compared to its competitors.

First and foremost, step is to understand the data in order to get familiar with structure of each label image. Moreover, it would be beneficial to spend some time to understand theory behind three pathologies (general abnormality, ACL tears and meniscal tears) the knee MRI scans belong to. In knee dataset, MRI are taken from three different planes axial, coronal and sagittal for each exam.

Figure 51: Different MRI planes in knee dataset (Irmakci et al., 2020)

Medical datasets are mostly stored in uncommon formats. The knee dataset is comprised of a collection of .npy files which can easily be read using python library Numpy. The images stored in this format to decrease their loading time and storage space. The images in knee dataset can easily be visualized by using matplotlib library. Next, we will explore the size, shape of knee images. We will Scale and reshape them into shape that a deep learning model expects.. Further, images will be normalized by using intensity standardization algorithms. Our training and validation sets consist of 1130 and 120 knee exams respectively. Three folders axial, coronal and sagittal are created to store images taken from different planes. We will create folders regarding to each class label in order to store data for each class in separate folders.

In second stage, we will apply different data augmentation techniques in order to enlarge our dataset and to ensure that class distribution is balanced. We aim to apply three different data augmentation techniques to assess their impact on classification of knee abnormalities. First, we will apply geometrical data augmentation which is commonly used in many computer

vision problems. In geometric data augmentation, different transformation such as rotation, shearing, Gaussian noise, translation etc. will be applied on data in order to create diversity in data. Further, we will use more sophisticated technique of data augmentations based on generative adversarial networks. For this purpose, we aim to use WGAN and PGGAN to generate high quality images. Recent studies have used both WGAN and PGGAN for medical data augmentation which leaded to promising results in classification task. WGAN consists of two networks: generator which generates images from random noise and a discriminator that can behave like critic which gives a scalar score in output to determine the realness of input images. For training, WGAN uses Wasserstein-1 loss function that is distance among distribution of real and generated images which helps to avoid the issues of gradient vanishing and mode collapse. Moreover, In WGAN, discriminator or critic function must have to sustain 1-Lipschitz property during training that is achieved by fixing weights within smaller regions. PGGAN are well known in generating high resolution images from random noise. In PGGAN, resolution of generated images increases progressively by adding new layers in generator and discriminator (Yi et al., 2019). For training of GAN algorithms, the loss of both discriminator and generator is calculated. The error of discriminator and generator is backpropagated in order to update their weights. It is noted that optimization of generator and discriminator is accomplished in two steps: We fix discriminator network and update the weights of generator network and vice versa. Training of GAN algorithms is very difficult because an equilibrium is searched between generator and discriminator in optimization process (Chollet, 2018). Hyperparameter optimization plays an important role on performance of deep learning models. We aim to use Bayesian optimization for hyperparameter tuning instead of random grid search due to long training time. Moreover, regularization will be applied on both discriminator and generator networks that helps to avoid overfitting and improves model performance. Our plan is to train WGAN to generate images of each MRI plane (axial, sagittal and coronal) of knee abnormalities (general abnormality, ACL tears and meniscal tears) separately. The same approach will be used to generate images using PGGAN. This part is partially inspired by work of (Onishi et al., 2019).

Third stage of project is building classification model. We will train a convolution neural network from scratch along with pretrained deep convolution neural network architecture DenseNet-121 to classify knee pathologies. We aim to train CNN for each MRI plane (axial, coronal and sagittal) of three classes separately. After that, final classification output will be obtained by integrating outputs of three CNN models which is partly inspired by (Onishi et al., 2019). We will train CNN models with geometrically augmented data, WGAN based augmented data and PGGAN based augmented data. First, we will build baseline models corresponding to each type of augmented data to get the idea of its performance. Hyperparameter tuning and cross validation would be performed on training data to select best model. We aim to use Bayesian optimization to get the optimal values of hyperparameters. Regularization will be used to avoid overfitting. Moreover, we will be fine-tuned pretrained DenseNet-121for each MRI plane (axial, coronal and sagittal) and output obtained by three fine-tuned DenseNet-121 model will be integrated to get final output. In order to assess the performance of DenseNet121 with different type of augmented data, we will use data augmented by geometric, WGAN and PGGAN based data augmentation techniques to fine tune pretrained DenseNet121 model.

Testing and critically evaluating the performance of classification models is last stage in our workflow. Testing model performance on unseen data is crucial step in deep learning. It determines the generalization performance of our model. Moreover, overfitting or underfitting of model is determined by testing its performance on unseen data. We have 120 MRI scans for testing. For evaluating and comparing the performance of different models, we will calculate F score, accuracy, Precision and Recall. In order to visualize the performance of different models, we will plot confusion matrix and ROC curve. After critically assessing the performance of models, we will get answers of our research questions.

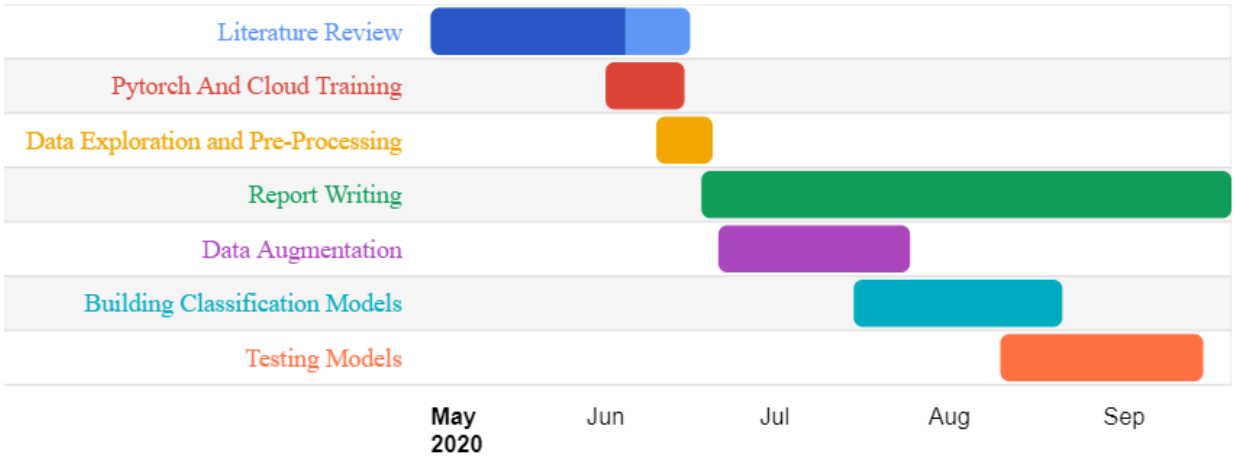
Our project is adoptable from the viewpoint of ethics and privacy rules. The dataset that will use for our research is publicly available for personal research and non-commercial purpose and forbidden for commercial use and monetization. Our project follows this requirement. The chosen dataset does not contain any personal information of patients, so data privacy is guaranteed. Moreover, our research does not require involvement of any external people, so it poses no trouble to anyone.

IV. WORK PLAN

Our project is expected to start around 1st June and to be completed by 1st October. The subsequent table shows the details of each task with its expected start and end date. This work plan might be affected by risks and other unexpected circumstances. That’s why, a flexible amount of time is organized at the end.

Task	Description	Expected Start date	Expected complete by date
Literature Review	Decide for domain and scope of project Outline the research questions Choose suitable dataset Choose appropriate techniques to implement	1/05/2020	16/06/2020
Training to use technical tools	Pytorch training Getting familiar with Google cloud environment	1/06/2020	15/06/2020
Data exploration and pre-processing	**Understanding data** Studying thoroughly about three type of abnormalities in knee data Visualizing knee MRI scans to Get familiar with knee pathologies structure in three MRI planes. **Data Preparation** -Resizing and scaling MRI scans according to requirement of deep learning models -Normalizing MRI scans by using Intensity standardization algorithm. -conversion of pixels to floating point tensors	10/06/2020	20/06/2020
Data augmentation	**Geometrical data augmentation** Apply different transformation on data such as rotation, translation, gaussian noise, adjusting brightness. **Training WGAN and PGGAN for augmentation of knee data** -Construct discriminator and generator network -Train GAN -Configuring Loss function and optimizer for generator and discriminator -Bayesian optimization for hyperparameter tuning -Regularization mechanism **Both WGAN and PGGAN will use to generate axial, coronal and sagittal images of three classes general abnormality, ACL tears and meniscal tears separately.**	21/06/2020	25/07/2020
Building Classification Model	**Training deep convolutional neural network from scratch** -Building baseline convolutional neural network -Configuration of optimizer, loss function -Validating using cross validation -Bayesian optimization for hyperparameter tuning -configuring regularization strategy -Training final convolutional neural network **Use pre-trained network for classification of knee pathologies** -fine-tune DEnseNet-121 using different type of augmented data **We will train a convolutional neural network from scratch along with fine-tune pre-trained network using data augmented by geometric, WGAN and PGGAN augmentation techniques.**	15/07/2020	21/08/2020

Testing and Evaluating the classifiers performance	**Use validation set to test model performance** -Investigate for underfitting and overfitting -Calculate accuracy, f score, recall and precision -Plot confusion matrix and Roc curves to visualize performance of models -Critically evaluate and analyse the model's performance in order to answer the questions	10/08/2020	15/09/2020
Report Writing	Plan and Structure the project report Write about the process that is carried out and critically evaluate the results obtained from our research.	18/06/2020	20/09/2020



V. RISKS

At last, we present the risks that may occur during our research. Table below describes associated risks, their nature, the possibility of their occurrence, their impact on our work and plan to reduce their effect. Possibility and Outcomes columns in table are classified as low, medium, high and very high.

Risks	Nature	Possibility	Outcomes	Plan to mitigate Risk
Difficulty with healthcare domain knowledge	Technical	High	medium	Seek help from supervisor
Hardware restrictions	Technical	High	High	Using Cloud machine learning services provided by AWS or GCP
Unexpected results of GAN algorithms	Technical	medium	Very high	Using other data augmentation techniques such as neural style transfer.
Sickness	Non-technical	Low	Ver High	Flexible time planning
Unavailability of supervisor	Non-technical	medium	medium	Seek assistance from other classmates or professors
Long training time of deep learning models than expected	Technical	high	Very high	Flexible time planning
Failure of hard disk derive	Technical	medium	Very high	Availability of another computer and work will be backed up or carried ou on online storage i.e. One Drive and Google Drive

Coronavirus lockdown can impact meeting with supervisor	Non-Technical	High	Medium	Plan with supervisor to have weekly project catch-ups on Microsoft teem or skype

VI. REFERENCES

Bien, N., Rajpurkar, P., Ball, R., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B., Yeom, K., Shpanskaya, K., Halabi, S., Zucker, E., Fanton, G., Amanatullah, D., Beaulieu, C., Riley, G., Stewart, R., Blankenberg, F., Larson, D., Jones, R., Langlotz, C., Ng, A. and Lungren, M., 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Medicine*, 15(11), p.e1002699.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J. and Greenspan, H., 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321, pp.321-331.

Karras, T., Aila, T., Laine, S. & Lehtinen, J. 2017, Progressive Growing of GANs for Improved Quality, Stability, and Variation. ArXiv, abs/1710.10196.

Arjovsky, M., Chintala, S. & Bottou, L. 2017, Wasserstein GAN. *ArXiv, abs/*1701.07875.

Litjens, G., Kooi, T., Bejnordi, B., Setio, A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B. and Sánchez, C., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis, 42, pp.60-88.

Wu, E., Wu, K., Cox, D. & Lotter, W. 2018, Conditional Infilling GANs for Data Augmentation in Mammogram Classification.

Onishi, Y., Teramoto, A., Tsujimoto, M., Tsukamoto, T., Saito, K., Toyama, H., Imaizumi, K. and Fujita, H., 2019. Automated Pulmonary Nodule Classification in Computed Tomography Images Using a Deep Convolutional Neural Network Trained by Generative Adversarial Networks. BioMed Research International, 2019, pp.1-9.

Yi, X., Walia, E. and Babyn, P., 2019. Generative adversarial network in medical imaging: A review. Medical Image Analysis, 58, p.101552.

Onishi, Y., Teramoto, A., Tsujimoto, M., Tsukamoto, T., Saito, K., Toyama, H., Imaizumi, K. and Fujita, H., 2019. Multiplanar analysis for pulmonary nodule classification in CT images using deep convolutional neural network and generative adversarial networks. International Journal of Computer Assisted Radiology and Surgery, 15(1), pp.173-178.

Chollet, F., 2018. Deep Learning with Python. pp.305-313.

Saad Ali, I., Farouk Mohamed, M., Bassyouni Mahdy, 2019. Data Augmentation for Skin Lesion using Self-Attention based Progressive Generative Adversarial Network. arXiv e-prints arXiv:1910.11960.

Irmakci, I., Anwar, S.M., Torigian, D.A. & Bagci, U. 2020, Deep Learning for Musculoskeletal Image Analysis.

Research Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

http://www.city.ac.uk/department-computer-science/research-ethics

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

PART A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

PART B: Ethics Proportionate Review Form. Students who have answered "no" to all questions in A1, A2 and A3 and "yes" to question 4 in A4 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in such cases that are considered to involve MINIMAL risk. The approval may be ***provisional*** – *identifying the planned research as* likely to involve MINIMAL RISK. In such cases you must additionally seek ***full approval*** from the supervisor as the project progresses and details are established. ***Full approval*** must be acquired in writing, before beginning the planned research.

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*

2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**

3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**