Multimodal Deep Learning Approach for Skin Cancer Detection using Image and Patient Metadata

MSc Research Project
Data Analytics

Sriram Iyer
Student ID: 23232064

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: …….Sriram Krishnan Iyer……

Student ID: ……23232064…….

Programme: …MSc in Data Analytics…… **Year:** …2024-25…..

Module: …Research Project…….

Supervisor: ……Jorge Basilio…….

Submission Due Date: ……12 December 2024…….

Project Title: …… Multimodal Deep Learning Approach for Skin Cancer Detection using Image and Patient Metadata…….

Word Count: ……8181…… **Page Count** …20……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: …Sriram Krishnan Iyer……

Date: …12 December 2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Multimodal Deep Learning Approach for Skin Cancer Detection using Image and Patient Metadata

Sriram Iyer
Student ID 23232064

Abstract

Melanoma a type of skin cancer is on the rise in the last few years due to long exposure to UV rays from the sun. Detection of melanoma is still done with visual inspection of the affected area by dermatologist but its hard to detect even by experts without additional details. With the rise of usage of deep learning algorithms it is possible to detect intricate patterns from images which could be missed by any human. But most of the machine learning studies in the domain focusses on unimodal (imaging)for detection of melanoma detection. Even though these algorithms perform well patient demographic and clinical information could also be used together with imaging for overall better diagnosis. Using such multimodal data on diagnosis could be the key to achieving better performing models with more learning capabilities.With this research aim to see how utilising such multimodal modal data effectively with hybrid model building approach helps to improve overall diagnostic accuracy of melanoma detection. We would be evaluating the multi modal approach to the standard unimodal approach across all the metrics to see how the overall accuracy and could be increased and if there is any improvement in the same.

1 Introduction

Cancer is one of the most deadliest diseases in the world as there no specific cure for the same. Skin cancer is a result of prolonged exposure to UV radiations from the sun. Skin cancer particularly melanoma is one of the most deadliest of cancer resulting in high death rate over last few years(Saleh et al. 2024) .It needs to be detected in early stages so that treatment can start early. Melanoma detection is heavily relied on analysis of image data from dermatologists but there may be patterns that a doctor may sometimes fail to identify which could be deadly for the patient. Even an experienced doctor may have a hard time detecting if a skin lesion is malignant or benign without a biopsy and additional details of patient.

Advancements in artificial intelligence and machine learning over the last decade has helped medical science a lot in terms of early diagnosis of diseases. With deep learning algorithms it is possible to process the images very high level and spot every single feature from large image datasets (Soto and Godoy (2024). Image models such as CNN has helped a lot for analysing skin lesions in a much deeper level and extract features for melanoma detection. Cancer detection over the years have generally relied much on imaging data for diagnosis and with the help of machine learning algorithms it has increased the accuracy of detection. Even though utilising data from one modality has improved the accuracy of detection it is equally important to utilise other data such as patient demographic, clinical data as they may provide more insights for early detection. There may be early signs of melanoma which may take time to show up in imaging which could be detected earlier by utilizing data from other modalities

such as csv data containing patient metadata. Most research in the domain focus more on utilising unimodal data for analysis ,but it makes more sense to utilize data of all modalities and not just imaging for better and accurate diagnosis. Hence in this research we will explore the idea of using multimodal data (imaging and patient data) and how utilising both may help with early and accurate diagnosis of melanoma.

1.1 Research Question

How can multimodal data such as imaging and patient data be effectively used with a hybrid multi-modal model building approach to impact the overall diagnostic accuracy than standard unimodal approach?

With this research firstly we aim to understand the nuances of utilizing multimodal data and how to handle them effectively. This would enable extraction of important features from both types of data and train a hybrid multimodal model for melanoma detection. We also plan on training a unimodal model trained with just image features for prediction of melanoma. So by comparing the multimodal model with a unimodal model will tell us as to how it impacts with overall accurate prediction of melanoma diagnosis.

In this research we be going through all the below phases during the implementation of the project so that we can find out how multimodal approach impacts diagnostic accuracy

1.2 Data Description

The data that we would be using for this research would be Multimodal Skin Cancer Dataset(MRA MIDAS) from Stanford website which is open source in nature. The dataset contains thousands of dermoscopic images of skin lesions which were gathered from clinic visits to dermatologists. The dataset along with the imaging also contains relevant patient metadata such as skin type, gender ,age, clinical results,etc which was also gathered during these visits and compiled in a structured data format(csv).

1.3 Exploratory Data Analysis and Preprocessing

In this phase we would be exploring to understand the data and how the attributes are related to each other and target variable. We would be checking on inconsistencies, imbalances, missing values in the data and find a way to resolve them so that it does not impact the information gain in later stages. Once the dataset is clean and we have decided on key features we would be finally encoding and normalising the variables . Image data would also be pre-processed by resizing ,ROI detection so that the model can find features easily from noise.

1.4 Model Building Phase

In this phase we would be deploying multiple deep learning models to learn features of both image and patient metadata. Model building would be done in 2 parts i.e utilising unimodal and multi-modal data(image and patient data) as to address the main problem in the research question. For Unimodal approach we would be using only the image data to train the model. For Multimodal approach we would be using same EffecientnetB0 model for extracting features from the imaging data. But to learn the patterns in patient metadata we would be using a ANN model which would be able to get insights from the same. Then finally features extracted from both models would be provide as input to a hybrid multimodal to predict melanoma cases.

1.5 Evaluation

In this final phase we would be evaluating the performance of both unimodal and multimodal approaches. Model would be compared based on accuracy, precision, recall for both classes of target variable. Also we would be checking False positives and false negatives closely along with the overall accuracy. These metrics will help us to know if the model performed good and the misdiagnosis cases which is very dangerous for any medical diagnosis research.

2 Related Work

Cancer is the one of the deadliest disease across the world and its detection at early stages is paramount for recovery. Over the years there has been many research into the cancer detection in general and with rise of machine learning algorithms it has helped a lot. Melanoma is one of the more common types of cancer and its hard to detect as given just a general image of affected skin lesion it's hard to figure out the same. Hence its really important to utilise all available data for accurate detection of the same. Here we will discuss multiple research done on cancer detection using machine learning techniques and their performance. We will go through various approaches and algorithms which were used and try to understand the overall impact of the same on model accuracy.

Jalall et al. (2023) tried using multiple machine learning techniques to compare their accuracy on lung cancer detection. In their research they were using data from only single modality and tried to compare how SVM, KNN and logistic regression performed on the same. From their research they found out SVM performed overall better in accuracy and also had very less false positives and false negatives. Even though their result was promising the size of the dataset used in the research was way smaller. Also the research only utilised patient metadata for the research and did not include any imaging data. For any cancer detection finding patterns in imaging data is just as important and utilising that data would logically help detecting cancer much more accurately.

For any cancer detection utilizing imaging data (CT, MRI scan) is very important as it may take months for a person to show symptoms externally. in their research utilized x- ray imaging data for cancer detection. Sreeprada and Vedavathi (2023) in their research they used a hybrid model OCNN-SVM which leverages deep learning to combine CNN and SVM techniques. They found out that the hybrid model provided better results compared to pretrained VGC16, ResNeT models with same learning rate. With this research we can see that using hybrid model compared to traditional approaches does help increase in accuracy compared to standard techniques. But even though this research explored the usage of a new hybrid approach for model building it still utilised just the imaging data and did not take in account the patient data. If that was added it could have learn the patterns shown in imaging and clinical diagnosis for possibly earlier and more accurate diagnosis.

With the rise of usage of CNN it has become much easier to learn from imaging data sets across all domains. Cabrejos-Yalán and Rodriguez (2024) in their research tried to use CNN for detecting various types of skin cancer from dermatological images. They found out that using CNN model the they achieved ROC of 0.99 and precison of 0.98 to detect various types of skin cancer . They used oversampling to tackle the issue of balancing all the cases of specific types of cancer so that the dataset is balanced. But with accuracy as high as 0.98 there could be issue of overfitting or model just learning features of one class which is not referenced much in their study. Oversampling which is used in the research is a great way to address the imbalance but it also creates load on memory and system as the dataset size is increased by a lot. Feature

detection is one of they aspects in imaging model and there are various methods that could have been explored like identifying Region of Interest before input to the model.

Extracting features from imaging is a very key part in melanoma detection as dermatologists make decision based on how affected skin lesion looks to see if it malignant or benign. Soto and Godoy (2024) proposed a novel solution of using active infrared thermography on the images to features can be extracted of just the affected area. They found out that when these features are provided as input to automated U-Net convolutional model and from this they achieved a 16% improvement in Jaccard index compared to other pretrained semi-automated approaches. Even though the results are very good the overall cost and compute to implement Infrared thermography based feature selection is very high. Also a hybrid deep learning model can be used in conjunction with a pretrained model like ResNet could be explored to extract more features from the images. The use of pretrained CNN models like ResNet,AlexNet is widely used in feature extraction for imaging data in medical domain due to their efficiency in detecting intricate patterns than normal CNN.

Availability of high- quality equipment's is not possible in most clinical settings hence quality of imaging are not really up to mark in most cases.Clark et al. (2024) leveraged transfer learning to address this gap for detecting Non small cell lung cancer from histopathological images. In their research they tried to pretrain the model first on high quality TCGA(The Cancer Genome Atlas) and used transfer learning to use that knowledge to train the same on low quality image dataset. They found out that by doing the same they achieved higher accuracy than the model just trained on the low resolution images. With this we can see that leveraging transfer learning and multiple datasets is a effective way to for a model to learn new patterns.

Kumaran S et al. (2024) in their research tried to use a ensemble model using three pretrained models ResNet50, VGG16, Inception V3 for their cancer diagnosis study. Instead of using just one model to extract the features in their research they let 3 separate models train on a specific features on the dataset. This way they combined the strengths of these models to effectively learn all the insights and then finally they used transfer learning to use all the features extracted from individual models. They achieved a accuracy of 98.18 % which is a big improvement in diagnostic accuracy compared to non transfer learning based models.

Saleh et al. (2024) in their research tried to compare all models accuracy utilising pretrained models such as ResNet, AlexNet, InceptionV3,Mobilenet,etc . They carried out research was carried out in a benchmark ISIC 2017 dataset which is most commonly used for image based classification models for skin cancer detection. They found out that using AlexNet gave the best results compared to all others pretrained models. With this we can clearly see that utilising a pretrained model instead of conventional CNN is much better for melanoma detection. But using just one form of data for classification has its limitations as the result is solely based on the analysis of affected skin lesion. With the usage of patient history, demographic data and a hybrid model building approach the overall accuracy could be increased and the resulting model would be able to make more precise decisions.

There has been some research in the recent years of exploring the multimodal approach to model building for Lung cancer detection. Subramanian et al. (2020) tried to use multimodal data which comprised of patient genomics and radiology imagining data in a effort to detect chances of resurgence of cancer. Even though they successfully implemented multimodal approach with neural networks they found out that the model did not perform better. One of the main reasons here could be the size of dataset as in their study had multimodal data of just

130 patients. Even though the research was promising it was limited by the availability of large dataset for learning more patterns and insights. For any model to perform better it requires a large size datasets and this is very important when it comes to multimodal approaches. There is a lack of available multimodal datasets across most domains as it requires a lot of effort to curate the same that unimodal data.

The usage of clinical data along with imaging data is very essential when it comes to neurological diseases like Parkinsons. Ganesh et al. (2023) has used data from all available data including images, sensor data ,patient records in their research for predicting Parkinsons disease. They found out that while using such combined data to train models like LSTM, Random forest it provided better accuracy than conventional approach. The resulting model would be trained on all features which makes it much more efficient than just using features from unimodal data.

With the review of available literature we can clearly see that cancer detection research has catapulted leaps and bounds with the advancements in Machine learning and Deep learning techniques. Pretrained models like ResNet, AlexNet have been used a lot for feature extraction due to their ease in use and better feature extraction from images for cancer research. Even though these models achieve high accuracy there is very few studies utilising other forms of data like patient metadata for Melanoma Detection. This is mostly because of the fact that unimodal data is readily available whereas proper multi modal datasets are very rarely available. Hence in this research we would be utilising multimodal data and building a hybrid model which takes decisions based on both imaging and patient metadata features.

3 Research Methodology

The core idea of this research is going to be on how the use of multimodal data could provide better results than the conventional unimodal approaches. Data is available in multiple forms and it makes more sense to use all available data types for prediction of high risk diseases such as Skin Cancer. So imaging data and tabular data containing patient metadata would be the type of data we would be using for the research. For this research the multimodal data used is Multimodal Image Dataset for AI Based Skin Cancer MRA MIDAS dataset from Stanford university website.

3.1 Data Description

- **midas_record_id**: Unique patient identifier
- **midas_file_name**: Name of the image file associated with patient
- **midas_iscontrol**: If a record is from control sample
- **midas_distance**: Distance from which image was taken
- **midas_location**: Location of the skin lesion (e.g. left arm ,chest)
- **midas_gender**: Sex of the patient
- **midas_age**: Patient's age.
- **midas_fitzpatrick**: patient's skin type and colour of their eyes
- **midas_melanoma**: If the patient has melanoma (Target variable)
- **midas_ethnicity**: Ethnicity of patient
- **midas_race**: Race of patient
- **clinical_impression_1**: First clinical diagnosis from a dermatologist

- **clinical_impression_2**: Second diagnosis from another dermatologist
- **clinical_impression_3**: Third diagnosis from a dermatologist.
- **length_(mm)**: Length of affected skin lesion
- **width_(mm)**: Width of the affected skin lesion.

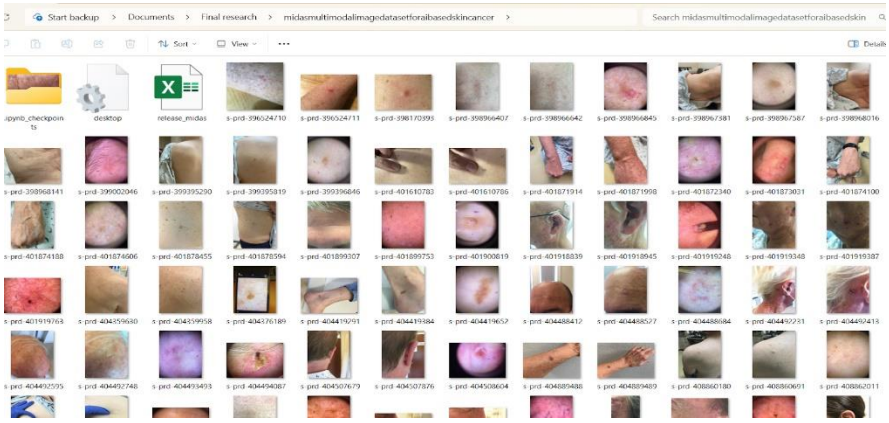

Figure 1 Image Data

As we can see from the above figure 1 the imaging data includes an array of all kinds of skin lesions which one would typically see in reality in clinics that doctors would examine for melanoma. The patient metadata includes an array of features such as age, gender, skin type ,clinical impressions etc which would provide additional information to learn for our model. Based on the structure of the data ,target variable that would be used to train the model is the feature midas_melanoma. The target variable is binary in nature with yes and no values for if the patient has melanoma or not given the features.

3.2 Exploratory Data Analysis and Preprocessing

For the patient metadata it has a multitude of features that provide more details as to the patient who had undergone a diagnosis for the skin lesion. The data has lot of missing values among the variables and these needs to be addressed accordingly. But before any of that we would be exploring the data to understand the variables ,how they are interdependent with each other and the target variable. This will help in getting a basic understanding on the patterns which are already present in the data and how they influence target variable. Also based on the analysis a proper way to address the missing values would be used so that it does not impact the integrity of the data and the model building entirely. The numerical columns such as length and width of the lesion which showed skewness and outliers would be handled by median imputation as it is not impacted much from outliers. For the other categorical columns which still have some NA values would be imputed with “unknown” as this will preserve the information intact without creating any unnecessary bias with other methods.

Based on the exploratory data analysis, feature selection would be done so that only the necessary variables would be used for model building phase. For we would be using a random forest classifier for us to analyse the importance of the variable. Based on this the top 10 features would be used for the next stage that is the model building phase. These features would be normalized and encoded prior to the model phase as the data needs to be scale free and unit free for the same.

As any data prone to inconsistencies and missing values first data would be pre-processed so that it is consistent prior to model building. As in this research we would be using both image and csv data, both of them needs to be preprocessed separately prior to model building phase.

As we have patient metadata associated to these images ,we would at first go through the image names in midas_filenames column and only go through the images that have entry in the column. This is a very essential step as it makes sure that we remove redundant or duplicate images of same record. For image data as we would be resizing all the images to 128 x 128 for dimensional uniformity across all images and better and faster feature learning. To remove noise and focus only on affected area from a particular image we would be using CLAHE for identifying Region of Interest. Contrast Limited Adaptive Histogram Equalisation(CLAHE) is a technique used to highlight or enhance contrast of images so that the affected area could be easily detected from a image. So with these preprocessing we ensure that we provide model proper input so that it can learn patterns faster and easily.

3.3 Model Building Stage

Once the data was preprocessed, two separate models were built to compare the effectiveness of using only image data versus combining image data with tabular metadata. This was done to see if using more data types together could provide better predictions for melanoma. The two models were:

3.3.1 Image-Only Model

For our unimodal approach we would be using EffecientNetB0 which is a pretrained deep learning model. EffecientNet is known to be very lightweight compared to other pretrained models like ResNet. One of the main reasons for using this over other models is that the size of the compiled model is very low and it provides almost same level of accuracy as other pretrained models with faster training time. EffecienetB0 is pre-trained on imagenet dataset which means it has weights already initialised for better feature extraction from the input images. Once the model is trained it would be tested in test data set to see how accurately the model is able to predict the class variable midas_melanoma

3.3.2 Multimodal Approach

For the multimodal approach which is the main part of the research we would be utilising both the image and patient metadata for model building . But since we are dealing with different data types we would first train these data separately with different models to extract the main features of the individual data types. For the imaging branch we would be using the same EffecientnetB0 model to train, but for patient metadata we would be using a ANN model to train on the patient metdata features. The output features of both of these individual models would be passed as input to our hybrid multimodal model which would have both imaging and patient data features. Once the multimodal model is trained we would be testing the same to see if there is any changes in accuracy from unimodal model.

3.4 Evaluation

In the evaluation stage we would be evaluating both the multimodal and unimodal model on metrics such as accuracy, precision, recall, F1 score. We would be checking confusion matrix and comparing the false positives and false negatives cases in particular as these are very important statistics for a medical domain. Based on how the model is predicting both positive and negative cases across both models we would be finally get a answer to our research question.

4 System Design

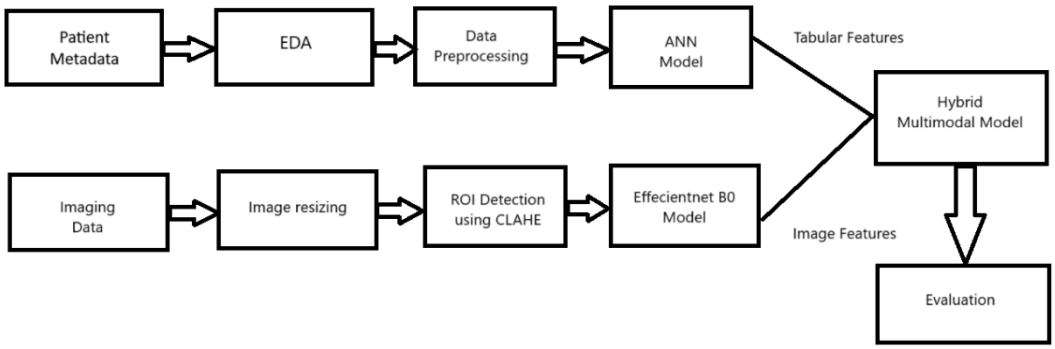

Figure 2 System Architecture Diagram

As seen from the above system design the first part multimodal implementation would consists of firstly preprocessing patient metadata and imaging data separately. Then the pre-processed images and patient features would be passed as input to EffecientnetB0 and ANN model respectively. This is done so that the key important features of data from both modalities could be extracted from the dataset. Then these tabular and image features would be passed as input to our multimodal model which would be able to predict target variable midas_melanaoma given both patient metadata and image features.

5 Implementation

So in this research as we are going to be using MRA MIDAS multimodal dataset for skin cancer from Stanford for training our models. The dataset comprises of both image and patient metadata for us to use to derive patterns from Machine learning models. But it is essential for us to understand the various features in the patient metadata and how they are related before we perform anything on the same. We also need to understand how these attributes affect one another and our target variable(midas_melanoma). So in the first part of the implementation, we will proceed with Exploratory data analysis for us to get to know the data that we would be using for the research.

5.1 Exploratory Data Analysis and Visualisations

There are in total of 18 attributes in the patient metadata from which 2 of them ‘Unnamed’ and ‘midas_record_id’ are irrelevant as they would not provide any meaningful information for our model. In this phase we will go through some of the visualisations and try to find some visible patterns that would help us understand the nature of the dataset. Firstly we will start to see how age impacts the melanoma diagnosis and for this we would be using a histogram to visualise the same.

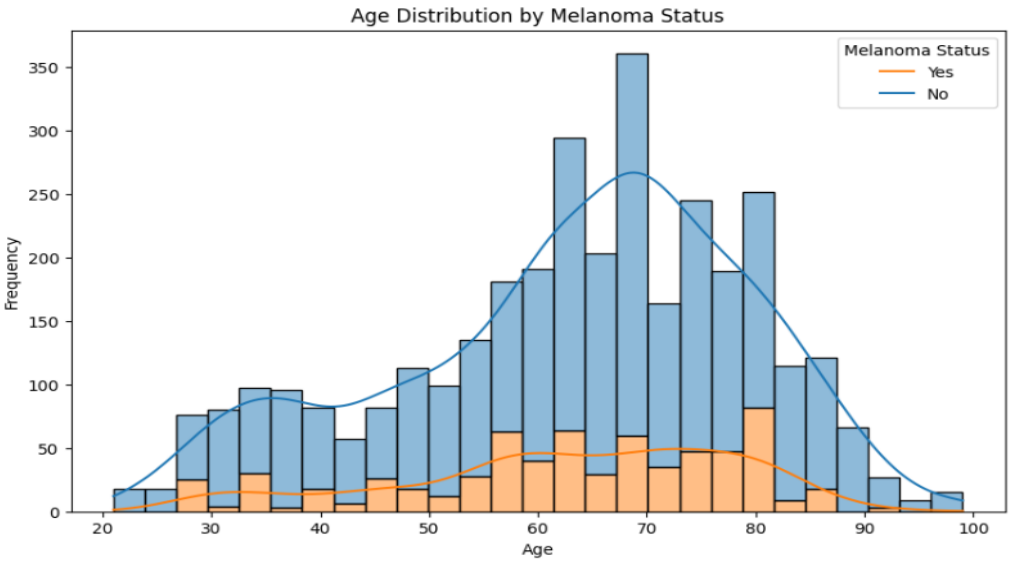

Figure 3 Histogram of Age with Melanoma

From the above figure 3 we can clearly see that people above the age of 50 show signs of melanoma more than the younger ones. This observation makes sense as older people would have weakened immune system and the prolonged exposure to UV rays throughout the years would make them easily susceptible to melanoma. So with this we can clearly see that age is one of the key factors that would be needed for melanoma diagnosis and needs to be part of the model building in later stage. One of the other things that we will be checking is how does gender impact in melanoma detection in our dataset . Generally cases melanoma is seen in men more than women and it is essential for us to see if that is the case here in our dataset.

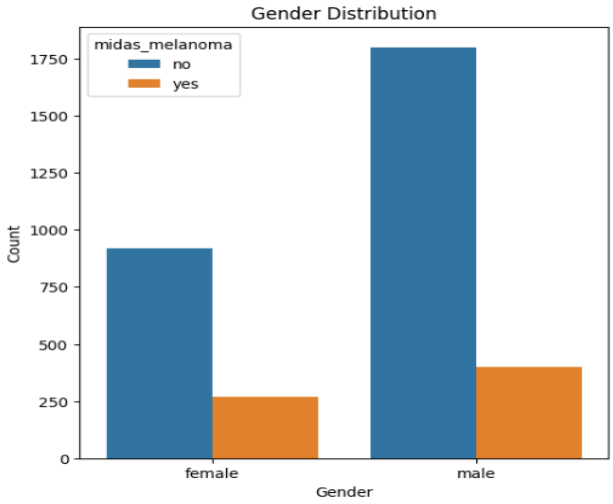

Figure 4 Histogram of Melanoma cases by Gender

As we can see from the histogram men have more melanoma cases reported than women and this is in line with our assumption. This makes sense as men in general do not take care of their skin as much as women and this makes them more exposed to UV rays which results in melanoma over the years. Since we are trying to understand different attributes, we would also be checking how certain skin types matters in melanoma diagnosis. So for this we would be plotting a histogram to see if there is a certain skin type that is more susceptible to melanoma.

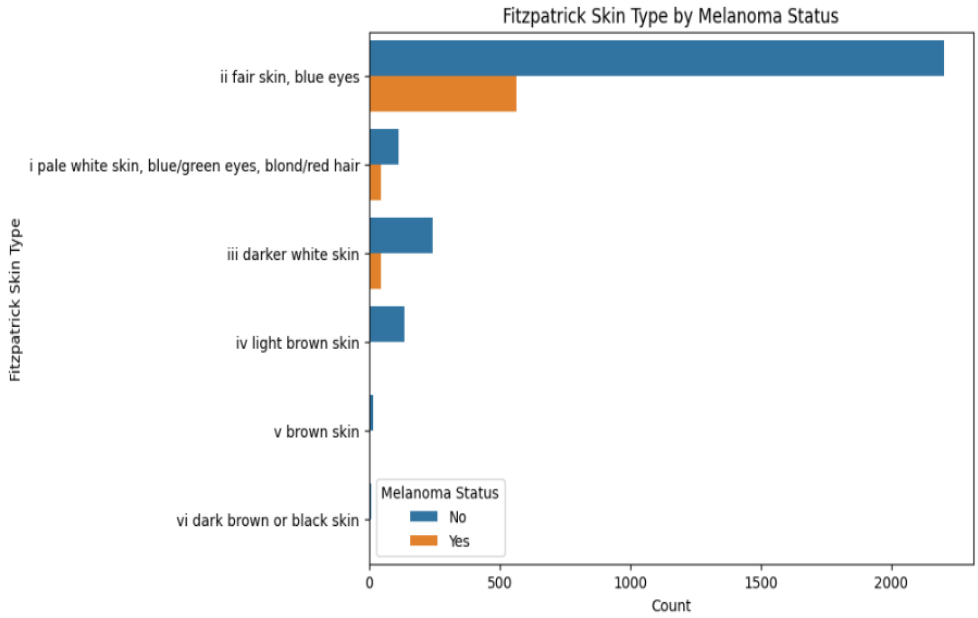

Figure 5 Melanoma Cases with Skin Type

As we can see from the histogram melanoma cases are seen mostly with people with fair skin than darker skin. This is a genetic factor which is not in control of anyone but this insight is very crucial and needs to be part of model building phase.With these graphs we have a basic idea about some of the factors in the dataset and how they impact our target variable.

Before we go any further we will first check for inconsistencies in the data i.e missing values,NA values etc in our dataset. These inconsistencies need to be addressed prior to any future actions as this will impact the learning of our model.When we checked the total number of NA values in the dataset we found out that there were many columns with high number of NA values in the dataset.

Also upon checking we found out that even the target variable midas_melanoma also had 30 missing values. Firstly we have removed the entire rows containing no target variable value because any imputation on target variable can cause unnecessary bias to the data. Variables such as midas_path,midas_pathreport have close to 500 NA values and these columns would be excluded as the total rows in the dataset itself is around 3000 rows.It did not make sense to impute around 500 values in a column as that would create inconsistency and cause the model to learn wrong patterns.

For the numerical columns we have length and width of the lesion having 17 NA values each which was resolved by using median imputation.This is because as checked in the histogram in the below figure 5 we can see there is a right skew in both of these columns. Median imputation does not effect the overall skewness of the data and causes minimal distortion in the data which does not create any outliers in the same.

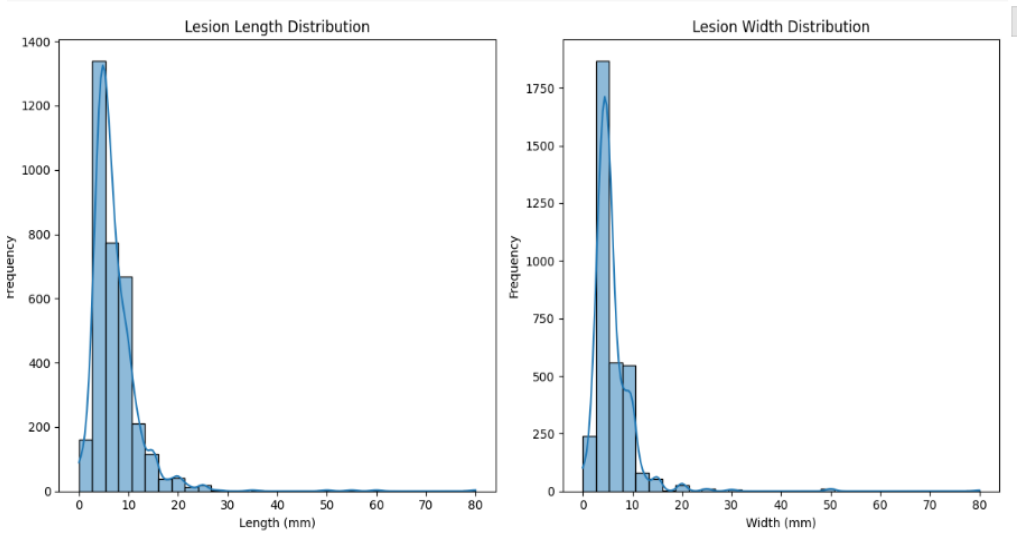

Figure 6 Histogram of Length and Width of Lesion

Remaining columns with NA values which needed to be addressed were midas_fitzpatrick, clinical_impression_1, clinical_impression_2 which were categorical columns.Mode imputation did not seem right as most cases in the dataset had benign diagnosis and this would make the dataset more imbalanced than it is already.Also dropping these columns would not work as these are one of the key attributes we need our Machine learning model to be trained. So the missing values were imputed with "Unknown" values as this maintained the overall structure and the dataset without creating any unneccessary bias or variance in the data.

5.2 Data Preprocessing and Feature Engineering

For Tabular Data

Once all the missing values are dealt with in the patient metadata, we then procced with understanding which features should be actually used in the next phase. At first we would be dropping columns Unamed, midas_record_id as stated earlier they do not provide any information related to data hence we would be dropping the columns. Midas_distance contains the distance from which the image was taken of the skin lesion ,we would not be using for model building as this does not help with providing more information on patient and we are preprocessing images separately and not with patient data. Also column midas_filename exists only to check and load the matching imaging data for the record ,after loading and preprocessing of the image data the column would be removed from the dataframe.

So excluding these columns we still need to find out top features that influences the target variable the most from the others. For this a random forest classifier is run on the target variable given the features to see which of them can be used to build the model.

Rank	Feature	Importance
1	midas_age	0.102852
2	width_(mm)	0.077259
3	length_(mm)	0.076017
4	midas_gender_male	0.020078
5	clinical_impression_3_11-malignant-melanoma	0.016871
6	clinical_impression_2_11-malignant-melanoma	0.014569
7	clinical_impression_1_11-malignant-melanoma	0.013095
8	midas_location_left upper arm	0.012144
9	clinical_impression_3_unknown	0.012038
10	midas_iscontrol_yes	0.011588

Table 1 Feature Importance

From the above table we can see that age is one of the key features with the highest importance value. Also length and width of the lesion are equally important and it makes sense too and the length and width of the affected skin lesion if more has higher chances of melanoma. The other attributes include gender ,clinical impressions ,midas_location ,midas_control. Most of these attributes were assumed in general to provide more insights and we can definitively say that these attributes should be used for model building stage. Clinical impressions include the doctors clinical diagnosis and this is also one of the key attributes besides patient demographic as it provides additional information for model to learn patterns.

Once we have decided on the key features that we would be using we now need to standardise and encode the values before feeding into a machine learning model.This stage is very essential and important part of preprocessing as string values cannot be entered to the model directly. It first needs to be encoded and in this project we have used label encoder library in python to encode the categorical values to numerical values.The numerical columns such as age ,length and width all follow different scale so we would standardising these values using StandardScaler() function to make it scale free and unit free. Once all of these transformations are done we checked to see if there is any class imbalance in our data. This needs to be checked at this very stage as if there is imbalance it may cause the model to only learn from one class and not the other.

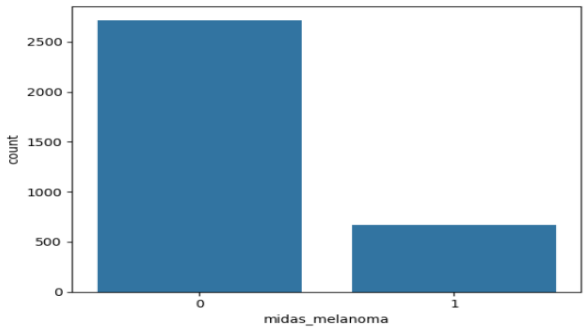

Figure 7 Count Plot of midas_melanoma

The above figure 7 shows us that there is heavy imbalance in our target variable midas_melanoma. This is common for most medical datasets as in any case there would lot less yes values than there is no values. To address this in our project we would be using a class weights to address the issue. With the class weights parameter it enables to give more or less importance to a specific imbalanced class. This ensures that the model learns patterns of true cases with the same importance as false values. Another reason for using class weights is that it does not increase the size of the dataset and handles the imbalance by assigning higher

weights to underrepresented class. Since we are dealing with multimodal data in this project which includes pre-processing images it utilises a lot of memory and resources. So if the dataset is oversampled it will essentially double the data size and that with the image data would create very high resource utilisation. This would cause long training time and kernel to crash and hence in this research we have used class weights to address the imbalance.

For Imaging Data-

Since the data is multimodal in nature image data also needs to be pre-processed before it is provided as input to Effecientnet model. The images in the dataset comprises of photos of affected skin lesions from various parts of the body. These were actual photographs submitted to dermatologists for their review and the metadata file which we discussed earlier had the clinical diagnosis of the doctors. These photographs in itself cannot be used in their raw form and need to be preprocessed prior to input to the model. So firstly the images would be filtered to ones that have a entry in the midas_filename column. Since we are building a multimodal model it is essential that only images with matching entry in patient metadata is loaded as this ensures that there is no error while combining features at the final stage of model.

Once the images are filtered and only the ones that have associated values are selected images are resized to 128 x 128 . The main reason to resize it to 128 x 128 is to reduce the memory and resource utilisation overhead of loading all the images. Since we would be running a image model as well as multimodal model and the size of the entire image dataset is around 3 GB it would utilise a lot of resources. So by resizing into lower dimensions it ensures faster model training time and efficient memory usage. Just resizing the images is not enough as model needs to find the affected skin lesion from the picture. Just providing the raw images to the model would not yield better results and so before the image is provided to the model we would be detecting Region of Interest(ROI) from the image. The region of interest is used to highlight the affected skin lesion from the image so that image model can check the same. In this project we would be using CLAHE(Contrast Limited Adaptive Histogram Equalization) for identifying the ROI from the image.

CLAHE is a technique which is used to enhance the quality of the image so that the areas with no lighting could be easily seen and reduce the overall noise in the data. With this technique the region of interest could be easily targeted from all the noise in the image .How this technique works is that first the image is converted into grayscale as clahe cannot work in coloured images. Then the image is passed through the clahe function which essentially converts the entire image in multiple small tiles and then works on reducing the noise across by interpolating the nearby tiles. Once all of this is done it is image is converted back to RGB format as this is required as input to the Efficientnet model. Then the converted image is resized again to 128 x 128 and then normalised so that the output could be stored in a numpy array and fed into the image model as a input.The below figure 8 shows how CLAHE works by adjusting contrast on the image such that main region of interest is clearly seen among the noise . The below figure shows how ROI looks in some of the images and highlights in small rectangle. We are able to see the affected area much clearly and most the noise is removed from the actual image. Resizing into a smaller dimension also helps a lot in this case as the affected area can be seen clearly and with this we have finally completed image preprocessing.

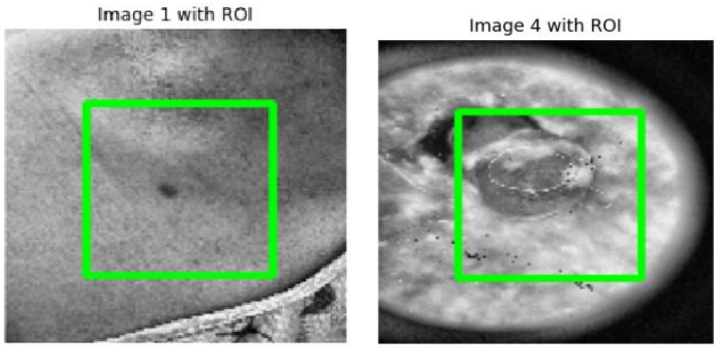

Figure 8 ROI detection using CLAHE

5.3 Model Building Phase

This phase would be divided into 2 sections as one part will show how the image only model (Unimodal) and the second part will show how the multimodal model is built and trained on the respective data. For image only model we would be using Effecientnet which is a pretrained model and used in lot of medical imaging datasets as it extracts features much more efficiently than standard CNN model. For the multimodal approach the we would be using the same Effecientnet Model for images but for the tabular data we would be using a ANN model to extract the features. The final multimodal model is a hybrid model having the features of both image and tabular model as input and predicting the target variable.

5.3.1 For Image only Model (Unimodal Approach)

As all the preprocessing of the images is done in the earlier stages ,we finally have a image data which is normalised numpy array which can be used as input to model.
In this research we would be using EffecientnetB0 Model for feature extraction and learning patterns from images.

Preprocessed images with ROI using CLAHE would be used as input to our EffecientNetB0 model. But since we are dealing with just images here we would be adding additional Squeeze and Excitation(SE) Block to our Model. SE block is added specially to get more contextual understanding of the features as we are not using any patient metadata model is limited on just features extracted from the model. Then output of the SE block is passed as input to pooling layer which is added to reduce the dimensionality of the model. Then the output of the pooling layer is fed into a dense layer of fully connected 128 neurons with dropout rate of 0.5. This ensures that the model does not just learn patterns of one feature and avoids overfitting. Then the final layer consists of just one neuron as we are dealing with binary classification with sigmoid activation function

With the above model architecture the image only model was trained with learning rate of 0.0003 and with adam optimizer. Callbacks are also added so that the model stops the training process if it does not see any decrease in validation loss and this ensures it stops training of the model if there is nothing more to learn.

5.3.2 Multimodal Model

For both imaging data we would be using the Efficientnet with some changes to image only model . But first we will explore on the Tabular model branch of patient metadata and how the ANN model is built for training.

Layer	Details
Input Layer	Patient metadata features as input
Dense Layer 1	128 neurons, Activation: ReLU, Kernel Regularizer: L2 (0.01)
Batch Normalization 1	Normalizes the output of the first dense layer
Dropout Layer 1	Dropout Rate: 30%
Dense Layer 2	64 nerons, Activation: ReLU, Kernel Regularizer: L2 (0.01)
Batch Normalization 2	Normalizes the output of the second dense layer
Dense Layer 3 (Output)	32 neurons, Activation: ReLU, Kernel Regularizer: L2 (0.01). Output as metadata_features.

Table 2 Patient Metadata ANN model architecture

The above table is for the ANN model build that we would be using for the tabular data.As we can see at first the data which was pre processed earlier and split into training and test data ,the training data is taken as input along with the total dimensions of the data.
This is the first input layer and from which data pases through first dense layer with fully connected 128 neurons. We have chosen relu as the activation function with l2 regularizer is used to prevent overfitting of the model.

Then the next layer is the normalisation layer to normalise the output of previous layer and in this model. In this we have chosen dropout value as 0.3 which means 30% of neurons would be dropped in every cycle and this makes sure that the model does not rely on just one feature. Next is a dense layer is of 64 neurons fully connected which has same activation and l2 regularisation as first dense layer. The final layer just contains 32 neurons from 62 with same parameters which is so that the outputs finally converge to binary classifier of either true or false values.
This code enables such that the model does not overfit much and learns unique insights every time and not just dependent on one specific feature.

For the image model we have the used EffecientnetB0 for feature extraction and the Neural network model . Firstly the weights of the effecientnetb0 model is set to same as that of imagenet which means that these are the weights learned from that dataset. This means that it makes it easy to learn general features same as that of imagenet model.

Layer	Details
Input Layer	Shape: (128, 128, 3) Size of resized preprocessed image
EfficientNetB0 Base	Pre-trained on ImageNet, include_top=False. All layers except the last 20 are frozen.
Global Average Pooling	Aggregates spatial dimensions of the feature map output from EfficientNet to a single vector.
Dense Layer 1	128 units, Activation: ReLU, Kernel Regularizer: L2 (0.01)
Batch Normalization 1	Normalizes the output of the first dense layer for stable and faster training.
Dropout Layer 1	Dropout Rate: 40%
Dense Layer 2 (Output)	64 units, Activation: ReLU, Kernel Regularizer: L2 (0.01). Output as image_features.

Table 3 EffecientnetB0 architecture

The above table shows the EfficientNetB0 model architecture used to train image dataset. Also most of the layers of the model is frozen and only the last 20 layers are trainable of the imagenet model. This makes it easier for model to learn the patterns from the input data as well as leverage the pretrained weights so that it avoids any overfitting. Next we have a pooling layer which is to reduce the dimensionality of previous layer and then its passed to a dense layer. In this layer there are 128 fully connected neurons with relu activation function and l2 regularisation. Then the output of this layer is normalised and then dropout is set to 40% which is to make sure that model does not just learn one pattern over and over and overfit. Then the final layer has 64 neurons which is to reduce the dimensionality and narrow in more on important features from the image.

Layer	Details
Input (Image Branch)	Input: image_input from EfficientNet branch (128x128x3), producing image_features.
Input (Tabular Branch)	Input: tabular_input from the tabular model branch, producing metadata_features.
Concatenate Layer	Combines image_features and metadata_features into a single vector called combined_features.
Dense Layer 1	64 units, Activation: ReLU, Kernel Regularizer: L2 (0.01).
Batch Normalization 1	Normalizes the output of the first dense layer.
Dropout Layer 1	Dropout Rate: 30%.
Output Layer	1 unit, Activation: Sigmoid, providing the final binary classification output.
Optimizer	Adam, Learning Rate: 0.0003.
Loss Function	Binary Cross-Entropy.
Callbacks	ReduceLROnPlateau (reduce learning rate on plateau), EarlyStopping (stop if no improvement).

Table 4 Multimodal Model Architecture

The above table shows the hybrid multimodal model architecture which is used to train on both features of the data. Once the features and insights are extracted from data of both modalities i.e image branch and tabular branch we would proceed to create a hybrid multimodal model. This model will make use of transfer learning as input to this model would be the features extracted from the previous 2 models. In the first layer itself both the image and metadata features would be concatenated and this fused input is passed through a dense layer which

utilises relu as all the other models. Dropout value is set to 0.3 and the final layer just contains one neuron as our class variable midas_melanoma is binary in nature with true or false values. For the final model we have added a callback which is to reduce the learning rate and early stopping. This is added so that the if validation loss stops improving from a certain point in a few epochs it will automatically reduce the learning rate. Early stopping is also added so that the model does not overfit and stops the the training of the model if learning gets stagnant.

Once the model is compiled we would be passing both the image and clinical metadata as input and start the training process. We have set the total number of epochs as 10 and once the model is trained we will be testing the same on our test dataset. Then we once we have the necessary statistics of Accuracy, F1 score , precison and recall we will compare the same with image only model and see how it performed in comparison.

6 Evaluation

Once both the image only and the multimodal model are trained would be testing the built model on test data to see how the model performed in comparison to training stage.We would critically evaluating both the image only and the multimodal model on metrics such as Precision ,Recall ,F1 score ,False Positives and False Negatives . Accuracy of a model is not always a key indicator as the model could be overfitted or could learn patterns of just one class variable. Hence for any medical diagnostic model it is necessary to see how the model performed overall in detecting both benign and malignant melanoma cases. This enables us to see how the both the models performed in detecting cases accurately and not misdiagnosing cases.

6.1 – Image Only EffecientNetB0 model

Epoch	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss	Learning Rate	
1	0.7457	0.5401	0.448	1.3464	7.50E-05	
2	0.8231	0.4216	0.496	1.612	7.50E-05	
3	0.8352	0.3817	0.272	0.8583	7.50E-05	
4	0.8577	0.3375	0.2032	1.0588	7.50E-05	
5	0.8739	0.294	0.2032	1.2266	7.50E-05	Early stopping triggered.

Table 5 Model Performance on Test data

The above table 5 shows the models performance throughout the 8 epochs of training. It could be seen that with the increasing epochs the training accuracy is increasing exponentially but the validation accuracy seems to be fluctuating and getting lower. This suggests the sign of overfitting in training data and due to this the callback function it stopped early at 5^{th} epoch.To analyse this further we would be checking the classification matrix to get the bigger picture of overall model performance detecting both classes of target variable.

Metric	No Melanoma	Melanoma	Overall
Precision	0.8	0.21	**Macro Avg: 0.51**
Recall	0.41	0.61	**Macro Avg: 0.51**
F1-Score	0.54	0.31	**Accuracy: 0.45**
Support	498	127	**Total: 625**
Confusion Matrix	**202 (TN), 296 (FP)**	**49 (FN), 78 (TP)**	-
Test Accuracy	-	-	**44.80%**
Test Loss	-	-	**1.3464**

Table 6 Statistics for Image Model on Test Data

From the above figure we can see that the that the model is giving us a overall accuracy of 44% on test data. Upon checking the confusion matrix we can see the reason why it performed poorly is that it almost predicted around the same as malignant and benign. The false positive rate very high as the model misdiagnosed around 296 as positive cases. But we can see that the false negatives rate is much less which means model missed on 49 cancerous cases. This accuracy overall is very low and without much information on patient metadata it makes sense how model performed so low just based on imaging features. Next we will see how the multimodal model performed in comparison.

6.2 Multimodal Model

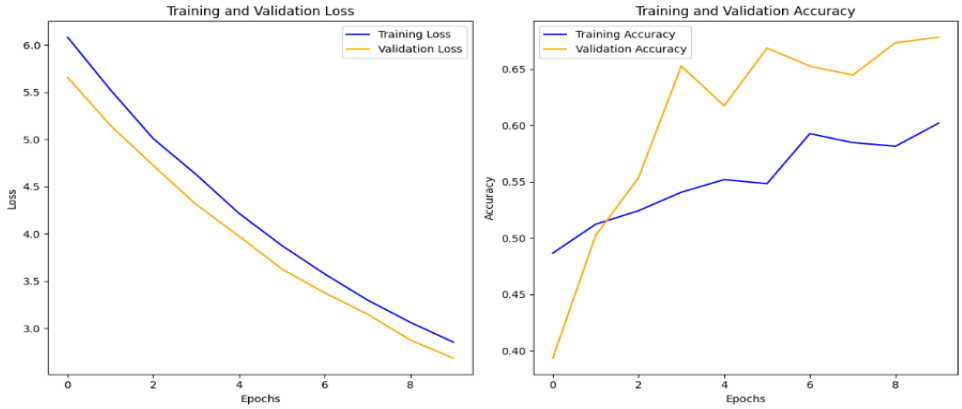

Figure 9 Training and Validation Epoch Plots

The above figure 9 shows graph of models performance throughout the 10 epochs of training. As seen from the graph the training and validation loss keeps on decreasing with every epoch. Both the losses are decreasing gradually which means that the model is not overfitting. Models accuracy over training and validation also seem to increasing with validation accuracy increasing more than training after first epoch. This could be due to addition of l2 regularisation and dropout value affecting the training of the model but not during validation phase. Next we need to check the classification matrix to see how the model performed in terms of predictions of both class variables.

Metric	No Melanoma (0.0)	Melanoma (1.0)	Overall
Precision	0.87	0.32	**Macro Avg: 0.59**
Recall	0.69	0.58	**Macro Avg: 0.63**
F1-Score	0.77	0.41	**Accuracy: 0.67**
Support	498	127	**Total: 625**
Confusion Matrix	**342 (TN), 156 (FP)**	**53 (FN), 74 (TP)**	-
Test Accuracy	-	-	**66.56%**
Test Loss	-	-	**3.1297**

Table 7 Multimodal Model Statistics

The above figure shows the performance on test data and here we can see that the multimodal approach is giving a overall 66% accuracy on prediction. But when we see the confusion matrix we can see that model is still having high false negatives and false positives. But when compared to image only model we can see significant improvement in false positive values which had 296 compared to multimodal which only had 156. But there is a slight downgrade in False negative values for which the image model had 49 compared to 53 of multimodal model. Having more false negatives are more deadlier than false positives is bad for any medical diagnosis as this mis diagnosis may cause the patient their life. But the difference is very minimal in this case and the high accuracy, precision and better balance of false positives to negatives makes the multimodal approach still better than just image only model.

6.3 Discussion

From this research we can see that there is clear improvement in better diagnosis of melanoma with multimodal approach than unimodal approach. Even though the overall accuracy is not that good with 66% this research shows that utilising multimodal data effectively can definitely result in better diagnosis. The accuracy of this multimodal approach can be increased further with addition of more features in patient metadata and better quality of the images. Having a subject matter expert like dermatologist help throughout the model building process would also help as they have more understanding of the disease and can aid us in steps that could be added to improve image or metadata pre-processing . Also integrating multiple sources of data for training would increase in finding more patterns and better diagnostic accuracy.

7 Conclusion and Future Work

The main goal of the research question was to see how the utilising multimodal approach for melanoma detection impacts on the overall diagnostic accuracy. For this we finally have a answer that using multimodal data like imaging and patient metadata results in higher accuracy to traditional unimodal approaches. The overall accuracy achieved by our hybrid multimodal model has clearly shown a improvement from the image only model.But the overall accuracy is very less and can still be improved upon by exploring other pretrained models like ResNet, AlexNet for features extraction.

In future research additional data from other sources could be added increasing the overall dataset size for model training with more variations. Also multiple pretrained models could be used to train on a specific aspect of a image and then the combined learning learnings could be passed as input to final model. This would ensure that no specific pattern or insight is missed from a image and used for the final model.

8 References

Cabrejos-Yalán, V.M. and Rodriguez, C. (2024) 'Convolutional Neural Network Model for Skin Cancer Diagnosis in a Dermatological Center', *Mathematical Modelling of Engineering Problems*, 11(11), pp. 2997–3005. Available at: https://doi.org/10.18280/mmep.111112.

Clark, M. *et al.* (2024) 'Transfer Learning for Mortality Prediction in Non-Small Cell Lung Cancer with Low-Resolution Histopathology Slide Snapshots', in *Studies in Health Technology and Informatics*. IOS Press BV, pp. 735–739. Available at: https://doi.org/10.3233/SHTI231062.

Ganesh, D., Gautam, A.K. and Bhambu, P. (2023) 'Multimodal Data Fusion and Machine Learning for Comprehensive Management of Parkinson's Disease in Healthcare', in *3rd IEEE International Conference on ICT in Business Industry and Government, ICTBIG 2023*. Institute of Electrical and Electronics Engineers Inc. Available at: https://doi.org/10.1109/ICTBIG59752.2023.10456100.

Jalall, S.K. *et al.* (2023) 'Supervised learning techniques for detection of Lung Carcinoma', in *Journal of Physics: Conference Series*. Institute of Physics. Available at: https://doi.org/10.1088/1742-6596/2571/1/012004.

Kumaran S, Y. *et al.* (2024) 'Explainable lung cancer classification with ensemble transfer learning of VGG16, Resnet50 and InceptionV3 using grad-cam', *BMC Medical Imaging*, 24(1). Available at: https://doi.org/10.1186/s12880-024-01345-x.

Saleh, N., Hassan, M.A. and Salaheldin, A.M. (2024) 'Skin cancer classification based on an optimized convolutional neural network and multicriteria decision-making', *Scientific Reports*, 14(1). Available at: https://doi.org/10.1038/s41598-024-67424-9.

Soto, R.F. and Godoy, S.E. (2024) 'An automatic approach to detect skin cancer utilizing active infrared thermography', *Heliyon*, 10(23). Available at: https://doi.org/10.1016/j.heliyon.2024.e40608.

Sreeprada, V. and Vedavathi, K. (2023) 'Lung Cancer Detection from X-Ray Images using Hybrid Deep Learning Technique', in *Procedia Computer Science*. Elsevier B.V., pp. 467–474. Available at: https://doi.org/10.1016/j.procs.2023.12.102.

Subramanian, V., Do, M.N. and Syeda-Mahmood, T. (2020). Multimodal fusion of imaging and genomics for lung cancer recurrence prediction. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, pp. 804-808. doi: 10.1109/ISBI45749.2020.9098545.