

Medical Image Reader powered by Artificial Intelligence

by

Tanvir Ahmed Palok

19301012

Syumum Ahmed

19101456

Golam Kibria Anim

23341034

Shahed Sharif Bhuiyan Ratul

23341059

Shahreear Alam

19301016

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
September 2023

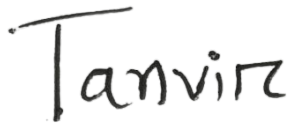
© 2023. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Tanvir Ahmed Palok
19301012



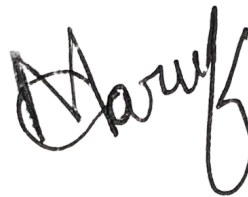
Syum Ahmed
19101456



Golam Kibria Anim
23341034



Shahed Sharif Bhuiyan Ratul
23341059



Shahrear Alam
19301016

Approval

The thesis titled “Medical Image Reader powered by Artificial Intelligence” submitted by

1. Tanvir Ahmed Palok(19301012)
2. Symum Ahmed(19101456)
3. Golam Kibria Anim(23341034)
4. Shahed Sharif Bhuiyan Ratul(23341059)
5. Shahrear Alam(19301016)

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on September, 2023.

Examining Committee:

Supervisor:
(Member)



Nabuat Zaman Nahim
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Golam Robiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Misdiagnosis in medical imaging is a critical concern, risking patients' health due to the pivotal role of radiologists' accuracy in diagnostics. Current cross-checking methods for radiologists' decisions are limited, potentially leading to errors and treatment delays. This study introduces a data processing technique and an advanced prediction system for improving disease detection accuracy in medical images. Our main goal is to contribute to healthcare by developing a system capable of achieving human-level or higher accuracy in disease detection across diverse medical image types. To achieve this, we utilize deep learning techniques, specifically Convolutional Neural Networks (CNNs), and leverage Transfer Learning with pre-trained models. Data processing plays a crucial role, given the importance of image availability and quality. We apply image enhancement techniques such as Histogram Equalization, Adaptive Histogram Equalization (AHE), and Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance image quality and augment a limited training dataset. The advanced ensemble approach significantly enhances the overall accuracy and reduces individual model variance. Validation of our approach using confusion matrices reveals that selective class-wise voting achieves the highest accuracy at 95.27% on the testing dataset. Additionally, our customized weighted voting approach achieves an accuracy of 94.07% on the test set. These results emphasize the effectiveness of our ensemble techniques in improving disease detection accuracy. Our ensemble techniques offer substantial accuracy improvements, promising more accurate and reliable medical diagnoses.

Keywords: Misdiagnosis, Deep learning, Ensemble learning, Confusion matrices, Selective class-wise voting, Histogram equalization, Adaptive histogram equalization, Contrast limited adaptive histogram equalization, Transfer learning.

Dedication

This study is dedicated to all those who lost their lives due to misdiagnosis, to all the patients who are dealing with various fatal diseases all around the world, to all the people who are facing many struggles in their day-to-day lives as a result of misdiagnosis. Hence, we hope our efforts and determination will pave the way of improved treatments for numerous diseases.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our supervisor Mr. Nabuat Zaman Nahim sir for his kind support and advice in our work. He guided us whenever we faced any trouble while doing the research.

And finally to our parents for their support. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iii
Abstract	iii
Dedication	iv
Acknowledgment	v
Table of Contents	vi
List of Figures	viii
List of Tables	x
Nomenclature	xi
1 Introduction	1
1.1 Medical Image and Diagnosis	1
1.2 Research Problem	3
1.3 Research Objectives	3
2 Related Work	4
3 Research Methodology	11
3.1 Data Processing	11
3.1.1 Histogram Equalization	11
3.1.2 Adaptive Histogram Equalization (AHE)	12
3.1.3 Contrast Limited Adaptive Histogram Equalization (CLAHE)	12
3.1.4 Image Normalization	13
3.1.5 Augmentation Techniques	13
3.2 Convolutional Neural Network (CNN)	15
3.3 Transfer Learning	15
3.3.1 EfficientNetV2S	15
3.3.2 InceptionResnetV2	17
3.4 Confusion Matrix	17
3.5 Ensemble Learning	18

3.6	Explainable AI	19
3.6.1	Grad-CAM	19
3.6.2	Back-Propagation	20
4	Implementation	21
4.1	Work Plan	21
4.2	Dataset Collection	21
4.3	Dataset Preparation	22
4.4	Data Processing	22
4.5	Training Models	24
4.6	Proposed Prediction System	25
4.6.1	Majority Voting	26
4.6.2	Selective Class-wise Voting	26
4.6.3	Customized Weighted Voting	27
5	Result Analysis	28
5.1	Learning Curves for Five Approaches	28
5.2	Analysis of the Confusion Matrices	30
5.3	Ensembled Accuracy	34
5.3.1	Validation of the Proposed Data Processing Method	36
5.3.2	Validation of Image Processing	36
5.4	Grad-CAM Visualization	37
6	Conclusion	40
6.1	Challenges	40
6.2	Future Scopes	41
	Bibliography	44

List of Figures

1.1	Severity of the dangers for the misdiagnosis cases according to the study published in 2017	2
2.1	Number of publications since 2010 till 2020 in the PubMed repository, containing keywords related to AI/ML/DL methods in the title and/or abstract.[2]	5
2.2	Accuracy of the models with and without applying image augmentation.[3]	6
3.1	Original X-ray image sample and its Histogram	11
3.2	Effect of applying Histogram equalization on a X-ray image and the Histogram of the processed X-ray image	12
3.3	Effect of applying AHE on a X-ray image and the Histogram of the processed X-ray image	12
3.4	Effect of applying CLAHE on a X-ray image and the Histogram of the processed X-ray image	13
3.5	Architecture of EfficientNetV2S. [21]	17
3.6	Architecture of InceptionResnetV2. [22]	18
4.1	Workplan of the study.	21
4.2	Flow-chart of the Data Processing Method.	24
4.3	Visualization of processed images with different types of parameters.	24
4.4	Flow-chart of the Proposed Ensemble Learning Method (Customized Weighted Voting).	26
5.1	Accuracy curve for Model 1	29
5.2	Loss curve for Model 1	29
5.3	Accuracy curve for Model 3	29
5.4	Loss curve for Model 3	29
5.5	Accuracy curve for Model 5	29
5.6	Loss curve for Model 5	29
5.7	Accuracy curve for Model 6	30
5.8	Loss curve for Model 6	30
5.9	Accuracy curve for Model 7	30
5.10	Loss curve for Model 7	30
5.11	Confusion Matrix for Model 1	31
5.12	Confusion Matrix for Model 3	32
5.13	Confusion Matrix for Model 5	32
5.14	Confusion Matrix for Model 6	33

5.15	Confusion Matrix for Model 7	33
5.16	Confusion Matrix for Customized Weighted Voting	35
5.17	Confusion Matrix for Selective Class-wise Voting Technique	35
5.18	A knee X-ray with identifiers of Osteoarthritis of knee	38
5.19	A knee X-ray with another set of identifiers to detect Osteoarthritis of knee.	38
5.20	Grad-CAM visualizations for another Osteoarthritis detection.	38
5.21	Grad-CAM visualizations of an X-ray image that is also detected 'Osteoarthritis' by the proposed model.	38
5.22	Grad-CAM visualizations of a chest X-ray image that is detected 'Covid' by the model.	39
5.23	Grad-CAM visualizations of a brain MRI that is detected 'Brain Tu- mor' by the model.	39

List of Tables

1.1	Different Types of Medical Imaging	1
3.1	LeNet-5 Layers and Components	16
4.1	Overview of the datasets used for the study.	22
4.2	Number of training samples in datasets <i>mir18</i> and <i>mir18_v2</i>	23
4.3	Overview of the trials of the study.	25
5.1	Record of correct predictions (%) per class for each model	34
5.2	Records of the performance (All accuracy) by the models were stored in a spreadsheet for analysis. A comparison chart is made for understanding the performances before and after applying our data processing techniques. * indicates the base layer of the model is kept trainable.	36

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AI Artificial Intelligence

ANN Artificial Neural Networks

CT Computerized Tomography

DL Deep Learning

ECG Electrocardiogram

ML Machine Learning

MRI Medical Resonance Imaging

PET Positron Emission Tomography

Chapter 1

Introduction

1.1 Medical Image and Diagnosis

With the advancement of science and technology, the medical sector has too become highly developed in the modern era. Broadly, medical sectors can be divided into three correlated parts, they are- diagnosis, analysis and treatment. After being concerned about certain symptoms of one's physical or mental health, he or she visits a doctor. The doctor typically advises some medical tests or medical images. A medical image is a visual representation of the internal body or organs. Each type of medial image machine uses different types of imaging techniques. Hence, in order to find out the appropriate inner visuals, doctors need a certain type of medical image for diagnosing certain diseases or medical problems. Three of the major medical images are familiarized below in a nut-shell:

Image type	Methods	Diagnoses
X-ray	Uses electromagnetic radiations to create images of bones, lungs and other structures	Lung diseases, osteoarthritis, fractures etc
MRI	Uses strong magnetic fields and radio waves to produce detailed images of soft tissues like brain, muscle etc.	Brain tumors, Alzheimer's disease, other tumors (pituitary, meningioma, glioma) etc.
CT Scan	Combines X-rays and computer processings for generating detailed cross-sectional images of the inner body.	Stone/Cyst/Tumor in kidneys, Lung cancer etc.

Table 1.1: Different Types of Medical Imaging

There are many other medical images, such as: Ultrasound, PET scan, ECG, Fluoroscopy etc. Understanding these kinds of images plays a significant role in detecting numerous diseases. The professionals who are trained for this work are called Radiologists. They perform certain actions on a patient's body under the instruction of a physician. Then, the produced image sample is seen, a decision is inferred after a thorough analysis and the report is passed to the doctor. Finally, the doctor provides the most suitable treatment to cure that patient. Here, the whole process relies upon the decision made from the image.

Sadly, this process will not be effective at all, if the diagnosis goes wrong at the very beginning. In many cases, misdiagnosis pushes a life towards death in the end. The study [1] by Seigal et al. (2017) highlights that 1325 claims out of 29,777 medical malpractice filed cases had 'Radiology' as the 'Primary Responsible Service' between the years 2010 and 2014. After a rigorous review, they found that 42% of the claims resulted in high severity clinical injuries including 235 deaths. These numbers are based on a 5-years study over 25 states of the United States of America only. So, it is unimaginable what effect this error from the radiology department has been creating since a very long time. Medical sector is not lagging behind only because of the misinterpretation of the images by the radiologists. The study shows that 23% of the cases had reasons for communication errors from both the ends, patients and radiologists.

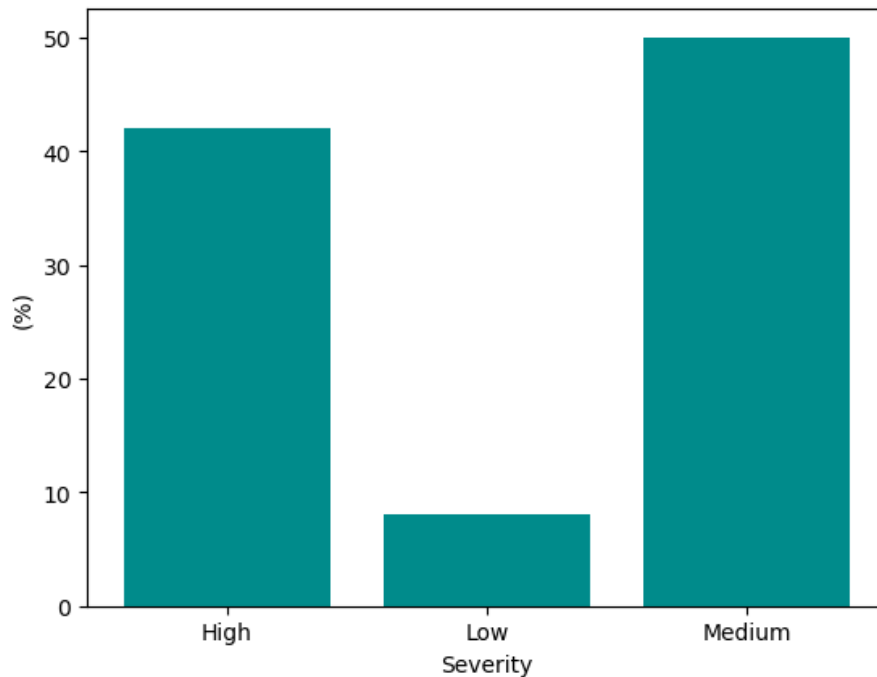


Figure 1.1: Severity of the dangers for the misdiagnosis cases according to the study published in 2017

The interpretations from the medical images are made by radiologists by using just their 'seeing capability'. Sometimes, the professional radiologists oversee a tiny nodule inside the organ which can eventually result in malignancy. The images contain many details that it is not impossible for even a professional to omit this mistake. In an X-Ray, the white/black intensity can help to deduce what is the severity of the danger, a small alien structure in the brain image can easily remain unnoticed by a human vision. A mistake made in the first phase of the disease can lead to a very dangerous case eventually. Every life in this world has the right to live in good health, so it is not expected that a tiny mistake will violate or sabotage some of the lives. Above all, 'to err is human'.

Artificial Intelligence has become the sensation in this era of technology. It has been giving us more and more in every aspect of our lives. Scientists are becoming accustomed to the usage of AI in their own fields of work. Medical department is no different than others. It is being flourished with various facilities with the help of

AI. Yet, some challenges are yet to be overcome. Our research has always focused on the challenge which is introduced above. We expect that the outcome of our research can be the beginning of building a deeply interconnected network between medical practice and AI.

1.2 Research Problem

As depicted in Figure 1.1, the risk factor for every misdiagnosis case is very high, the problem remains in between the radiologists' decision and the doctors' treatment as per the decision. There are currently a very few ways to cross-check the decision made by the radiologist. Most often, the samples are passed to other radiologists for their decisions. In this way, the problem remains unsolved. Again, not every radiologist can analyze all kinds of medical images. Only the trained individuals can perform this action. Hence, hiring trained radiologists or training more people to be good radiologists may or may not enhance the current state of the diagnosis. Moreover, there is no strong universal technical system which can read all kinds of medical images by itself. An automated system can find patterns from a number of samples. Later, it can use the information in reading newer samples. Also, in this way, the system can detect anomaly cases easily. Therefore, if a system can be built which will work between the radiologists and doctors in order to cross-validate the decisions made by the radiologist, it will be able to send a comparison report to the doctor. Finally, the doctor can see both the radiologist's and system's prediction over the sample. Thus, being quite sure of the disease, the doctor can provide proper treatment to the patient from the earliest phase of diagnosis.

1.3 Research Objectives

The main goal of this study has always been to contribute to the healthcare department with a great impact. So, we aim that our study will be helpful in building a system that can detect diseases at a human level or more than that if possible. The system should have the ability to read multiple types of medical images and predict the diseases with the highest accuracy possible. If this system can be developed with an immense amount of data, it can serve the purpose accurately in the diagnosis section. There are advancements occurring in medicine technologies as well. So, with the help of the prediction system we aim to build in future, the healthcare department can successfully flourish with the development it needs now. Moreover, using the prediction system, more and more radiology training programs can be arranged. Summing all of these together, the major problems in the medical diagnosis can be eradicated in future.

Chapter 2

Related Work

Medical imaging is an important part of many clinical uses, including the detection, diagnosis, tracking, and study of different medical diseases. In the past few years, the contribution of machine learning has changed the way of analyzing medical images, moving it from human observation to automatic methods. Experts have been using deep learning too in order to come up with new ways to diagnose and track the status of diseases, which is improving medical practices gradually. In 2021, Barragán-Montero et al. published a review on the usage of AI, ML and DL technologies on medical image analysis [2]. Figure 2.1 highlights how AI technologies have been impacting medical image analysis on the basis of a review on the basis of 10 years of PubMed publications. Researchers are constantly looking into DL methods, which are quickly becoming a popular area of study. DL models have been made and improved to handle complicated medical image datasets. These models use their abilities to learn complex patterns and traits from a vast amount of image data. This makes it possible for detecting the patterns and classifying into some labels with great accuracy. The application of DL in medical image analysis has opened up new ways to make better diagnoses and come up with personalized treatment plans. As DL models can be improved, there are doors to simplify the paths research process, to speed up any diagnosis, and to help healthcare workers in making more informed precise decisions. The DL models that have been made so far have shown that they have a lot of promise to improve how medical images are interpreted. As the field keeps improving, the ongoing study and growth in DL-based medical image analysis holds a lot of potential in it. When these methods are combined with a lot of medical imaging data, it opens the door for improvements in early disease detection, accurate diagnoses, and better treatment strategies. By using the power of DL, medical workers can get useful insights from medical images. This can change the way healthcare is done and, in the end, improve the results in every case.

Two approaches have often been applied to deal with this kind of classification based problems. They are Transfer Learning and Data Augmentation. Training a deep learning model from scratch can be very time-consuming and expensive, especially if a vast amount of data is needed. Transfer learning allows to use a pre-trained model that has already been trained on a huge dataset. In this way, a lot of time and money can be saved. Data Augmentation increases the number of samples in the dataset by making slight variations in the already existing samples. These

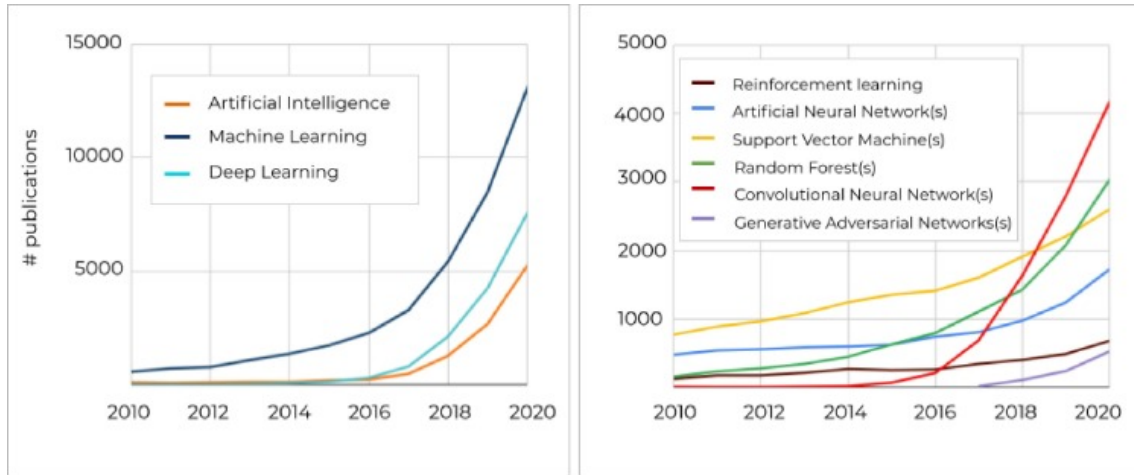


Figure 2.1: Number of publications since 2010 till 2020 in the PubMed repository, containing keywords related to AI/ML/DL methods in the title and/or abstract.[2]

variations can be tweaked a little or more by tuning the parameters for respective techniques. Some of the most used augmentation techniques for image classification are Rotation, Gaussian Blur, Flipping, Noise Injection etc.

Shyni and Chitra (2022) conducted a comparative study [3] of X-ray and CT images in COVID-19 detection using image processing and deep learning techniques. They showed that applying image augmentation bumped up the accuracy of the models highly. Figure 2.2 shows the increase in the accuracy of the models after applying various image augmentation techniques.

The study [4] by Hussein et al. developed a hybrid architecture of CLAHE and CNN which outperformed traditional methods by roughly 20% in terms of accuracy. The authors evaluated three different CNNs for classifying lung diseases from CXR images. The first one was a support vector machine (SVM), which achieved an accuracy of 68%. The second network being a pre-trained VGG19 network achieved an accuracy of 84% after applying CLAHE. The third network was a custom-designed CNN, which achieved an accuracy of 91%. The authors found that the accuracy of the CNN networks decreased as the size of the dataset increased because of the dataset being heavily imbalanced. Despite the limitations, they believe that the proposed model can be used in hospitals, medical clinics, and radiology clinics to assist specialists in identifying lung diseases. The model is able to classify three different types of lung diseases, including COVID-19, pneumonia and tuberculosis. The authors suggested that future research should focus on improving the accuracy of the model and on evaluating the model on larger datasets.

Taresh et al. presented a study [5] on the use of transfer learning to train CNN models for the automatic detection of COVID-19 from CXR images. The authors used a dataset of 5,000 chest X-ray images to train and evaluate three different trained CNNs: VGG16, ResNet50, and MobileNet applying Transfer Learning. The results showed that all of the networks were able to achieve high accuracy in detecting. VGG16 achieved the highest accuracy of 98.28%, followed by ResNet50

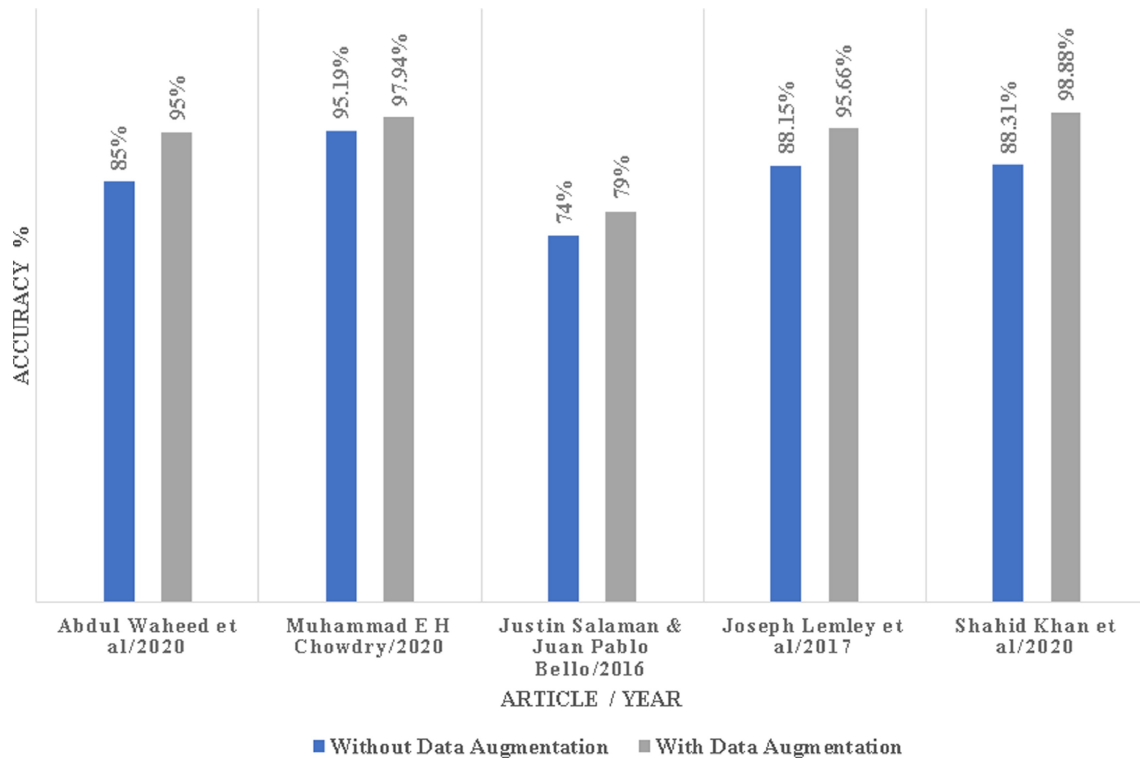


Figure 2.2: Accuracy of the models with and without applying image augmentation.[3]

(98.12%) and MobileNet (97.96%). The authors concluded that transfer learning is an effective way to train CNNs for COVID-19 detection, and that the CNNs they trained are able to achieve high accuracy on both large and small datasets. The authors believe that the results of the study are still significant, and that they provide evidence that transfer learning can be used to train effective CNNs for COVID-19 detection.

Yimer et al. did a research [6] with the aim of improving the diagnosis of lung diseases from CXR images using an AI-based multi-class classification approach. They proposed a method using the Xception model which was trained using labeled image data with data augmentation. They had pre-processed the images with the Median filter to remove salt and pepper noise from the images and CLAHE for enhancing the contrast. The results showed that the Median filter effectively removed noise without compromising edge preservation. The CLAHE method outperformed global histogram equalization and AHE in enhancing image contrast, making it the chosen technique for their study.

Khan et al. presented a study [7] proposed a DL model based on the Xception architecture for detecting COVID-19 cases from CXR images. The authors evaluated the model two different datasets and found outstanding performance on both. They experimented the model by executing 4-class, 3-class, and binary class classification tasks. The model achieved accuracies of 89.5%, 94.59%, and 99%, respectively. Additionally, the model achieved an accuracy of 90% on the second dataset. The recall for COVID-19 cases was also high, indicating a low number of false negatives, which

is crucial in minimizing missed COVID-19 cases. Compared to CovidNet, VGG19 and other CNN models, CoroNet demonstrated higher accuracy in different classification tasks. However, the study faced some challenges including getting the access to a larger and diverse dataset. In summary, the proposed CoroNet model based on the Xception architecture showed superior performance in detecting COVID-19 cases from chest X-ray images, outperforming other state-of-the-art deep learning models.

Rajpurkar et al. developed a CNN named CheXNet [8] which outperformed radiologists on this task, with an F1 score of 0.435 (95% CI 0.387, 0.481). This points that CheXNet was able to correctly identify pneumonia in CXR images with an accuracy of 43.5%, which is significantly higher than the average accuracy of radiologists (38.7%). The authors also extended CheXNet to classify 14 different thoracic diseases, and found that it achieved state-of-the-art results on all 14 classes. This means that CheXNet is able to accurately identify a wide range of thoracic problems, including pneumonia, consolidation, edema, effusion, emphysema, and fibrosis. Comparing CheXNet to the previous state-of-the-art algorithms they found that CheXNet outperformed this model on 13 classes. This suggests that CheXNet is capable of learning more complex relationships between features than those models. The authors trained and evaluated CheXNet on the ChestX-ray14 dataset which contains over 100,000 CXR images with 14 different thoracic pathologies. The authors used the bootstrap method to construct 95% confidence intervals for the F1 scores.

Kumar et al. discussed in their study [9] that the development of an Android app named "Disease Prediction using Artificial Intelligence" (DPAI), which allows users to predict diseases and view disease trends. The app utilizes ML models for predicting diseases, and real-time coefficients and intercepts are fetched from a Firebase database. The accuracy and transparency of the predictions are displayed to the users. The authors splitted the dataset into 65:10:25 for training, cross-validation and testing respectively. They evaluated their proposed model's performance against existing ML models on binary classification problems for predicting diabetes, heart disease, and COVID-19. The proposed model outperformed all the existing models by 1.2746% in terms of accuracy and 1.3926% in terms of F-measure. Overall, the results demonstrate the effectiveness of the DPAI app and its superior performance in disease prediction compared to existing models.

The research [10] done by Saeedi et al. had an aim to develop two DL networks and six machine learning techniques for classifying MRI images into brain tumor categories: glioma, meningioma, pituitary gland tumor, and healthy brain. They used a dataset containing 3264 T1-weighted contrast-enhanced MRI images. Comparisons were made with studies using a well-known MRI dataset for tumor detection, and their study utilized a larger dataset with four categories. The proposed 2D CNN achieved 96.47% training accuracy and 93.45% validation accuracy, while the convolutional auto-encoder achieved 95.63% training accuracy and 90.93% validation accuracy. Precision, recall, and F-measure values for both networks were analyzed for the four classes. Comparisons were made with studies that employed other neural networks for tumor classification, and the proposed 2D CNN performed better, achieving higher accuracy. Additionally, six ML techniques were applied for tumor

classification, and the proposed methods outperformed previous studies in terms of accuracy. Despite the success of the proposed models, a limitation was the small size of medical image databases, which restricted availability for training deep neural networks. Data augmentation techniques were used to address this challenge. Overall, the research presented two effective DL networks and showed promising results in brain tumor classification, with potential applications in cancer detection using MRI or CT scans.

Mohsen et al. presented that the performance evaluation of their proposed methodology [11] was based on a comprehensive set of evaluation metrics, including average classification rate, average recall, average precision, average F-Measure, and average area under the ROC curve (AUC). These metrics were computed for all four classes: normal, glioblastoma, sarcoma, and metastatic bronchogenic carcinoma tumors. Upon thorough evaluation and comparison with other classifiers, the DNN classifier emerged as the most prominent performer, exhibiting superiority across all measured performance criteria. It achieved the highest scores in all the metrics. KNN with $K=1$ and LDA also demonstrated commendable performance, validating their potential in this context. On the other hand, KNN with $K=3$ and SMO displayed relatively lower performance in the given evaluation metrics. These findings underscore the remarkable capabilities of the DNN classifier in accurately classifying brain MRI images, hinting at its profound implications for clinical applications and medical research.

Gaur et al. highlighted an in-depth comparison [12] of methodologies and state-of-the-art models for the analysis of brain MRI image and classification of brain tumors. The comparison shows the effectiveness of DL-based approaches, particularly CNNs, when combined with advanced techniques like PCA and KSVM, leading to improved accuracy and reduced computation time in brain tumor diagnosis. The authors emphasized upon the importance of explainable AI (XAI) in medical image analysis. Hence, they used SHAP and LIME methods to provide the visualization of the learning of CNN models. The visuals explain the contribution of individual features and important regions for classification. The CNN model achieved a training accuracy of 94.64% and an overall test accuracy of 85.37%, with 26 wrong predictions. K-fold cross-validation demonstrated almost 100% training accuracy.

Alsubai et al. chose DL models, CNN and CNN-LSTM, to apply their proposed technique for brain tumor classification [13]. The performance evaluation of the CNN model achieved a training accuracy of 99.4%, validation accuracy of 98.3%. The evaluation of the CNN-LSTM model shows that it outperformed the CNN model with training and validation accuracy of 99.8% and 98.5%, respectively. The CNN model achieved an accuracy of 98.6%, precision of 98.5%, recall of 98.6%, and F1-measure of 98.4%. On the other hand, the hybrid CNN-LSTM model performed even better, with an accuracy of 99.1%, precision of 98.8%, recall of 98.9%, and F1-measure of 99.0%. The pre-processing steps are the same as those used in both of the techniques. In summary, the proposed CNN-LSTM model exhibits outstanding results in brain tumor classification, outperforming previous techniques and demonstrating its potential for accurate and efficient diagnosis.

Khan et al.[14] proposes two deep learning models for brain tumor classification. One of them is 23-layered CNN and the other Fine-tuned CNN with VGG16. The 23-layered CNN achieved an overall prediction accuracy of 97.8% on the *Figshare* dataset and 100% on the *Harvard Medical* dataset. The "Fine-tuned CNN with VGG16" achieved 100% accuracy on dataset 2 although some overfitting issue was seen. Existing state-of-the-art methods were outperformed by the proposed models for binary and multi class tumor classification. Data imbalance was addressed as one of the major challenges in executing the study. The models demonstrated high accuracy, precision, recall, and F1-measure. Comparisons with other studies showed the superiority of their proposed models.

Noreen et al.[15] used DL models, specifically DenseNet and Inception-v3 architectures and presented a study on the classification of brain tumor. Specific hyper-parameters, such as a learning rate of 0.0001, 100 epochs, and an Adam optimizer were applied while training the models. The authors split the dataset into 80:20 for training and testing respectively. They explored various bottom layers of the Inception-v3 and DenseNet201 blocks to extract features for brain tumor classification. The results of this study demonstrate that a combination of features extracted from various layers of Inception-v3 and DenseNet201 significantly improves classification accuracy compared to individual block feature extraction methods. The proposed approaches achieve high accuracies of up to 99.51%. The ROC curves indicate outstanding performance for glioma and pituitary tumor classes, while meningioma being less effective. The study further analyzes feature maps and discusses the challenges of classifying brain tumors due to their diverse shapes, sizes, and positions within MR images.

Yildirim et al. presented in their study [16], a DL model was developed to detect kidney stones from CT images of abdomen. They used 1453 CT images splitting them into 80:20 for the training and validation purposes. After training for 40 epochs, the model's performance was evaluated on 346 unseen test images. The model scored an accuracy of 96.82% and a recall of 95.76%. To show the interpretability of the model's decision-making process, they applied Grad-CAM to visualize the regions of interest on the images where the model focused to obtain the prediction label. The model successfully detected kidney stones in most cases, however, some images were misclassified due to factors like the presence of rib tips, calcified areas, and other organs in the image. The authors expect future work could involve collecting images from different sources to validate the model's performance in diverse settings. Additionally, both axial and sagittal planes of image could be used to evaluate the model's performance.

The study conducted by Shakeel et al.[17] presented that the Improved Profuse Clustering Technique (IPCT) and Deep Learning with Instantaneously trained neural networks (DITNN) can improve lung cancer detection and classification system using CT images. They tested both these approaches in respect of various segmentation metrics, outperforming other segmentation methods like fuzzy c-means, global threshold, watershed, and Sobel. The IPCT method gained high scores in all the metrics. For the classification of lung cancer, different techniques were compared, including Radial Basis Neural Network (RBNN), Convolution Neural

Networks (CNN), Hopfield Neural Network (HNN), Learning Vector Quantization (LVQ), and the proposed DITNN. The DITNN achieved the lowest error rate in detecting lung cancer. The deep learning approach in DITNN demonstrated its effectiveness in feature extraction and matching. The DITNN outperformed other methods in all metrics, achieving an overall precision of 98.43%, recall of 98.36%, and F1-measure of 98.42%. This study's outcome suggests that the DITNN approach is highly reliable for lung cancer detection and classification.

Islam et al.[18] reported that the Swin Transformer can become one of the most effective options for diagnosing kidney diseases from CT images. The authors highlighted a comparison of six different models including InceptionV3, EANet, Resnet50, CCT, VGG16, and Swin Transformers for classification of 3 types of kidney problems on CT images using all the major evaluation metrics by predicting the models on unseen data. They employed Tenfold cross-validation to obtain averaged results. Among the models, the Swin Transformer achieved the highest accuracy of 99.30%. With an accuracy of 61.60%, InceptionV3 showed the worst performance. The Swin Transformer gained the highest recall for kidney cyst, normal, stone, and tumor class images, with values of 0.996, 0.981, 0.989, and 1, respectively. It was particularly effective at detecting kidney stones and tumors. Resnet50 created the least impact in detecting kidney tumors and stones, with recalls of 0.295 and 0.462, respectively. In terms of precision, the Swin Transformer also achieved the highest values for all classes, with an average score of 0.992. The highest F1-score for all classes was observed in the Swin Transformer model, with values of 0.996, 0.998, 0.985, and 0.996. The GradCam analysis of the InceptionV3, VGG16, and Resnet models provided the visuals of the models' decision-making processes. The Swin Transformer showed focused attention on small regions of interest which led to more accurate predictions than others.

Gharaibeh et al.[19] stated that DL has shown promising results compared to traditional ML methods in classifying kidney tumors from CT images. Their study consisted of 3 parts including classification and segmentation of tumors based on CT images. They used 109 CT scans for their study. They found that usage of ML techniques achieved highest accuracy of 95%. On the contrary, DL approach scored an accuracy of 97.3% on 369 CT scans aiming to classify between the same classes. Their segmentation studies focused on detecting the tumor nodules within the kidney and they achieved accuracy rates of 97.7% and 96.9% using V-Net and 3D U-Net both based on 210 CT scans. Some studies combined classification and segmentation models which scored accuracy rates of 90-99% on 300 CT scans. They also experimented a multi-model study using SVM, CNN, and InceptionV3, which achieved an accuracy rate of 93.39% on 196 CT scans. The authors report a challenge that they faced is the limited availability of medical image data. This limitation led to risks of overfitting which reduced the performance of the models. The authors expect future works based on this work should include trials of using smaller models and a proper augmentation of the images.

Chapter 3

Research Methodology

3.1 Data Processing

The main material required for our study is vast amount of medical images. After a thorough lookup online, we found several images. However, the amount of the images were not enough for a neural network to properly learn the patterns from the images. Moreover, the quality of them were not good enough to distinguish the correct information needed for classification. Hence, we applied some image processing techniques to enhance the quality of the images and then some augmentation techniques to increase the amount of total trainable images. All these techniques are briefly introduced in the following subsections.

3.1.1 Histogram Equalization

Histogram equalization is a technique for enhancing the contrast of an image. It increases the global contrast of an image with close contrast values by spreading out the most frequent intensity values. Thus, this technique increases the local contrast value where the value is notably low. When working with colorful images, applying histogram equalization separately to the Red, Green, and Blue (RGB) components can affect the color balance of the image [20]. Figure 3.2 represents an image from our dataset where we applied Histogram Equalization and the Histogram for the image.

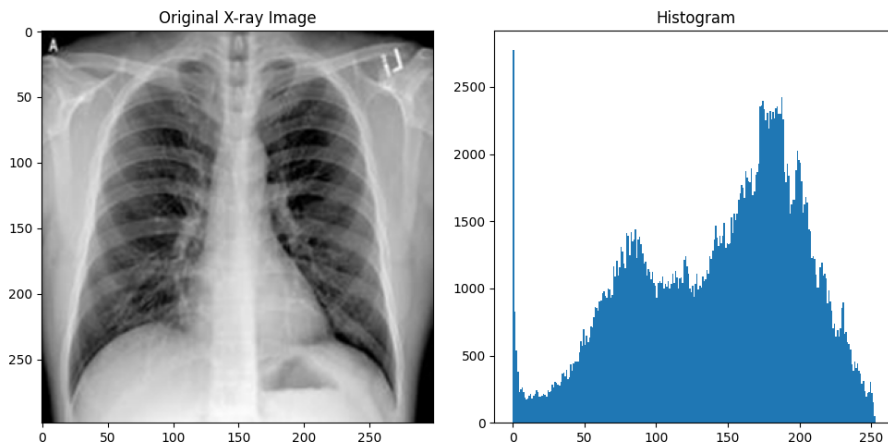


Figure 3.1: Original X-ray image sample and its Histogram

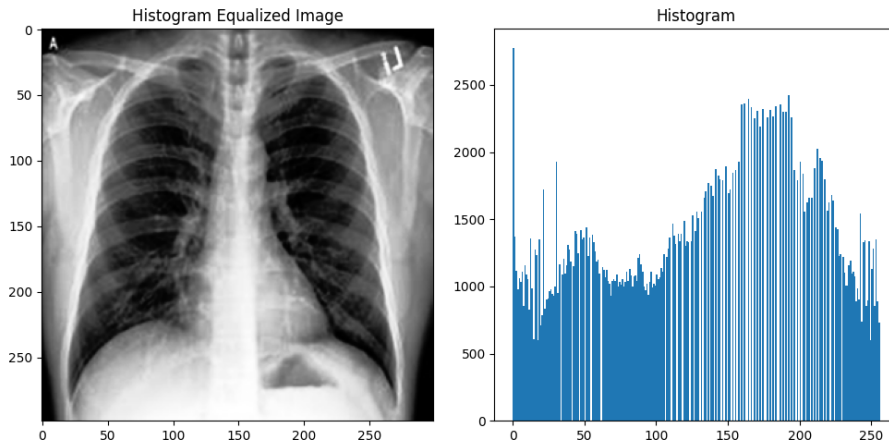


Figure 3.2: Effect of applying Histogram equalization on a X-ray image and the Histogram of the processed X-ray image

3.1.2 Adaptive Histogram Equalization (AHE)

Unlike ordinary histogram equalization, Adaptive Histogram Equalization, shortly known as AHE, generates multiple histograms of an image. Each histogram corresponds to a specific part of that image. After a computation taking all the generated histograms, this technique adjusts the brightness of the image. This ultimately enhances local contrast and defines edges in different regions of the image. Moreover, it can adapt to the equalization process for the characteristics of each area, which result in more refined and region-specific contrast enhancement [20]. Figure 3.3 represents an image from our dataset where we applied AHE and the final Histogram for the image.

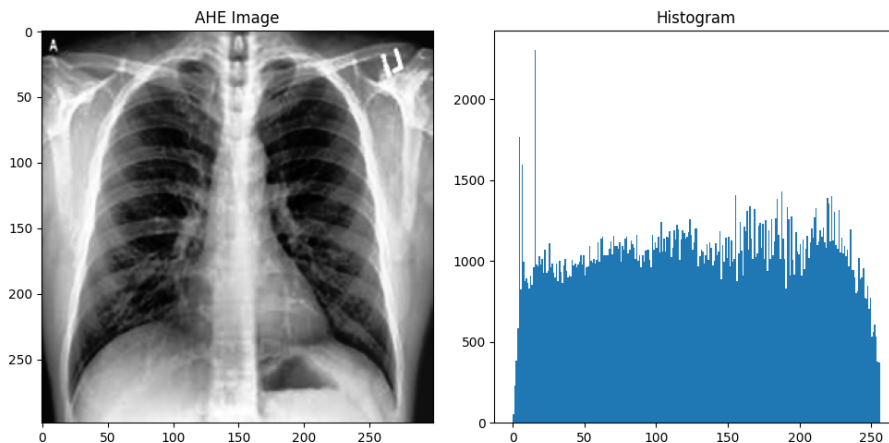


Figure 3.3: Effect of applying AHE on a X-ray image and the Histogram of the processed X-ray image

3.1.3 Contrast Limited Adaptive Histogram Equalization (CLAHE)

Contrast Limited Adaptive Histogram Equalization, also known as, CLAHE, distinguishes itself from regular adaptive histogram equalization with its ability to limit the contrast adjustment. Using this technique, the contrast limiting process is ap-

plied to each local region of the input image. This produces the derivation of a specific transformation function. CLAHE was primarily introduced to encounter the challenge of over-amplification of noise that could be generated with traditional AHE methods. CLAHE has the ability to effectively control that amplification of contrast by limiting the contrast to each region so that prevention of the adverse impact of noise exaggeration can be addressed. This unique characteristic makes CLAHE particularly well-suited for applications in which the preservation of image details is crucial, as it maintains better balance between contrast enhancement and noise management. Hence, CLAHE has found widespread utilization in various image processing tasks, such as medical imaging, surveillance, and other fields where enhancing local contrast while minimizing noise artifacts is of paramount importance [20]. Figure 3.4 represents an image from our dataset where we applied CLAHE and the final Histogram for the image.

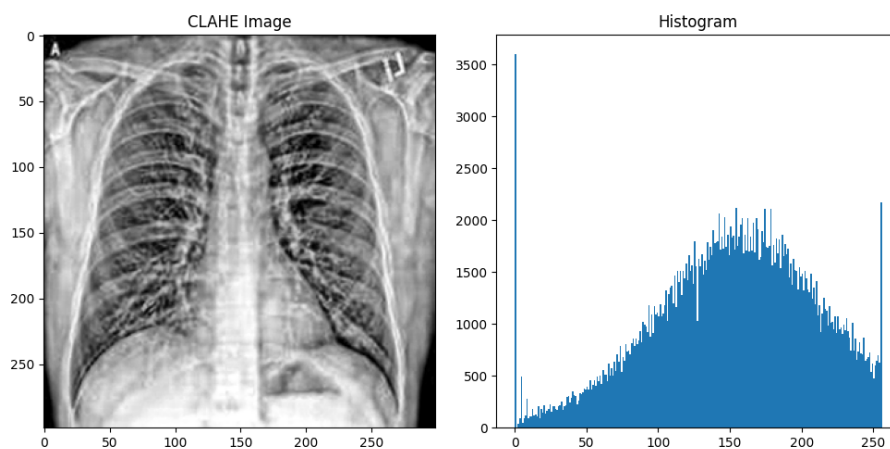


Figure 3.4: Effect of applying CLAHE on a X-ray image and the Histogram of the processed X-ray image

3.1.4 Image Normalization

Image Normalization is a process in image processing that changes the range of pixel intensity values. This process aims to convert an input image into a range of pixel values that are more normal to the senses, hence the term normalization. The linear normalization of a digital image is performed following this formula:

$$\text{Output_channel} = 255 \cdot \frac{\text{Input_channel} - \min}{\max - \min}$$

For a greyscale image, we only need to normalize using one channel. However, for an RGB image that contains 3 color channels, we need to normalize every channel using this formula.

3.1.5 Augmentation Techniques

By augmenting the training dataset with diverse variations of the original images, these techniques introduce valuable variations that make the model familiarised with immense amount of image which gives the model robustness and increase its impact

on prediction. There are numerous techniques for augmenting images. Following is a brief discussion on the techniques that we have applied on our training samples. Overall, image augmentation techniques are a fundamental section in the workflow of a deep learning study as it allows models to learn a wider range of features and achieve superior performance on various tasks.

Rotation

It is one of the positional augmentation techniques. Rotating images introduces geometric variations of the image to the model. Rotated images enable the model to recognize objects from different perspectives.

Gaussian Filter

Gaussian filter helps to reduce the amount of noise and smooth out pixel values in an image. This is particularly useful for reducing the impact of small, irrelevant details or noise on the model's performance. It aids the model in focusing on the essential features of an object and improves its ability to generalize to new, noisy inputs.

Scale

Scaling images to different sizes allows the model to recognize objects at various scales. Particularly when objects in real-world scenarios may appear at different distances or sizes, this technique comes in handy. By training with scaled images, a model can gain a better view of objects varying dimensions.

Shear

A model may have to recognize objects in distorted or skewed scenarios. Sheared images are beneficial in this particular scenario. It helps in enhancing the model's robustness over other forms of geometric deformations of the input image.

Jitter

Jittering refers to adding random noise to the pixel values of an image. This assists the model to become tolerant to noisy variations of the image.

Flipping

Flipping is another positional image augmentation technique. Flipping can be done by two ways- Horizontal or vertical. Horizontal or Vertical Flipping mirrors the image and enables the model to recognize objects irrespective of their left-right or up-down orientation respectively. This augmentation is particularly useful for tasks where object orientation does not affect the overall context.

Sharpen

Sharpening an image enhances the edges and fine details. This technique can make the model more sensitive to object boundaries and improve its ability to understand the pattern more accurately.

3.2 Convolutional Neural Network (CNN)

A type of deep neural network known as a convolutional neural network (CNN) is typically used to analyze visual images. It utilizes an exceptional strategy called Convolution. Convolution is a mathematical operation on two functions that results in the creation of a third function that describes how one function is altered by the other. Multiple artificial neuron layers build a CNN. Each layer in a CNN generates a number of activation functions for the next layer when an image is input. Typically, basic features like horizontal or diagonal edges are extracted in the first layer. This output is sent to the next layer, which looks for more complicated features like combinational edges or corners. It is able to identify even more complex features, such as faces, objects, and so on, as we progress deeper into the network. The classification layer generates a set of confidence scores (values between 0 and 1) that indicate the image's likelihood of belonging to a "class" based on the activation map of the final convolution layer. For instance, the output of the final layer of a CNN that detects horses, dogs, and cats is the possibility that the input image contains any of those animals. CNN models have several common layers. The LeNet-5 CNN architecture is the pioneer to modern CNNs. A discussion regarding its 7 layers are highlighted in table 3.1

3.3 Transfer Learning

Transfer Learning is a widely used technique in the modern realm of AI, especially in the domain of deep learning. It involves leveraging the knowledge gained from a pre-trained model on one specific task and then applying that knowledge to improve the performance on a different and new model. The pre-trained model works as a feature extractor for the new task. By leveraging the learned representations from the previous task, the model can generalize better on the new task, even when data for the new task is limited or scarce. This is particularly advantageous because it allows for the efficient reuse of knowledge and effectively reduces the need for large amounts of training data for the new task. Moreover, it saves a lot amount of time in the training phase as well. Hence, it has become a fundamental approach across many domains in classification problems.

3.3.1 EfficientNetV2S

Figure 3.5 contains the architecture of EfficientNetV2S model. It is one of the two pre-trained models that we kept in our study.

Table 3.1: LeNet-5 Layers and Components

Layer	Description
Input Layer	An input image with a shape of 32x32 pixels.
Convolutional Layer	This layer extracts various features from the input image. Sliding filters of size 5x5 are applied to the input image, and the dot product is computed between the filter and corresponding parts of the image. This produces 6 feature maps of size 28x28x6. The output of this layer provides information about corners and edges of the image.
Pooling Layer	The pooling layer, a fundamental layer of CNN, is used to solve computer vision tasks, including image segmentation, image classification, and image detection. This layer works by downsampling the previously generated feature map, which has two types: max pooling and average pooling. Max pooling proceeds to preserve the most essential features and reduce the spatial dimensions. Meanwhile, average pooling helps reduce sensitivity to noise in the input. A pooling layer is implemented in each channel of the input feature maps, resulting in a reduced-size feature map. In a nutshell, the layer reduces the spatial dimension of the input feature maps. This downsampling helps improve translation invariance, computational complexity, select essential features, etc.
Convolutional Layer	This layer performs another convolution with 16 filters of size 5x5 on the feature maps from the previous layer, followed by another Pooling Layer. This further reduces the size of feature maps to 5x5x16.
Fully Connected Layer x2	The neurons and the weights and biases that make up the Fully Connected (FC) layer are used to connect the neurons between two distinct layers. In most CNN architectures, these layers come before the output layer and make up the last few layers. The FC layer receives the flattened input image from the preceding layers. The classification process gets under way at this point. This layer is a fully connected convolutional layer with 120 filters of size 5x5. Each of the 120 units in this layer is connected to the 400 (5x5x16) units from the previous layer. The layer performs numerical operations and serves as a precursor to the classification process. This layer follows another FC layer next to itself with 84 units, providing additional computations for classification.
Output Layer	The seventh layer, a softmax output layer with N possible classes based on the number of classes in the dataset. The final output of the network after passing through the previous layers will be the predicted class probabilities for the input image.
Dropout Layer	Overfitting occurs when a model performs poorly when applied to fresh data while it performed well on the training data. A dropout layer is used to solve this issue, in which a few neurons are removed from the neural network during the training process, resulting in a smaller model. For example, 30% of the nodes in the neural network are removed from the network at random if a dropout of 0.3 is applied. A ML model's performance is enhanced by dropout because it simplifies the network and prevents overfitting.
Activation Functions	The activation function at the network's end determines which model information is to pass forward and which is to fire back. The web gets non-linearities as a result of this. The activation functions commonly used are the ReLU, Softmax, tanH, and Sigmoid functions, and there is a function for each part. In a CNN model, the sigmoid function is used for binary classification. The softmax part is used for multi-class types.

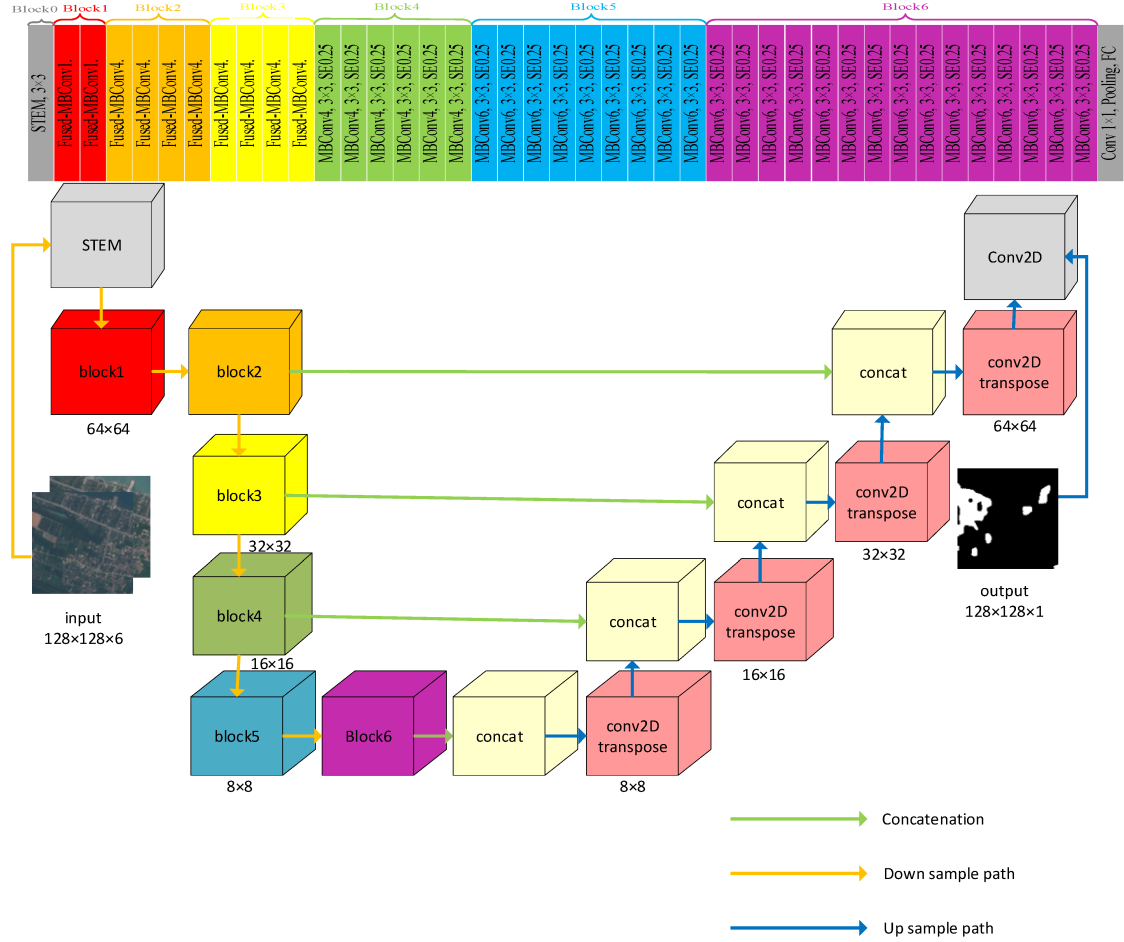


Figure 3.5: Architecture of EfficientNetV2S. [21]

3.3.2 InceptionResnetV2

Figure 3.6 contains the architecture of InceptionResnetV2 model. It is one of the two pre-trained models that we kept in our study.

3.4 Confusion Matrix

The confusion matrix is an evaluation technique to justify a model’s performance in classification tasks by breaking the prediction into four categories: true positive, true negative, false positive, and false negative. Usually, the method assists in computing accuracy, precision, f1 score, and recall. The hypothesis obtained through the technique helps fine-tuning and also helps to make the models more efficient, resulting in more vigorous decision-making ability. In addition, the diagonal value of the confusion matrix represents the count of the model’s correct prediction, true positive and true negative values. True positive values indicate the samples belong to positive classes that are also classified correctly as positive class instances. Meanwhile, true negative values are those values that belong to the negative classes that are classified correctly as negative class samples. Moreover, the values above the diagonal represent the samples that belong to the negative class but the model

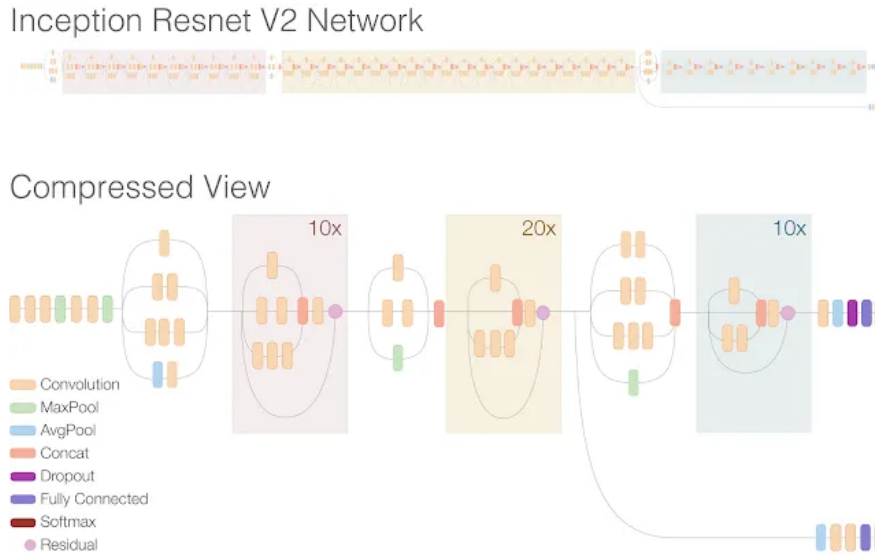


Figure 3.6: Architecture of InceptionResnetV2. [22]

classified them as positive class; meanwhile, the values below the diagonal represent samples that belong to the positive class but are classified as negative class instances. Thus, by visualizing the confusion matrix we can interpret models classifying ability per class as well as get an overall view of the result.

3.5 Ensemble Learning

Ensemble Learning in the context of deep learning refers to a powerful technique where multiple deep learning models are combined to create a single, stronger predictive model. The idea behind ensemble learning is that by leveraging the diverse strengths and weaknesses of individual models, the overall performance and generalization capability can be significantly improved. In the context of deep learning, ensemble learning can be achieved in various ways, such as- Bagging, Boosting, Stacking. Ensemble Learning can lead to improved performance and robustness by reducing overfitting, capturing complementary patterns in the data, and enhancing the model's ability to generalize to unseen data. It is particularly effective when individual models have varying strengths, as the ensemble can leverage these diverse aspects to achieve better overall performance. However, it's important to note that ensemble learning can be computationally expensive and may require additional resources for training and inference. Hence, this method should be in consideration where accuracy of the prediction needs more priority than the complexity of the whole system.

3.6 Explainable AI

Explainable Artificial Intelligence (XAI) is a bunch of cycles and strategies that permits human clients to understand and believe the outcomes and results made by AI calculations. Explainable artificial intelligence (XAI) is utilized to depict a AI model’s learning. It describes model exactness, decency, straightforwardness and results in simulated intelligence controlled direction. XAI is urgent for an association in incorporating trust and certainty while putting computer based intelligence models into creation. Man-made intelligence reasonableness likewise assists an association with taking on a mindful way to deal with man-made intelligence improvement. Furthermore, not even the specialists or information researchers who make the calculation can comprehend or make sense of what precisely is going on inside them or how the simulated intelligence calculation showed up at a particular outcome. There are many benefits of understanding how an AI empowered framework has prompted a particular result correctly.

3.6.1 Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is a cutting-edge technique that leverages the gradients of target concepts within a deep-learning model to produce localization maps. These maps highlight the crucial regions within an image that contributed to the model’s predictions. Grad-CAM is renowned for its versatility and accuracy, making it a valuable tool for interpreting complex CNNs. Our research applied Grad-CAM to the EfficientNetV2-S model to visualize the regions of interest in medical images that influenced the model’s diagnostic decisions. By doing so, we aimed to shed light on the ‘black box’ nature of deep learning models in medical image analysis. Results and Insights Initial Grad-CAM Visualization We began by generating Grad-CAM heatmaps for the last convolutional layer (‘conv2d_1’) in our model. This layer was expected to provide the most accurate visual explanation of the classified object. However, the initial results could have been more precise, with the heatmap encompassing multiple regions, including parts of the background and unrelated things. This led us to question what the model was considering when diagnosing. Analyzing All Model Layers To gain a deeper understanding, we extended our analysis to visualize Grad-CAM heatmaps from all model layers. We observed distinct patterns: Early layers (blocks 1 through 3) detected contours and borders. Middle layers began identifying relevant concepts. Late layers incorporated spatial information from early and concept-based leads from the middle layers. This analysis revealed the model’s progression in identifying critical features within images. This research demonstrates the significant potential of Grad-CAM as a tool for enhancing the interpretability of deep-learning models in medical image analysis. By visualizing the decision-making process, we gained insights into model strengths and weaknesses, which allowed us to improve model performance. Our findings highlight the iterative nature of machine learning, where models are continually refined and validated. These techniques will contribute to more transparent and trustworthy AI models in medical diagnosis, ultimately benefiting patients and healthcare professionals. We plan to refine our models further, explore additional interpretability techniques, and collaborate with medical practi-

tioners to ensure the practical application of AI in healthcare. Transparency and continuous improvement are the keys to unlocking the full potential of deep learning in medical image analysis. Please note that you should adapt this report to your specific research findings, dataset, and model architecture. Additionally, include experimental results, statistical analysis, and references as necessary to provide a comprehensive and scientifically sound research report.

3.6.2 Back-Propagation

Deep learning models have demonstrated remarkable potential in revolutionizing the field of medical image analysis. However, the inherent complexity of these models often renders them enigmatic, akin to impenetrable “black boxes”. This opacity presents a significant challenge in comprehending the decision-making processes that underlie their diagnostic capabilities. In this context, Backpropagation emerges as a fundamental and versatile technique. While traditionally applied for training neural networks, it also serves as a powerful instrument for enhancing the interpretability of these models. Backpropagation provides a unique vantage point to unravel the intricacies of deep learning’s inner workings. The Backpropagation process involves a series of interconnected steps. It commences with a forward pass, where input data traverses the intricate network of layers within the model. These layers apply a cascade of transformations to the input, ultimately leading to the model’s prediction. Central to Backpropagation is the computation of a loss function, which quantifies the disparity between the model’s prediction and the ground truth. The primary objective during training is to minimize this loss, a task facilitated by Backpropagation. The crux of Backpropagation, the backward pass, is where gradients are meticulously calculated. These gradients represent the sensitivity of the loss function to minute changes in each parameter and input variable. Armed with these gradients, optimization algorithms, such as gradient descent, iteratively adjust the model’s parameters—comprising weights and biases—toward minimizing the loss. While Backpropagation’s primary function remains model training, it uniquely positions itself to offer insights into the model’s decision-making process. By inspecting the gradients associated with input medical images, researchers and practitioners gain valuable insights into the model’s sensitivity to various regions within the image. Positive gradients pinpoint regions within the image that, when enhanced, would increase the loss, thereby underscoring the clinical significance of those features. Conversely, negative gradients signify areas where the model places less reliance, potentially revealing subtleties and complexities in the diagnostic considerations. Backpropagation emerges as an indispensable asset in the ongoing pursuit of transparent and interpretable deep-learning models for medical image analysis. It facilitates a deeper understanding of feature importance and model sensitivity, advancing the field and ultimately benefitting patients and healthcare professionals.

Chapter 4

Implementation

4.1 Work Plan

Figure 4.1 contains a flow-chart depicting our work-plan of the whole study.

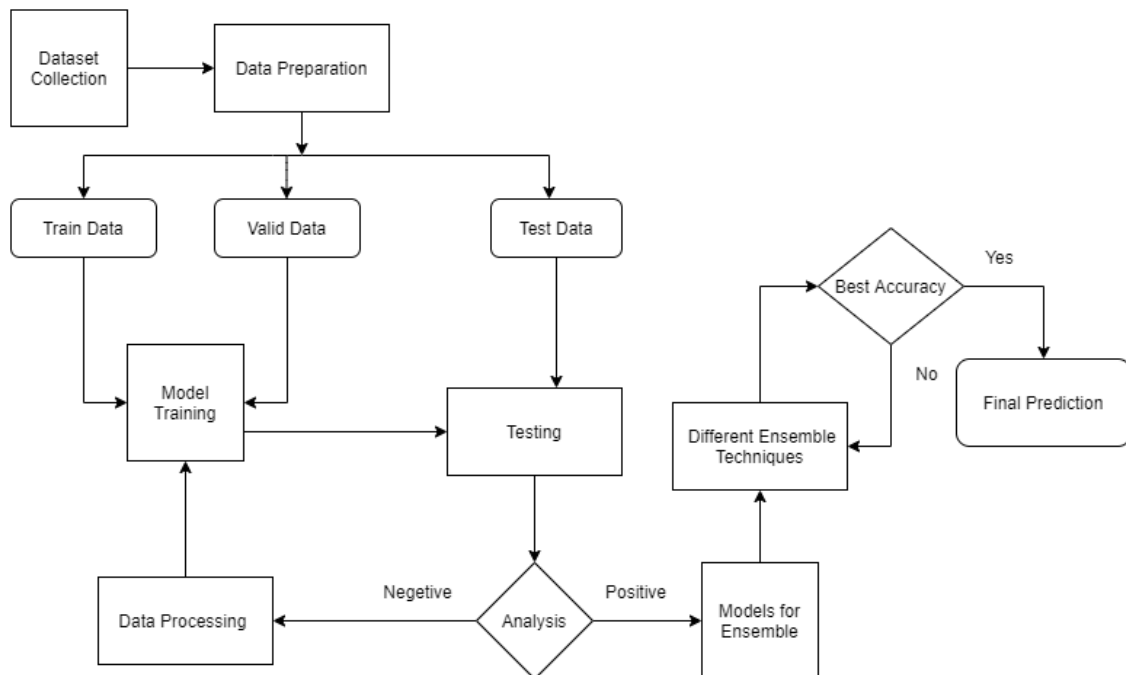


Figure 4.1: Workplan of the study.

4.2 Dataset Collection

For our study, we needed a vast amount of medical image for training the models. We collected three types of medical images: X-ray, MRI and CT images. All the data for this research can be found in table 4.1

Image Type	Dataset Name	Reference	Total Samples	Taken Samples
X-Ray	COVID-19 Radiography Database	[23], [24]	21,165	21,165
X-Ray	Knee Osteoarthritis Severity Grading Dataset	[25]	9,786	8,260
MRI	Brain Tumor Classification (MRI)	[26]	3,264	3,264
MRI	Br35H :: Brain Tumor Detection 2020	[27]	3,861	3,000
CT Image	CT KIDNEY DATASET: Normal-Cyst-Tumor and Stone	[18]	12,446	12,446
CT Image	The IQ-OTH/NCCD lung cancer dataset	[28]–[30]	1,190	1,097

Table 4.1: Overview of the datasets used for the study.

4.3 Dataset Preparation

After collecting all the datasets from publicly accessible online sources, the classes for each dataset was noted. Combining all of them, we found 18 classes including the medical images containing of both the patients and normal ones. We splitted all the images in a ratio of 70:30 for training and validation purposes respectively. Then, half of the validation samples were moved to a different directory storing them as unseen testing samples. In total, our dataset consisted of 18 classes, 34,178 training images. We named this dataset as *mir18*. Notably, *mir18* is highly imbalanced dataset. We created another dataset by applying a custom-built method to reduce the the imbalance among classes, which is discussed in the next section. This gave us another dataset containing a total of 74,366 images in 18 classes. We named this dataset as *mir18_v2*. This new dataset had less imbalance among the classes. Notably, among the 18 classes, 13 of them pertain to various diseases, while the remaining 5 classes describe the normal condition of the organs- Brain, Lungs, Kidney and Knee. Table 4.2 is added below for showing the number of training samples in each dataset.

4.4 Data Processing

In order to handle this highly imbalanced dataset (*mir18*), we have developed a new way of Data Processing combining Image Processing and Data Augmentation. In our train data, we have many classes under 1000 and some classes over 1000 even 7134. Methodically, this type of data cannot leverage the training of our models. Besides, the model will encounter overfitting as well as cannot classify those classes under 1000 with a satisfying accuracy. Due to this reason, we have applied a combination of different Image Processing Techniques and Normalization to the classes under 1000

Class Labels	<i>mir_18</i>			<i>mir18_v2</i>		
	Train	Valid	Test	Train	Valid	Test
-						
Benign Lung Cancer	84	18	18	3024	18	18
Brain Tumor	1050	225	225	4000	225	225
Covid	2531	542	543	4000	542	542
Kidney Cyst	2596	556	557	4000	556	557
Glioma Tumor	578	123	125	4000	123	125
Lung Opacity	4208	901	903	4208	901	903
Malignant Lung Cancer	392	84	85	4000	84	85
Meningioma Tumor	575	123	124	4000	123	124
No Brain Tumor	1326	284	285	4000	284	285
No Lung Cancer	291	62	63	4000	62	63
No Osteoarthritis	2277	487	489	4000	487	489
Normal Kidney	3553	761	763	4000	761	763
Normal Lung	7134	1528	1530	7134	1528	1530
Osteoarthritis	3503	749	755	4000	749	755
Pituitary Tumor	578	124	125	4000	124	125
Kidney Stone	963	206	208	4000	206	208
Kidney Tumor	1598	342	343	4000	342	343
Viral Pneumonia	941	201	203	4000	201	203

Table 4.2: Number of training samples in datasets *mir18* and *mir18_v2*.

samples to expand those classes to 1000 samples per class. In this method, for every original image we get 4 different images with different techniques applied including the original image. The motivation behind applying this method is basically Image Processing Techniques will create new samples which may derive new features that can help the model to learn new patterns. In the next stage, we applied a series of augmentation techniques on the train data to make them expand to a certain threshold (4,000). This method will introduce a large number of images that may assist the model in learning new patterns but may not be as useful as the first stage we applied. Figure 4.2 shows the workflow of the custom-built image processing function. We approached another method of preprocessing the images. In this method, we found out which images were specifically being predicted as wrong labels. Most of these images were X-ray images. A new augmentation approach suitable for X-ray was applied on these images. We kept this new dataset as *mir18_v3*. Although, we found very less accuracy of the models trained on this dataset in the prediction phase. Some outputs from the processing technique on a single raw X-ray image can be found in figure 4.3. Finally, we resized all the images to a dimension of (224,224,3) with a batch size of 32.

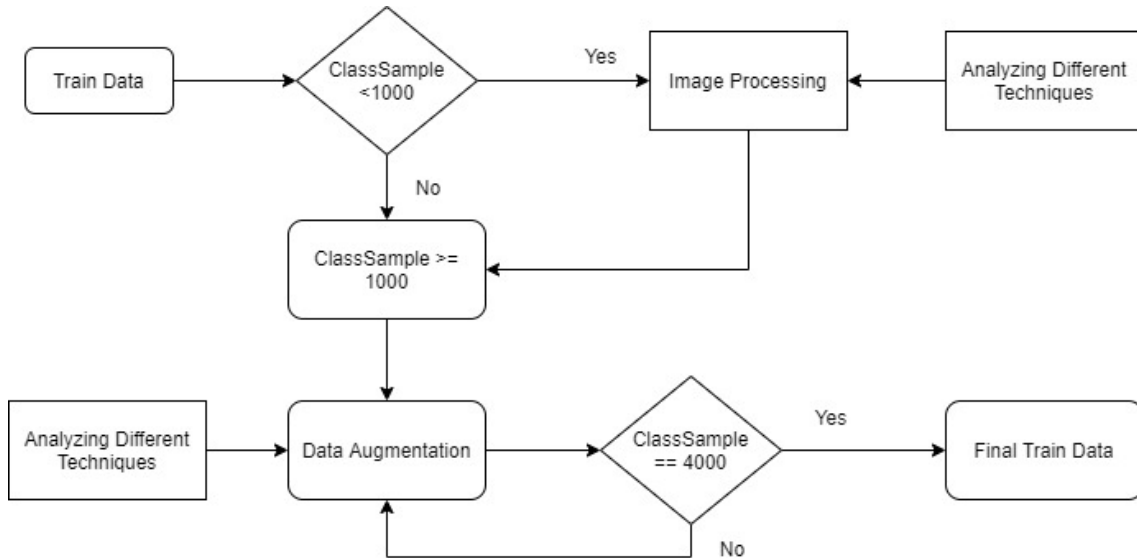


Figure 4.2: Flow-chart of the Data Processing Method.

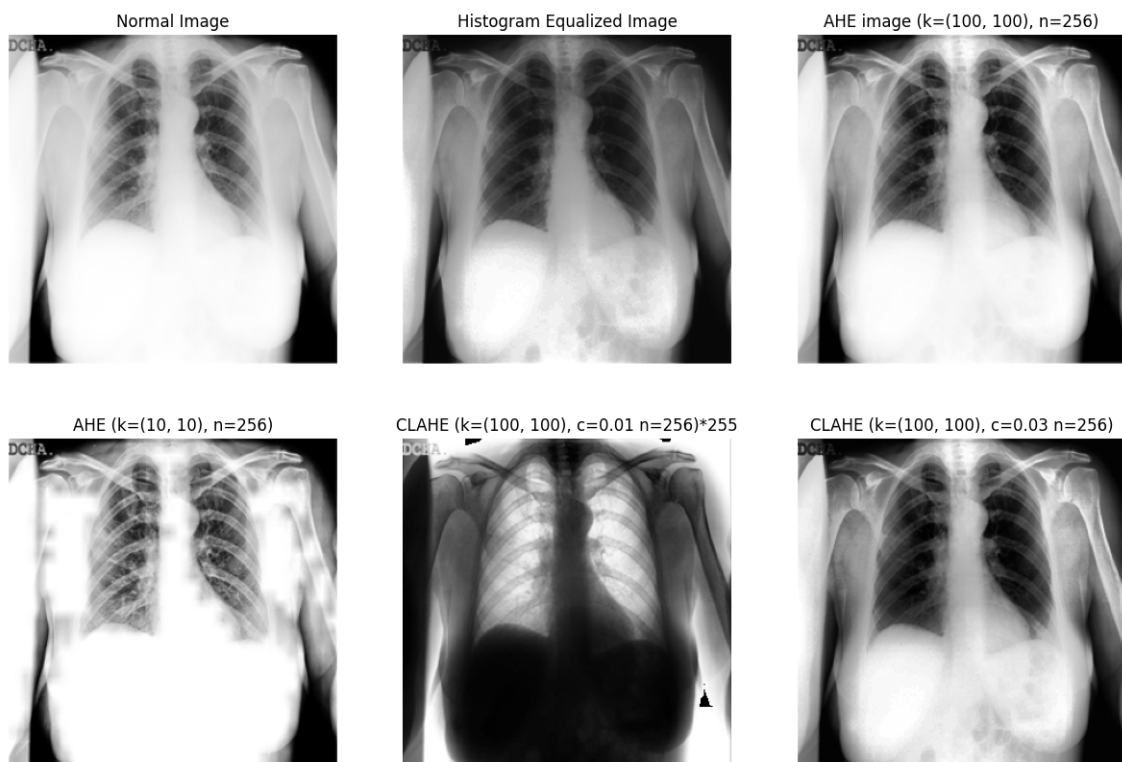


Figure 4.3: Visualization of processed images with different types of parameters.

4.5 Training Models

Our research is carried on a system with the following system configurations and software. Python 3.9.13 is used and implemented using Jupyter Notebook 6.4.12. and Tensorflow v2.10.0, respectively, on Intel(R) Core(TM) i5-8400 CPU @ 2.80GHz with 16 GB RAM. We used NVIDIA Gefore RTX 3060 GPU for an efficient training

purpose.

After a thorough research, we picked several pre-trained models and added 2 convolutional layers and a dropout of 0.5 onto them. For our study, we have used several pre-trained models such as: VGG16, VGG19, DenseNet121, ResNet50, EfficientNetV2S, InceptionResNetV2. We used all the models' stored weights from *imagenet* and trained them using both of our datasets. After checking the predictions, we finally picked two of them. These models are EfficientNetV2S and InceptionResNetV2. While using these models, we followed two approaches. First approach being the typical keeping the base layers of the models frozen and in the other approach we kept the layers trainable. In some cases, we found that keeping the layers trainable had a good impact on both the training accuracy and the cross-validations. Also, it helped in keeping the validation loss very low in numbers. Hence, we finally came up with two approaches of training using two CNN models on two different datasets. Table 4.3 portrays the overview of all our trials and the naming convention for better understanding of the Result Analysis section.

Model #	Pre-trained Model Name	Dataset used for training	Base Layer
Model 1	EfficientNetV2S	<i>mir18_v2</i>	Trainable
Model 2	EfficientNetV2S	<i>mir18_v3</i>	Frozen
Model 3	InceptionResNetV2	<i>mir18_v2</i>	Trainable
Model 4	InceptionResNetV2	<i>mir18_v3</i>	Trainable
Model 5	EfficientNetV2S	<i>mir18</i>	Trainable
Model 6	InceptionResNetV2	<i>mir18</i>	Trainable
Model 7	EfficientNetV2S	<i>mir18</i>	Frozen

Table 4.3: Overview of the trials of the study.

4.6 Proposed Prediction System

Deep learning neural networks are nonlinear methods. They offer increased flexibility and can scale in proportion to the amount of training data available. An oversight of this flexibility is that these models learn through a stochastic training algorithm which means that they are sensitive to specific training data and may find a different set of weights each time they are trained, which causes different predictions. Basically, this functionality of the neural networks is referred to as having a high variance and it can be frustrating to develop a final model to use for making predictions.

Spectating table 5.1, we can see the difference in prediction of multiple models. Some models are performing well in some classes while performing worst in other classes. For example, Model 7 has better accuracy in predicting classes like 10, 17 and the least predictions in some other classes. For further clarification, if we see Model 1 and Model 3, for which the same training samples were used, Model 1 has performed better than model 3 whereas Model 6 (Same architecture as model 3 but was trained on the raw images) performed better than Model 3 as well as giving compatible predictions with Model 1 for some specific classes. To overcome this issue,

we have introduced an ensemble learning method which is basically a combination of multiple models whereby applying different ensemble techniques like voting, bagging, boosting, stacking etc., the final prediction is generated. This final prediction not only reduces the variance but also can give far better predictions than any single model. In our case, we have combined five models which were trained differently. The reason behind choosing this model was mainly their class-based accuracy and overall accuracy. We have analyzed various scopes of our problem and spectating these models' results on the test dataset, we have gained some insight into their learning patterns which emphasizes combining them for ensemble prediction. We have used three different ensemble techniques. They are 'Majority Voting', 'Selective Class-wise Voting' and 'Customized Weighted Voting'. Figure 4.4 is a flow-chart which describes our proposed method of getting the ensembled learning.

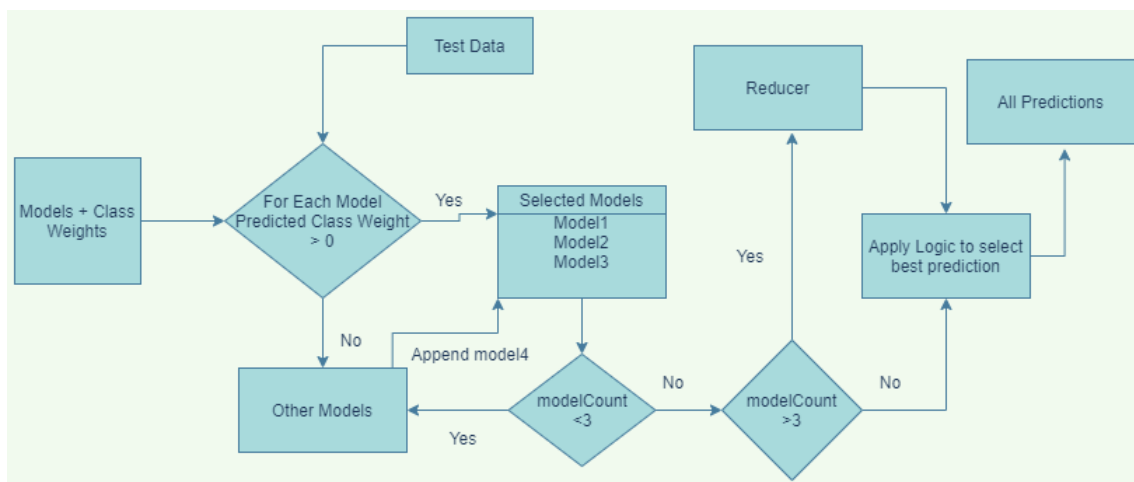


Figure 4.4: Flow-chart of the Proposed Ensemble Learning Method (Customized Weighted Voting).

4.6.1 Majority Voting

Firstly, we have tried majority voting which is a very general voting technique. In this method, we take the predictions from each model for a particular test data. After that, we check for the class which has the most votes by our CNN models to make the final prediction.

4.6.2 Selective Class-wise Voting

Furthermore, we have introduced a new type of voting where according to the actual label we select the best model for prediction. This is a completely customized approach to show the best prediction of the combined models for each label. Though this approach will not give accurate predictions for a random sample or a new test set. Because of this, we have tried a new way where neglecting the actual label for selecting the model, we have used the prediction of our best model (maximum average accuracy) to select the best model for a particular class to maximize accuracy which may work on a random sample or a new test set.

4.6.3 Customized Weighted Voting

To introduce a better approach, we have applied a customized where predefined weight gives priority to the best model for voting. After analyzing the results of these models, we specify a weighted array of values for each model containing values from 0 to 7. For every class, based on the model performance we have assigned a weight. For every model, there is a weighted array for 18 classes. The combination of the weight is most important here. Now, for each random sample, every model will give its prediction. Based on the weight assigned for that predicted class the model's prediction and weight will be stored in a dictionary with the model name as a key. After that, we removed the model's prediction from the dictionary which has a 0 weighted value. Now iterating over that dictionary, different logic is set to get a better prediction. Firstly, there can be a set of 3 models or less with the best weight for a particular class. Besides, before removing the models from the dictionary we stored the best-performing model's output in a separate dictionary. If after removing the model with 0 weight in the specific class, the dictionary becomes empty then we will take the best-performing model's prediction. We have also used a reducer function to keep the models in the dictionary limited to 3. This method is most suitable for a random sample or a group of samples without any type of label assigned to it.

Chapter 5

Result Analysis

The main purpose of our approaches is to gain better accuracy on all the 18 classes. In the beginning, we trained all the selected models on the *mir_18* dataset. This dataset was highly imbalanced in a sense that it was a mix of multiple datasets from multiple sources available online. Besides, some types of images were very rare to collect, for example: Benign Lung Cancer has only 120 samples on the whole dataset. In this case, we only used 84 samples for training, 18 for validation and 18 for testing purposes. After training on the base image we get 3 best model including EfficientNetV2S (keeping base layers frozen), EfficientNetV2S (keeping base layers trainable), InceptionResnetV2 (trainable) with an test accuracy of 88.78%, 92.68%, 93.77%. But the main drawback here is that, from the beginning of the training we've encountered overfitting . We have applied early stopping to lessen the overfitting and improve generalization of the model during the training process. However, due to the highly imbalanced class distribution of our dataset, the overfitting issue still remains. The training accuracies were 92.07%, 96.75%, 99.28% whereas the validation accuracies were 89.63%, 92.96%, 93.93% respectively. In order to overcome overfitting for a highly imbalanced dataset like ours, we have applied the data processing techniques in a different way.

5.1 Learning Curves for Five Approaches

A set of training and loss curves for our five approaches can be found in the figures 5.1 to 5.10. These curves provide valuable insights into the performance and convergence of the training process. The training curves show how various metrics evolve over time, while the loss curves indicate how the model's loss function changes during training. Please take a moment to review these curves, as they can provide a deeper understanding of the model's behavior and performance throughout the training process.

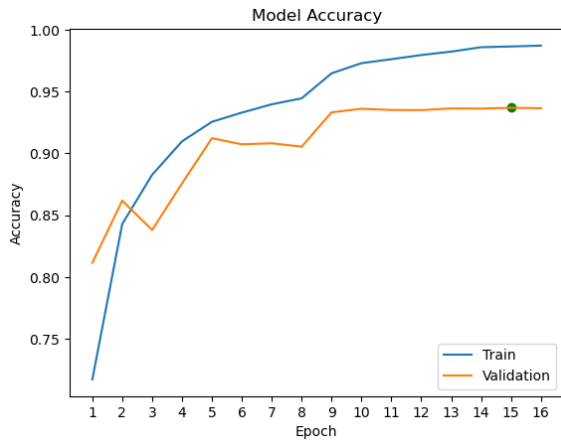


Figure 5.1: Accuracy curve for Model 1

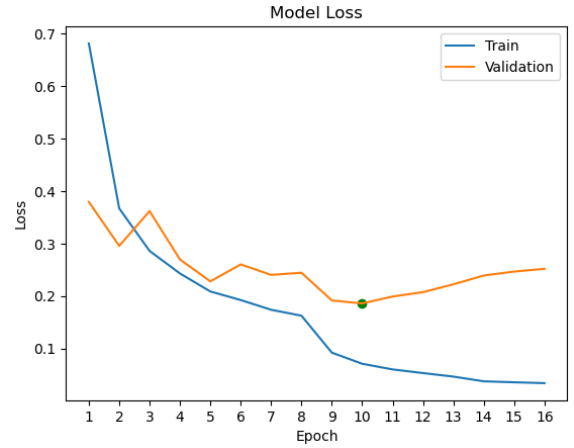


Figure 5.2: Loss curve for Model 1

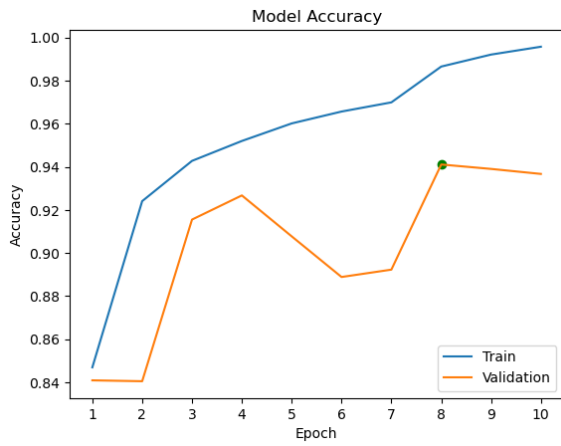


Figure 5.3: Accuracy curve for Model 3

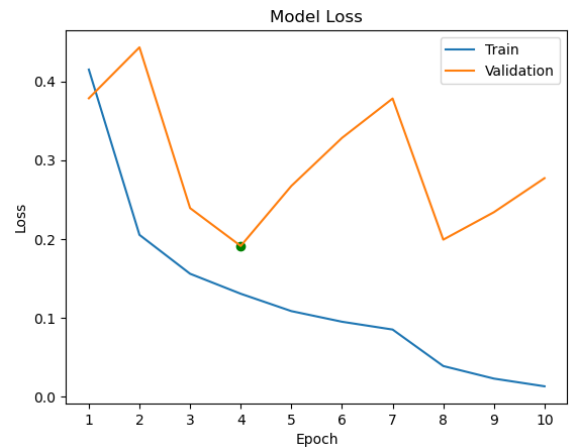


Figure 5.4: Loss curve for Model 3

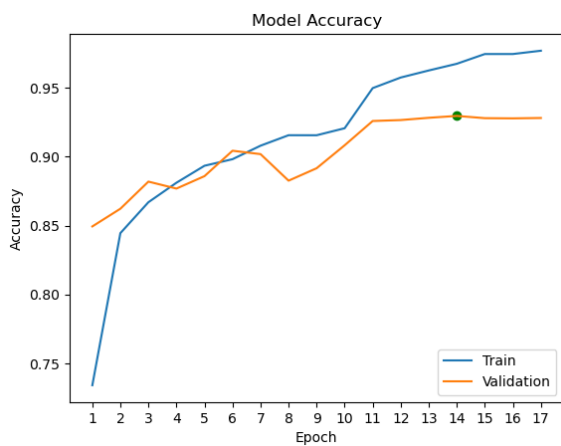


Figure 5.5: Accuracy curve for Model 5

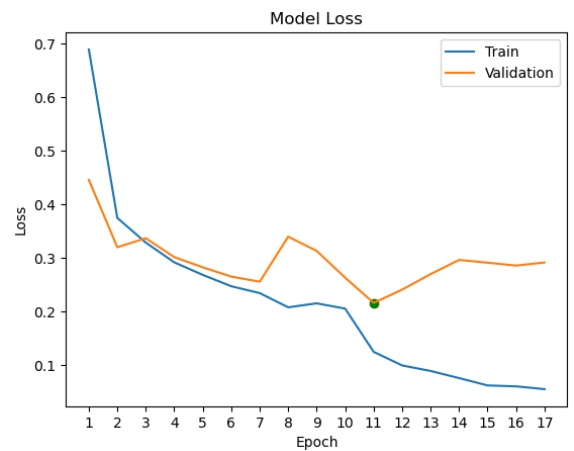


Figure 5.6: Loss curve for Model 5

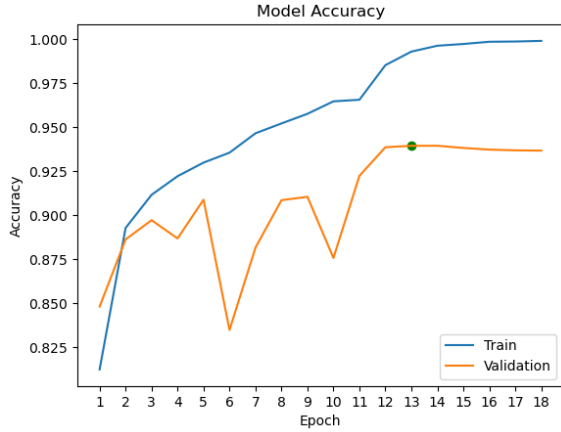


Figure 5.7: Accuracy curve for Model 6

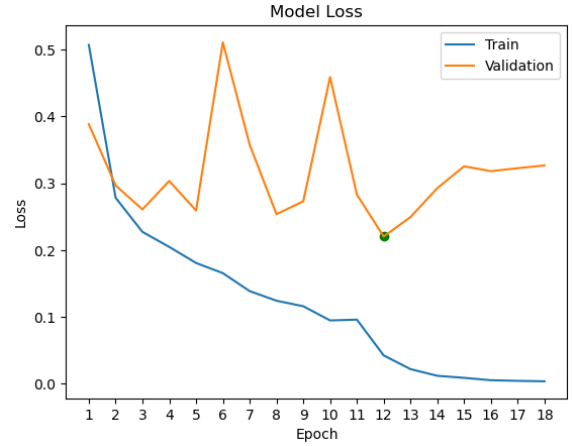


Figure 5.8: Loss curve for Model 6

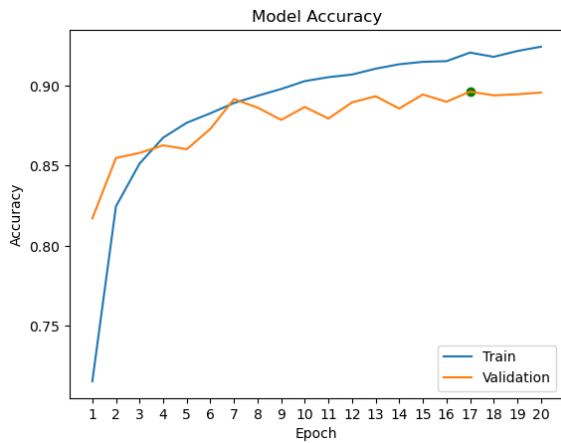


Figure 5.9: Accuracy curve for Model 7

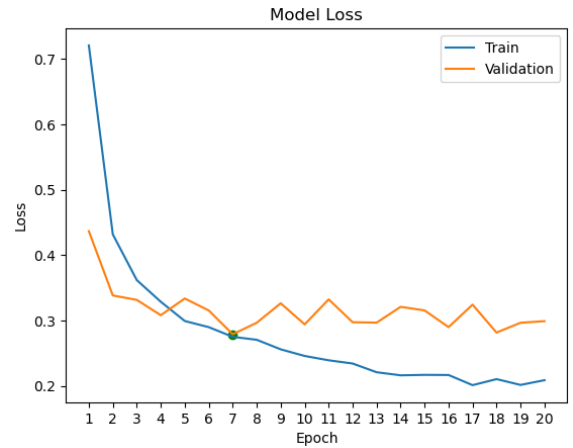


Figure 5.10: Loss curve for Model 7

5.2 Analysis of the Confusion Matrices

In our study, we have trained multiple models with *mir18*, *mir18_v2*, *mir18_v3* datasets to find an optimal solution to our problem. In order to analyze their result, we carefully examined the confusion matrices generated from these models to determine which ones were more effective at classifying maximum classes.

The confusion matrices of the model that we visualize in figures 5.11 to 5.15 are generated from the training on different datasets [explained in 4.3]. This matrix proves the effectiveness of our data processing techniques. For example, if we spectate the confusion matrix of model 1 and model 5, we can see the performance difference in the ‘Benign Lung Cancer’ class. By applying the proposed data processing, the same model has correctly classified 17 samples out of 18 whereas without data processing this model was unable to classify any sample. Besides, the overall performance in classifying almost every class has increased. Moreover, we tried to find out some models that have performed better in classifying ‘Osteoarthritis’ and ‘No

Osteoarthritis’ which is the major problem that we face till now. Due to the close relation between the images of these two classes, our models were unable to classify them correctly. Though they were classifying between the domains of Osteoarthritis they were giving false predictions. If one model predicts Osteoarthritis better, the same model is predicting ‘No Osteoarthritis’ worst. By analyzing these problems that we interpret from the confusion matrices, we tried to come up with some solutions to solve this issue.

For solving this problem, we then used a unique logical weighted method to prioritize a single approach for classifying each class using the ensembling techniques. This method involved calculating the percentages of correct predictions made by all the approaches in our study. The records are showed in Figure 5.1.

To narrow down our choices, we selected the top five approaches with the highest overall performances. These five approaches were given priority when building our prediction system. This step was crucial because comparing the confusion matrices for all the approaches played a significant role in our study. Figures 5.11 to 5.15 portrays the matrices for each approach.

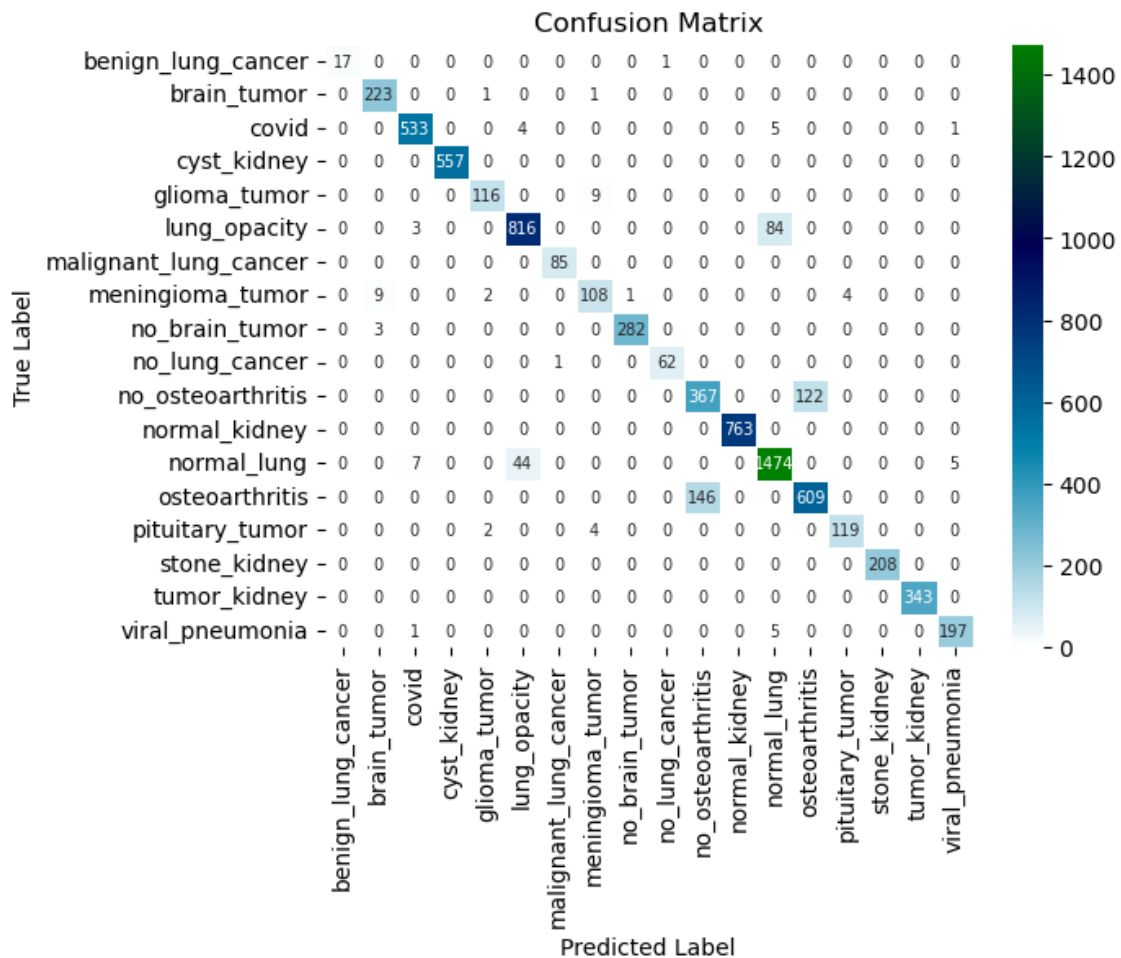


Figure 5.11: Confusion Matrix for Model 1

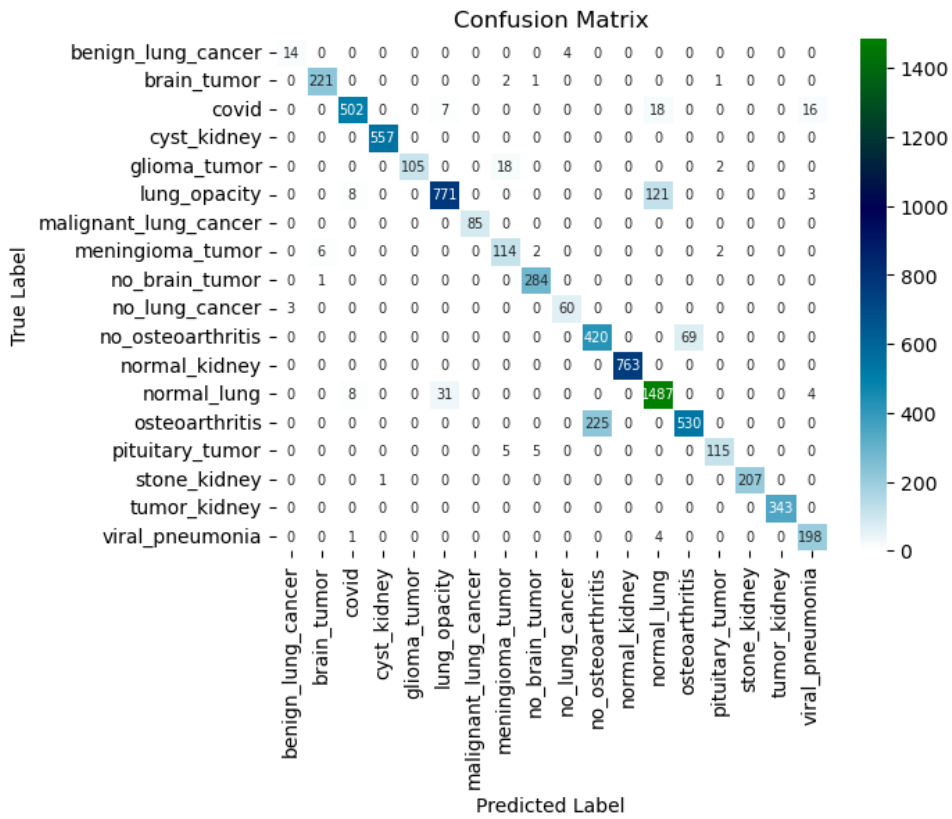


Figure 5.12: Confusion Matrix for Model 3

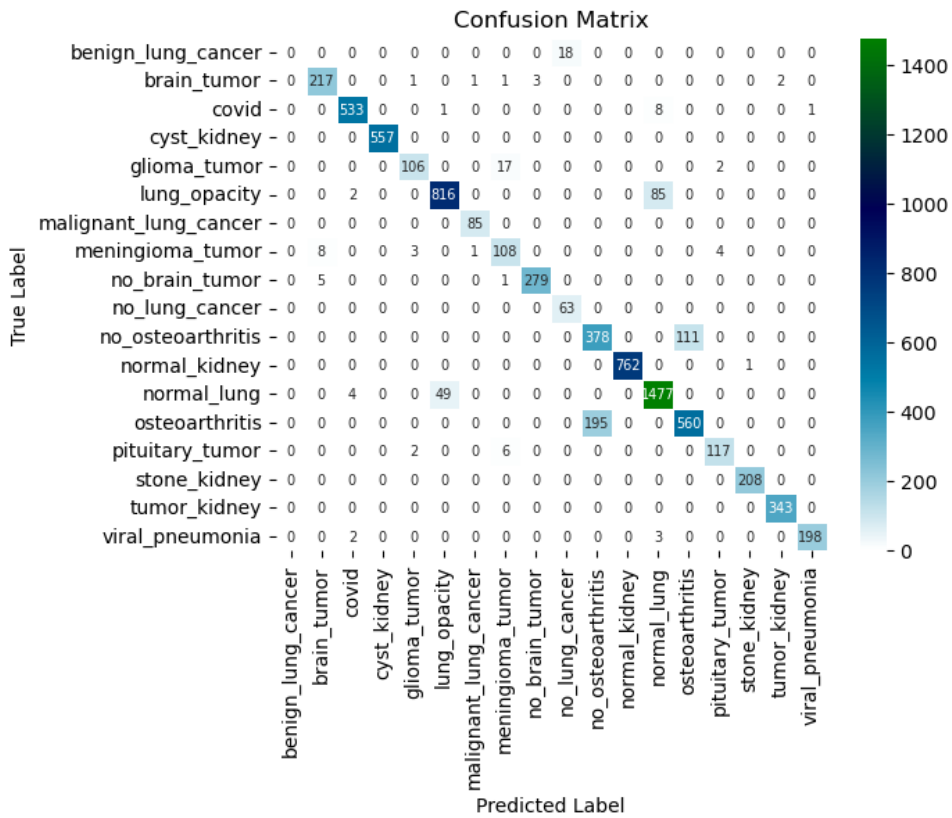


Figure 5.13: Confusion Matrix for Model 5

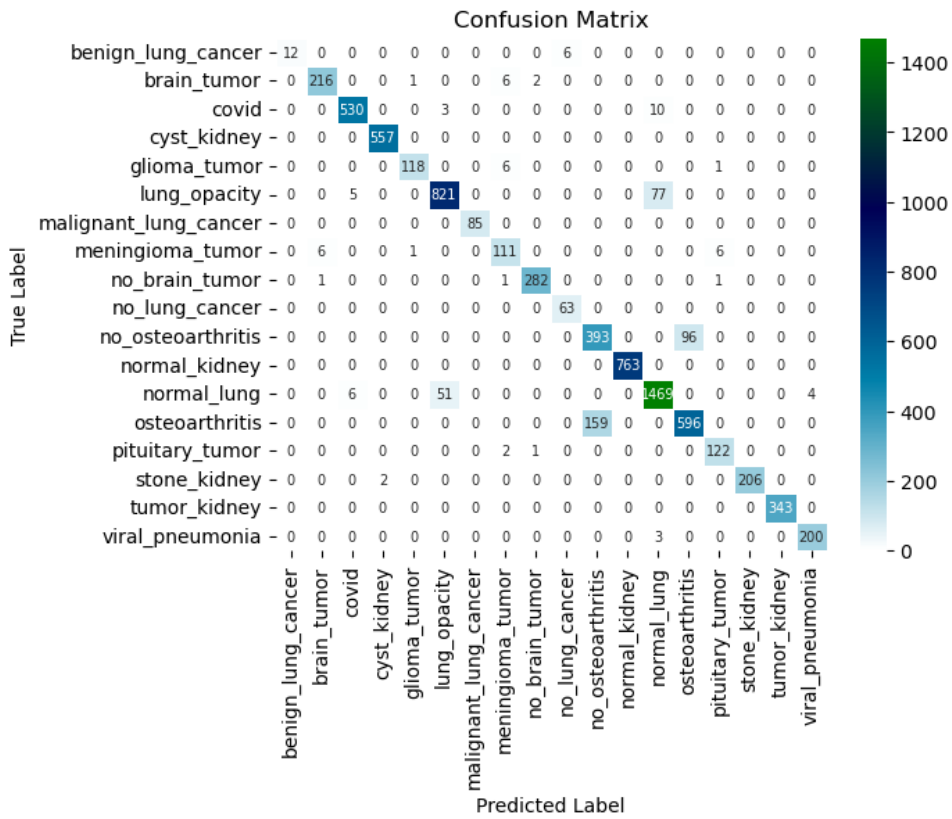


Figure 5.14: Confusion Matrix for Model 6

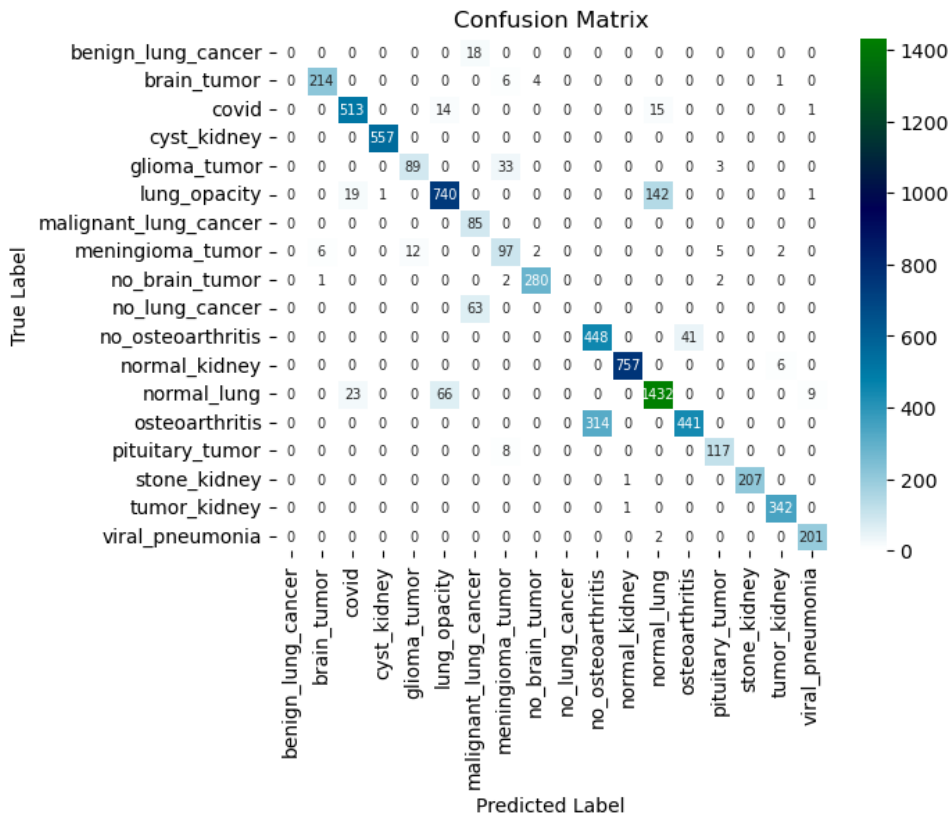


Figure 5.15: Confusion Matrix for Model 7

5.3 Ensembled Accuracy

In order to gain a more satisfying prediction, the ensemble techniques have worked to achieve better accuracy. In terms of majority voting, the ensemble models have an accuracy of 93.75. By applying “Sum Rule Ensemble”, the accuracy increases to 94.02%. In next we tried to predict using the “Mean Argmax” ensemble which gains an accuracy of 94.02% as well. After analyzing our used models performance in various sectors we have tried to implement a solution of our own. Firstly, we used a “Selective Class-wise Voting” technique where the first method (Actual label-wise selection of model) has an accuracy of 95.27% and the second method (Best prediction-wise selection of model) has an accuracy of 93.6%. To overcome this problem and handle random samples or a set of random samples our proposed method of ensemble “Customized Weighted Voting” is introduced which has gained a better accuracy of 94.07%. Though the accuracy difference is very less, combining the best weight for the models per class can increase this accuracy which can outperformed “Selective Class-wise Voting” technique (Actual label-wise selection of model) which gained an accuracy of 95.27%. Figure 5.16 and 5.17 shows the confusion matrices using both the approaches.

Label	Class Name	Model 1	Model 3	Model 5	Model 6	Model 7	MAX
0	Benign Lung Cancer	94.44	77.77	0	66.67	0	94.44
1	Brain Tumor	99.11	98.22	96.44	96	95.11	99.11
2	Covid	98.15	92.45	98.16	97.6	94.48	98.16
3	Kidney Cyst	100	100	100	100	100	100
4	Glioma Tumor	92.8	84	84.8	94.4	71.2	94.4
5	Lung Opacity	90.36	85.38	90.36	90.92	81.95	90.92
6	Malignant Lung Cancer	100	100	100	100	100	100
7	Meningioma Tumor	87.09	91.94	87.09	89.52	78.23	91.94
8	No Brain Tumor	98.94	99.65	97.89	98.95	98.23	99.65
9	No Lung Cancer	98.41	95.24	100	100	0	100
10	No Osteoarthritis	75.05	85.89	77.3	80.37	91.62	91.62
11	Normal Kidney	100	100	99.86	100	99.21	100
12	Normal Lung	96.33	97.19	96.53	96.01	93.59	97.19
13	Osteoarthritis	80.66	70.2	74.17	78.94	58.41	80.66
14	Pituitary Tumor	95.2	92.0	93.6	97.6	93.6	97.6
15	Kidney Stone	100	99.5	100	99.04	99.52	100
16	Kidney Tumor	100	100	100	100	99.71	100
17	Viral Pneumonia	97.04	97.54	97.54	98.52	99.01	99.01

Table 5.1: Record of correct predictions (%) per class for each model

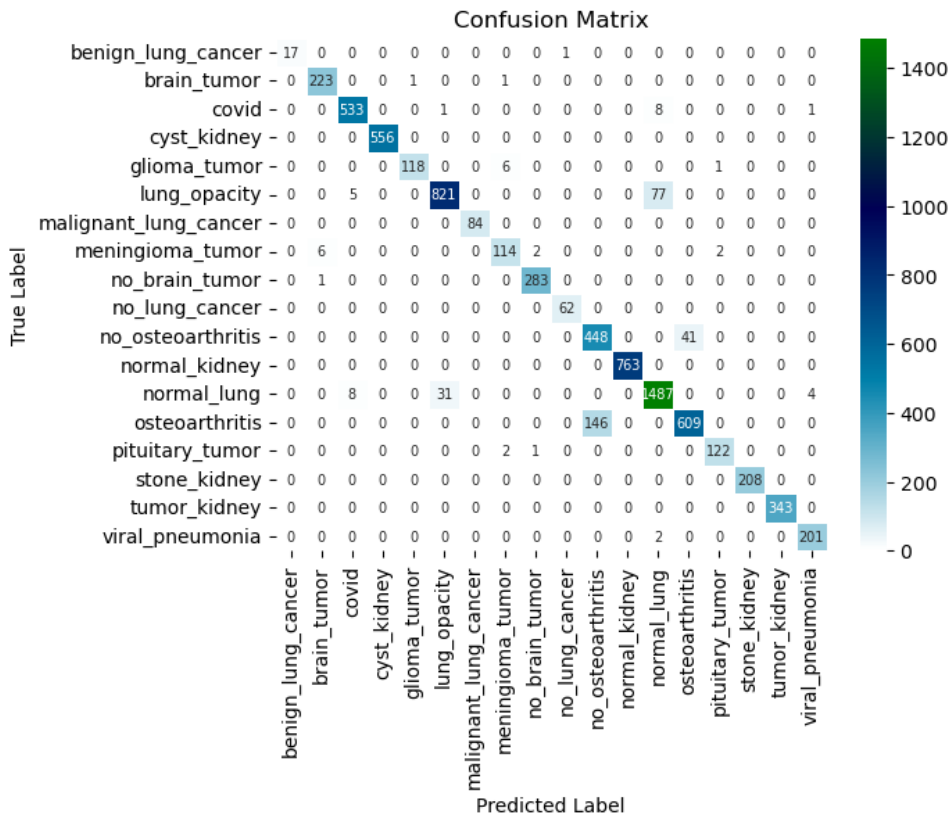


Figure 5.16: Confusion Matrix for Customized Weighted Voting

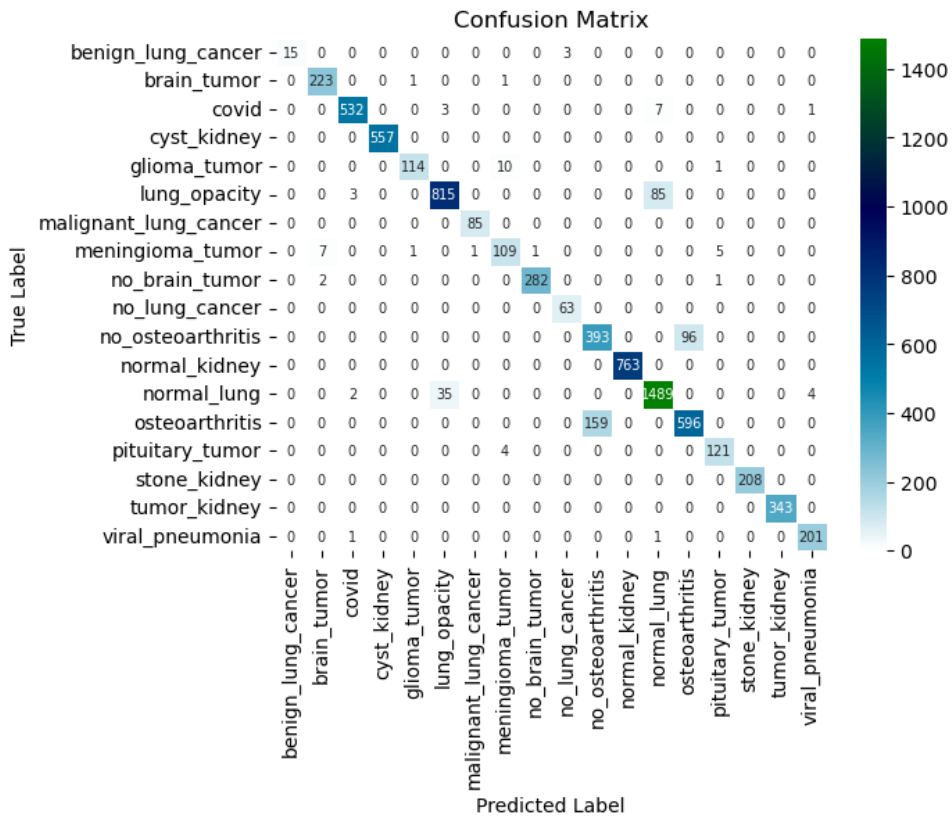


Figure 5.17: Confusion Matrix for Selective Class-wise Voting Technique

5.3.1 Validation of the Proposed Data Processing Method

Application of our custom data processing techniques has improved the training from the beginning. The training accuracy and validation accuracy have had less difference from the beginning because of our approach. After applying these methods, the validation as well as testing accuracy increased for most of the models. EfficientNetV2S (frozen), EfficientNetV2S (trainable), InceptionResnetV2 (trainable) have gained a test accuracy of 91.43%, 93.66%, 92.26% and validation accuracy of 91.58%, 93.69%, 94.11% respectively. To clarify, table 5.2 indicates all the models accuracy in percentages for each class before and after data processing. One of the most important results which we achieved with this method is the results of Benign Lung Cancer. Analyzing the confusion matrices of the best 3 models, in the *mir18* dataset, EfficientNetV2S (frozen), EfficientNetV2S (trainable), InceptionResnetV2 (trainable) have classified only 0, 0, 12 samples respectively among 18 whereas after applying our data processing approach, the accuracy has increased to 15, 17, 14 respectively. Besides, we tested DenseNet121 and VGG19 which have classified 7 and 4 samples in *mir18* dataset but both of them classified 14 samples after applying the technique. This analogy indicates that the high efficiency of our data processing method for a dataset with very less data. Besides, we can also interpret from the analysis that, models following EfficientNet architecture cannot learn and extract features from less data whereas models built following Inception-Resnet architecture can give better performance from less data. Moreover, applying this method if training samples are increased, EfficientNet architected models can learn patterns from training data as well as extract positive features more accurately than models following Inception-Resnet architecture.

Model	Before Processing (%)			After Processing (%)		
	Train	Valid	Test	Train	Valid	Test
-						
DenseNet121	90.9	87.67	85.79	94.14	88.05	87.37
EfficientNetV2S	92.07	89.63	88.78	96.3	91.58	91.43
EfficientNetV2S*	96.75	92.96	92.68	98.64	93.69	93.66
InceptionResNetV2*	99.28	93.93	93.77	98.66	94.11	92.26
VGG19	92.77	87.03	86.41	96.71	89.65	88.97

Table 5.2: Records of the performance (All accuracy) by the models were stored in a spreadsheet for analysis. A comparison chart is made for understanding the performances before and after applying our data processing techniques. * indicates the base layer of the model is kept trainable.

5.3.2 Validation of Image Processing

After training and testing, we have specified the wrong predicted images for all the models and applied intersection on all models wrong predicted samples and store them in a different dataset. This dataset indicates the images which no models could classify correctly. After that, we applied histogram equalization, AHE, CLAHE on all the samples and made three different datasets for three processing techniques.

Now, we tested the three models on this dataset and for EfficientNetV2S (trainable) which gained a better accuracy than others, scored an accuracy of 21.52%, 28.14% and 36.1% respectively on the Histogram Equalized images, AHE images and CLAHE images. This research indicates how efficiently the image processing techniques have worked to give positive features to the model which helped to classify the images more correctly in the test set.

5.4 Grad-CAM Visualization

According to Lloyd-Jones (2020), osteoarthritis has the potential to impact any synovial joint within the human body. The joints most frequently affected by this condition include the hands, wrists, hips, knees, and feet. One of the hallmark manifestations of osteoarthritis is the distinctive X-ray findings it produces. These radiographic features encompass a narrowing of the joint space, the development of osteophytes (commonly referred to as bone spurs), cortical irregularities, and/or sclerosis of the articular surface, as well as the formation of sub-cortical cysts, also known as geodes. It is noteworthy that while these features can manifest individually, it is quite common for two or more of these signs to be concurrently present in cases of osteoarthritis. Figures 5.18 and 5.19 can give a brief information about how Radiologists find out the possibility of Osteoarthritis in knee from an X-ray; notably a Consultant Radiologist has described this in a website [31]. We tested our models and got a very convincing accuracy in classifying ‘Osteoarthritis’. Although, we noted that our models were predicting incorrectly when classifying a X-ray image between ‘Osteoarthritis’ or ‘No osteoarthritis’. Using different techniques of Grad-CAM, we found our model learnt the patterns in the images very well. Figures 5.20 and 5.21 are the visualization for 2 X-ray images of knees from our testing dataset. These images were completely unknown by our model. Firstly, the model predicted correctly on the both images and the Grad-CAM visualizations show that the model took the data mainly on the appropriate portion of the X-ray images. Hence, the model found the pattern to classify the images correctly as ‘Osteoarthritis’. Figures 5.22 and 5.23 contain two other Grad-CAM visualizations for correctly classifying ‘Covid’ and ‘Brain Tumor’.

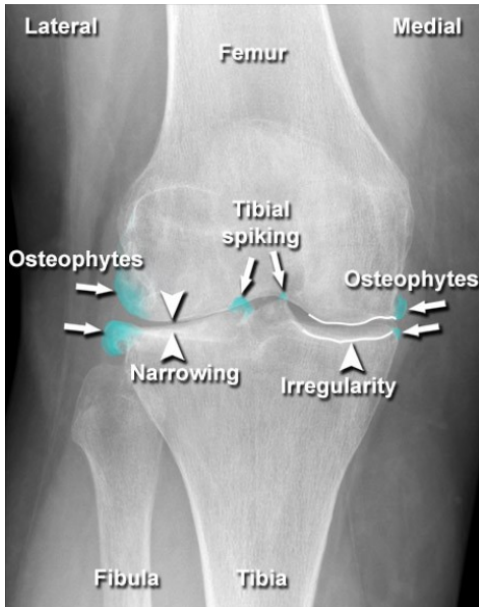


Figure 5.18: A knee X-ray with identifiers of Osteoarthritis of knee

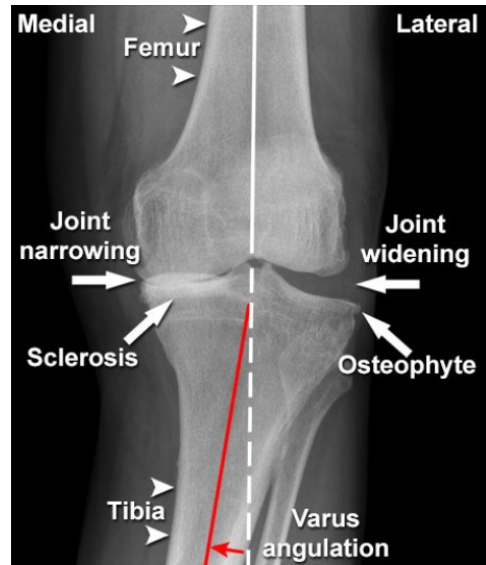


Figure 5.19: A knee X-ray with another set of identifiers to detect Osteoarthritis of knee.

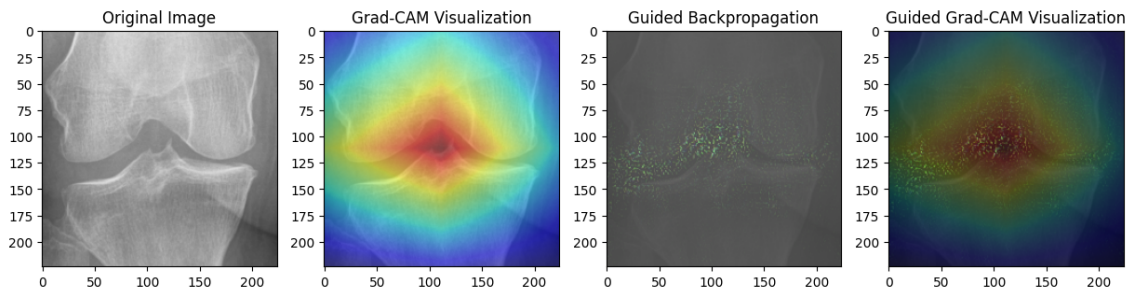


Figure 5.20: Grad-CAM visualizations for another Osteoarthritis detection.

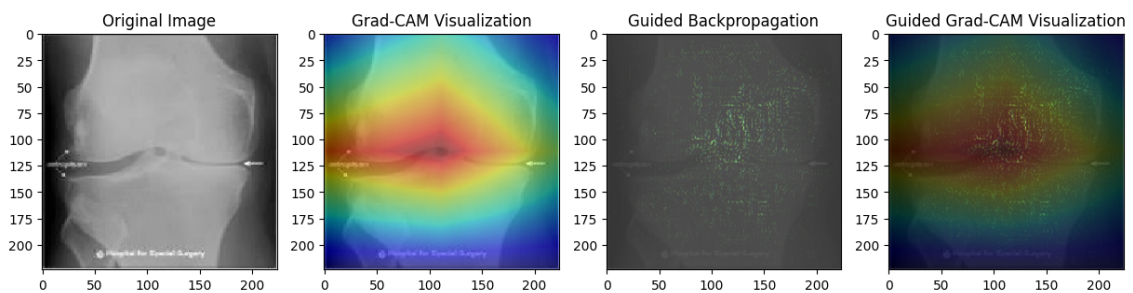


Figure 5.21: Grad-CAM visualizations of an X-ray image that is also detected 'Osteoarthritis' by the proposed model.

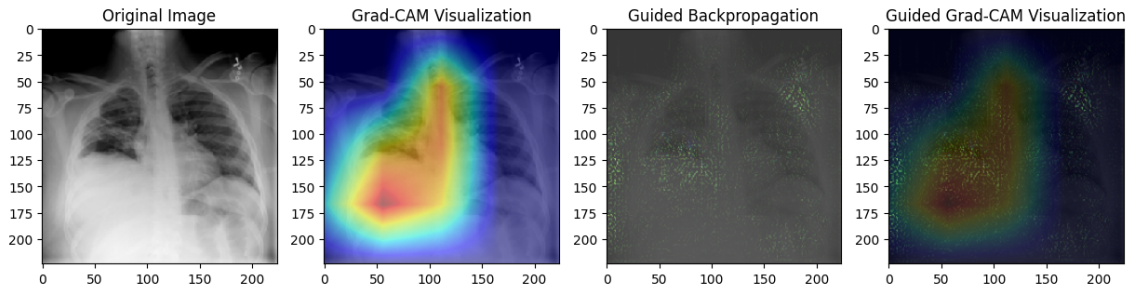


Figure 5.22: Grad-CAM visualizations of a chest X-ray image that is detected ‘Covid’ by the model.

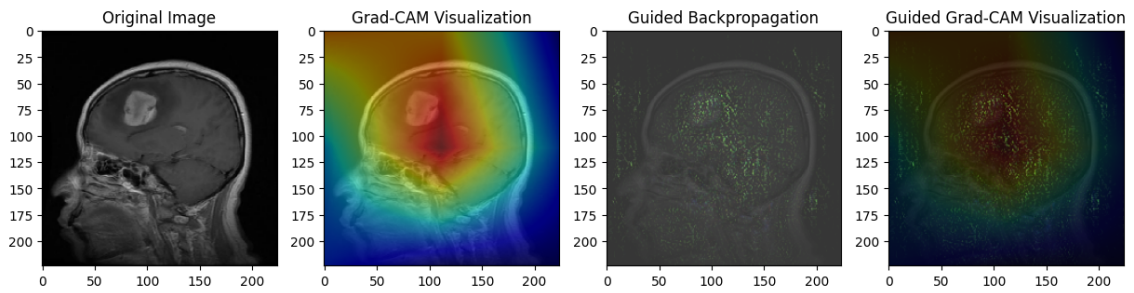


Figure 5.23: Grad-CAM visualizations of a brain MRI that is detected ‘Brain Tumor’ by the model.

Chapter 6

Conclusion

In this study, we have tackled a crucial issue within the realm of medical diagnosis, namely the substantial risk of misdiagnosis and the limited mechanisms for cross-validation within radiology. Our principal goal has been to make a valuable contribution to the healthcare sector by constructing a robust system capable of accurately detecting diseases through the analysis of medical images. The overarching aim was to bridge the gap between radiologists' decisions and the subsequent treatment plans proposed by doctors, ultimately leading to improved patient care. Our research journey encompassed a comprehensive array of methods, including data processing, deep learning models, transfer learning, ensemble techniques, and explainable AI. We methodically curated and processed datasets comprising medical images, enhancing their quality and expanding the pool of training samples. We harnessed the power of convolutional neural networks (CNNs) and transfer learning to develop models proficient at analyzing a wide spectrum of medical images and making precise predictions. The significance of ensemble learning in our research cannot be overstated. By amalgamating multiple models and introducing inventive ensemble techniques such as Selective Class-wise Voting and Customized Weighted Voting, we managed to construct a prediction system that surpassed the performance of individual models. This system not only mitigated variance but also substantially elevated the accuracy of disease classification across 18 distinct classes.

In summation, our research represents a substantial stride toward improving disease diagnosis in the medical domain. The fusion of advanced data processing, cutting-edge deep learning models, and innovative ensemble techniques lays the groundwork for more precise and efficient disease detection systems. As we continue to address challenges and explore future avenues, we envisage a healthcare sector wherein AI plays a pivotal role in augmenting patient care and saving lives.

6.1 Challenges

In our research, we have tried to introduce a process to classify multiple diseases with different types of medical images. We have tried to come up with a unique data processing method which might enhance the accuracy of a dataset with limited samples. Besides, we have introduced an ensemble technique which can perform better than many traditional ensemble techniques depending on the dataset and the

performance of the ensemble models. While we have achieved notable milestones in our research, we have also encountered some significant challenges. Firstly, one of the foremost difficulties we faced was collecting a sufficient amount of data per class. Data scarcity in some classes created a highly imbalance in the dataset. Due to this, we faced rigorous difficulties in handling these highly imbalanced medical image dataset. Some classes were severely challenging for our models to learn effectively. Another persistent concern was overfitting, which persisted despite the implementation of techniques like early stopping basically because of this highly imbalanced raw data. The skewed distribution of data played a role in exacerbating this problem. Moreover, we were unable to classify Osteoarthritis as precisely as other diseases because some ‘Osteoarthritis’ and ‘No Osteoarthritis’ images were very close and very difficult to distinguish which created a dilemma for the models. Furthermore, in the evolving landscape of AI-driven medical diagnosis, the need for model interpretability and transparency has become increasingly crucial. Future research should delve deeper into Explainable AI to foster trust and facilitate better understanding between AI systems and healthcare professionals. This will be pivotal in ensuring the successful integration of AI in healthcare practice.

6.2 Future Scopes

In order to extract a more sustainable and effective system that can help to reduce misdiagnosis, we hope to improve the methodologies we used. Firstly, in the future there would be more relevant data which will reduce the data scarcity problem that we faced and might move overfitting by reducing the imbalance nature of our dataset. Besides, we would like to classify more diseases with a proposed approach which will make our system more precise and accurate in multiple disease diagnosis. Moreover, classes like ‘Osteoarthritis and ‘No Osteoarthritis’ which were very difficult to precisely classify can be tackled by introducing some other techniques for processing the images or by tuning the hyperparameters. By improving and restructuring some methods, we hope to introduce a far better system which will help in medical science in future.

The research we’ve conducted paves the way for exciting future avenues of exploration. One such avenue is the refinement of advanced data augmentation techniques tailored specifically for medical images. This could enrich our training data, making it more diverse and informative. Staying abreast of the latest developments in transfer learning and incorporating newer pre-trained models could elevate disease detection to even greater heights in terms of performance. Additionally, the integration of information from multiple sources, such as combining X-rays with clinical data or other medical tests, offers the promise of a more comprehensive and holistic approach to diagnosis. The development of real-time prediction systems that can aid healthcare professionals during patient examinations and decision-making processes represents a significant leap forward in medical technology. Lastly, conducting rigorous clinical trials to assess the real-world impact of AI-assisted diagnosis in healthcare settings is imperative. This will ensure that such technology is not only effective but also safe and reliable in practical healthcare scenarios.

Bibliography

- [1] D. Siegal, L. M. Stratchko, and C. DeRoo, *Diagnosis*, vol. 4, no. 3, pp. 125–131, 2017. DOI: doi:10.1515/dx-2017-0025. [Online]. Available: <https://doi.org/10.1515/dx-2017-0025>.
- [2] A. Barragán-Montero, U. Javaid, G. Valdés, D. Nguyen, P. Desbordes, B. Macq, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman, S. Michiels, K. Souris, E. Sterpin, and J. A. Lee, “Artificial intelligence and machine learning for medical imaging: A technology review,” *Physica Medica*, vol. 83, pp. 242–256, 2021, ISSN: 1120-1797. DOI: <https://doi.org/10.1016/j.ejmp.2021.04.016>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1120179721001733>.
- [3] H. Mary Shyni and E. Chitra, “A comparative study of x-ray and ct images in covid-19 detection using image processing and deep learning techniques,” *Computer Methods and Programs in Biomedicine Update*, vol. 2, p. 100 054, 2022, ISSN: 2666-9900. DOI: <https://doi.org/10.1016/j.cmpbup.2022.100054>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666990022000064>.
- [4] F. Hussein, A. Mughaid, S. AlZu’bi, S. M. El-Salhi, B. Abuhaija, L. Abualigah, and A. H. Gandomi, “Hybrid clahe-cnn deep neural networks for classifying lung diseases from x-ray acquisitions,” *Electronics*, vol. 11, no. 19, p. 3075, 2022. DOI: 10.3390/electronics11193075.
- [5] M. M. Taresh, N. Zhu, T. A. Ali, A. S. Hameed, and M. L. Mutar, “Transfer learning to detect covid-19 automatically from x-ray images using convolutional neural networks,” *International Journal of Biomedical Imaging*, vol. 2021, pp. 1–9, 2021. DOI: 10.1155/2021/8828404.
- [6] F. Yimer, A. Tessema, and G. Simegn, “Multiple lung diseases classification from chest x-ray images using deep learning approach,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, pp. 2936–2946, Oct. 2021. DOI: 10.30534/ijatcse/2021/021052021.
- [7] A. I. Khan, J. L. Shah, and M. M. Bhat, “Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images,” *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105 581, 2020. DOI: 10.1016/j.cmpb.2020.105581.
- [8] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Y. Ding, A. Bagul, C. P. Langlotz, K. S. Shpanskaya, M. P. Lungren, and A. Y. Ng, “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning,” *CoRR*, vol. abs/1711.05225, 2017. arXiv: 1711.05225. [Online]. Available: <http://arxiv.org/abs/1711.05225>.

- [9] N. Kumar, N. Narayan Das, D. Gupta, K. Gupta, and J. Bindra, “Efficient automated disease diagnosis using machine learning models,” *Journal of Healthcare Engineering*, vol. 2021, pp. 1–13, 2021. DOI: 10.1155/2021/9983652.
- [10] S. Saeedi, S. Rezayi, H. Keshavarz, and S. R. Niakan Kalhori, “Mri-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques,” *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, 2023. DOI: 10.1186/s12911-023-02114-6.
- [11] H. Mohsen, E.-S. A. El-Dahshan, E.-S. M. El-Horbaty, and A.-B. M. Salem, “Classification using deep learning neural networks for brain tumors,” *Future Computing and Informatics Journal*, vol. 3, no. 1, pp. 68–71, 2018, ISSN: 2314-7288. DOI: <https://doi.org/10.1016/j.fcij.2017.12.001>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2314728817300636>.
- [12] L. Gaur, M. Bhandari, T. Razdan, S. Mallik, and Z. Zhao, “Explanation-driven deep learning model for prediction of brain tumour status using mri image data,” *Frontiers in Genetics*, vol. 13, 2022. DOI: 10.3389/fgene.2022.822666.
- [13] S. Alsubai, H. U. Khan, A. Alqahtani, M. Sha, S. Abbas, and U. G. Mohammad, “Ensemble deep learning for brain tumor detection,” *Frontiers in Computational Neuroscience*, vol. 16, 2022. DOI: 10.3389/fncom.2022.1005617.
- [14] M. S. Khan, A. Rahman, T. Debnath, M. R. Karim, M. K. Nasir, S. S. Band, A. Mosavi, and I. Dehzangi, “Accurate brain tumor detection using deep convolutional neural network,” *Computational and Structural Biotechnology Journal*, vol. 20, pp. 4733–4745, 2022. DOI: 10.1016/j.csbj.2022.08.039.
- [15] N. Noreen, S. Palaniappan, A. Qayyum, I. Ahmad, M. Imran, and M. Shoaib, “A deep learning model based on concatenation approach for the diagnosis of brain tumor,” *IEEE Access*, vol. 8, pp. 55 135–55 144, 2020. DOI: 10.1109/access.2020.2978629.
- [16] K. Yildirim, P. G. Bozdog, M. Talo, O. Yildirim, M. Karabatak, and U. Acharya, “Deep learning model for automated kidney stone detection using coronal ct images,” *Computers in Biology and Medicine*, vol. 135, p. 104 569, 2021. DOI: 10.1016/j.combiomed.2021.104569.
- [17] P. M. Shakeel, M. Burhanuddin, and M. I. Desa, “Lung cancer detection from ct image using improved profuse clustering and deep learning instantaneously trained neural networks,” *Measurement*, vol. 145, pp. 702–712, 2019. DOI: 10.1016/j.measurement.2019.05.027.
- [18] M. N. Islam, M. Hasan, M. K. Hossain, M. G. Alam, M. Z. Uddin, and A. Soyly, “Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography,” *Scientific Reports*, vol. 12, no. 1, 2022. DOI: 10.1038/s41598-022-15634-4.
- [19] M. Gharaibeh, D. Alzu’bi, M. Abdullah, I. Hmeidi, M. R. Al Nasar, L. Abualigah, and A. H. Gandomi, “Radiology imaging scans for early diagnosis of kidney tumors: A review of data analytics-based machine learning and deep learning approaches,” *Big Data and Cognitive Computing*, vol. 6, no. 1, p. 29, 2022. DOI: 10.3390/bdcc6010029.
- [20] S. Sudhakar, *Histogram equalization*, Jan. 2021. [Online]. Available: <https://towardsdatascience.com/histogram-equalization-5d1013626e64>.

- [21] M. Gomroki, M. Hasanlou, and P. Reinartz, “STCD-EffV2t unet: Semi transfer learning EfficientNetV2 t-unet network for urban/land cover change detection using sentinel-2 satellite images,” *Remote Sensing*, vol. 15, no. 5, p. 1232, Feb. 2023. DOI: 10.3390/rs15051232. [Online]. Available: <https://doi.org/10.3390/rs15051232>.
- [22] Z. Elhamraoui, *InceptionResNetV2 Simple Introduction — zahraelhamraoui1997*, <https://medium.com/@zahraelhamraoui1997/inceptionresnetv2-simple-introduction-9a2000edcdb6>, [Accessed 30-07-2023].
- [23] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. Abul Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughair, M. S. Khan, and et al., “Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images,” *Computers in Biology and Medicine*, vol. 132, p. 104319, 2021. DOI: 10.1016/j.compbiomed.2021.104319.
- [24] M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, and et al., “Can ai help in screening viral and covid-19 pneumonia?” *IEEE Access*, vol. 8, pp. 132665–132676, 2020. DOI: 10.1109/access.2020.3010287.
- [25] P. Chen, *Knee osteoarthritis severity grading dataset*, 2018. DOI: 10.17632/56RMX5BJCR.1. [Online]. Available: <https://data.mendeley.com/datasets/56rmx5bjcr/1>.
- [26] S. Bhuvaji, A. Kadam, P. Bhumkar, S. Dedge, and S. Kanchan, *Brain tumor classification (mri)*, 2020. DOI: 10.34740/KAGGLE/DSV/1183165. [Online]. Available: <https://www.kaggle.com/dsv/1183165>.
- [27] *Br35H :: Brain Tumor Detection 2020 — kaggle.com*, <https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection?select=Br35H-Mask-RCNN>, [Accessed 30-07-2023].
- [28] H. F. Al-Yasriy, *The iq-oth/nccd lung cancer dataset*, 2020. DOI: 10.34740/KAGGLE/DS/672399. [Online]. Available: <https://www.kaggle.com/ds/672399>.
- [29] H. F. Al-Yasriy, M. S. AL-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, “Diagnosis of lung cancer based on CT scans using CNN,” *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 2, p. 022035, Nov. 2020. DOI: 10.1088/1757-899x/928/2/022035. [Online]. Available: <https://doi.org/10.1088/1757-899x/928/2/022035>.
- [30] H. F. Kareem, M. S. AL-Huseiny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, “Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 3, p. 1731, Mar. 2021. DOI: 10.11591/ijeecs.v21.i3.pp1731-1738. [Online]. Available: <https://doi.org/10.11591/ijeecs.v21.i3.pp1731-1738>.
- [31] D. G. Lloyd-Jones, *Imaging of musculoskeletal disorders*, D. R. Smith, Ed., Jan. 2020. [Online]. Available: <https://www.radiologymasterclass.co.uk/tutorials/musculoskeletal/imaging-joints-bones/osteoarthritis>.