Design and Implementation of a Multi-modal Framework for Disease Detection and Classification Using Machine Learning

By

Omar Abukhalaf, Omar Tubeileh & Yazan Alsharif

Supervised by

Dr. Rajaa Alqudah

Submitted in partial fulfillment of the requirements for the degree of

BACHELOR OF SCIENCE
in

Computer Engineering

at

PRINCESS SUMAYA UNIVERSITY FOR TECHNOLOGY

Amman, Jordan

Spring 2024/2025

This is to certify that I have examined

this copy of an engineering documentation by

Omar Abukhalaf, Omar Tubeileh & Yazan Alsharif

And have found that it is complete and satisfactory in all respects,
And that any and all revisions required by the final Examining Committee have been made

---

Dr. Rajaa Alqudah (Supervisor)

Dr. Awos Kanaan (Department Head)

Acknowledgments

The team would like to extend its heartfelt gratitude to Dr. Rajaa for her exceptional guidance, unwavering support, and invaluable insights throughout this project. Her expertise and dedication were instrumental in shaping our work and enabling us to achieve our goals.

Abstract

This project presents the design and implementation of a multi-modal machine learning framework to improve disease detection and classification by integrating medical imaging and clinical text data. The proposed framework employs deep learning models—ResNet50, DenseNet121, and EfficientNet—for analyzing CT and MRI scans, alongside language models such as DistilBERT and BiomedBERT for processing diagnostic reports. The dataset was sourced from MedPix 2.0, a publicly available medical database containing paired images and textual case descriptions. To address the class imbalance and enhance model generalization, various augmentation techniques were applied to both modalities. Multiple fusion strategies, including early and late fusion, were explored to effectively combine visual and textual features. Model performance was evaluated using precision, recall, F1-score, and AUC-ROC metrics. The best-performing configuration—EfficientNet combined with Random Forest using weighted averaging late fusion—achieved an accuracy of 96.62% and F1-scores above 0.95. These results highlight the potential of multi-modal learning in advancing AI-driven diagnostic systems.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Overview and Objectives

The rapid growth of Artificial Intelligence (AI) in the healthcare field has opened new ways to improve disease detection and diagnosis [14]. Healthcare systems worldwide generate large amounts of medical data, such as clinical data, patient demographics, and imaging. There exists an opportunity to use this data for faster and more accurate diagnoses. Traditional disease diagnosis relies on a single data type, such as images, patient history, or lab results. Complex diseases with overlapping symptoms often require a more extensive approach with multiple data types. These diseases include cancer, cardiovascular diseases, and neurological diseases.

Image-based systems offer powerful insights and patterns in finding abnormalities, but they lack the context that structured data offers. Similarly, text data may highlight abnormal findings, but they lack the critical information found in medical images. Relying on a single data source may not be accurate and lead to a wrong diagnosis, highlighting the need for systems that use multiple data types. This is why multi-modal machine learning that utilizes diverse data types has emerged as a solution.

The core objective of the framework proposed in this project is to analyze medical images such as CT scans, MRIs, and X-rays, along with text-based clinical data such as patient symptoms and patient history. By using Machine Learning algorithms, the framework will analyze the multi-modal data inputs to detect diseases with high accuracy. This integration doesn't only improve the capabilities but also makes sure that insights from the different data sources aren't overlooked.

1.2 List of Design Requirements

1- A minimum of two ML Algorithms shall be implemented and rigorously evaluated on performance.

2- The system shall utilize a dataset that includes both medical imaging data and corresponding text-based clinical reports such as (such as patient symptoms, patient history, and lab results.

3- The system shall process medical images from various sources, such as MRI, CT, and X-ray images.

4- The system shall evaluate performance using a minimum of four metrics

5- The system shall ensure all data is handled according to privacy regulations, such as The Health Insurance Portability and Accountability Act (HIPAA) for medical data protection, ensuring patient confidentiality.

6- The system shall support various medical image formats (e.g., DICOM, PNG, JPG) and accommodate variations in lighting, resolution, and overall image quality during the preprocessing stage.

7- Data augmentation shall be used to improve the balance of the dataset and increase the model's robustness.

The implementation of at least 2 distinct Machine Learning Models is a fundamental design requirement, ensuring an accurate disease diagnoses framework. Utilizing multiple algorithms, such as Convolutional Neural Networks (CNN), to process the medical images and transformers to process the text-based data greatly improves the system's strength. Additionally, having multiple algorithms created opportunities for ensemble techniques, where a combination of predictions from several different models improves the system's accuracy.

To improve the model's robustness and to fix the class imbalance issues that may arise, data augmentation techniques will be employed. For medical images, techniques such as rotation, flipping, cropping, scaling, and noise addition will be used to enhance the dataset. For the text-based data, techniques such as synonym replacement, paraphrasing, and random insertion will be applied. This will increase the number of samples in the data while maintaining their meaning. These techniques not only provide balance to the dataset but also expose it to a wider range of variations and reduce the risk of overfitting.

The system evaluation using multiple metrics is important to ensure its' reliability across all categories. At least 3 metrics will be used to provide a comprehensive assessment. Accuracy is the most common metric; it measures the number of correctly classified instances against the total. However, accuracy alone can be misleading, which is why more metrics are needed. Precision and

Recall will be used to try and minimize the number of False Positives while ensuring the model can identify all cases. By combining the 2 above, the F1-score provides a balanced measure. All these metrics provide a well-rounded evaluation of the system's diagnosis abilities.

1.3 List of Design Constraints

1- The whole system shall not cost more than 200JDs.

1.4 List of Engineering Standards

This framework includes strong design requirements to ensure applicability. These include support for various medical image formats such as DICOM as well as PNG, and JPEG images. Additionally, compliance with privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) ensures data protection and patient confidentiality.

Adopting the ISO/IEC TR 24029-1:2021 [15] is essential to ensure the reliability of the AI model in medical diagnosis. This standard provides guidelines for fairness, transparency, and robustness, which are critical in healthcare applications where accuracy is crucial. By following these guidelines, the framework can handle diverse scenarios and deliver consistent performance across varying conditions.

1.5 Load Distribution

The tasks were distributed among the members equally. Firstly, each member wrote four literature review papers. Yazan cleaned the data from garbage and special characters, as well as moved the location category from the descriptions file into the Case Topic file. Omar Abukhalaf handled the categorization of the data as well as balancing the data. Omar Tubeileh handled the data image augmentation. All three members were involved in the design and implementation of the model. Lastly, each person wrote the documentation related to his work. Table [1] below provides a summary.

Table 1: Work distribution

Task	Task Distribution (%)			Level of Completion
	Omar Abukhalaf	Yazan Alsharif	Omar Tubeileh	
Literature Review	33.3%	33.3%	33.3%	100%
Data Cleaning	20%	20%	60%	100%
Data Balancing	60%	20%	20%	100%
Data Image Augmentation	20%	20%	60%	100%
Model Design	33.3%	33.3%	33.3%	100%
Documentation	33.3%	33.3%	33.3%	100%

1.6 Design Overview

Figure 1 below is a high-level diagram that shows the overall flow of the project.

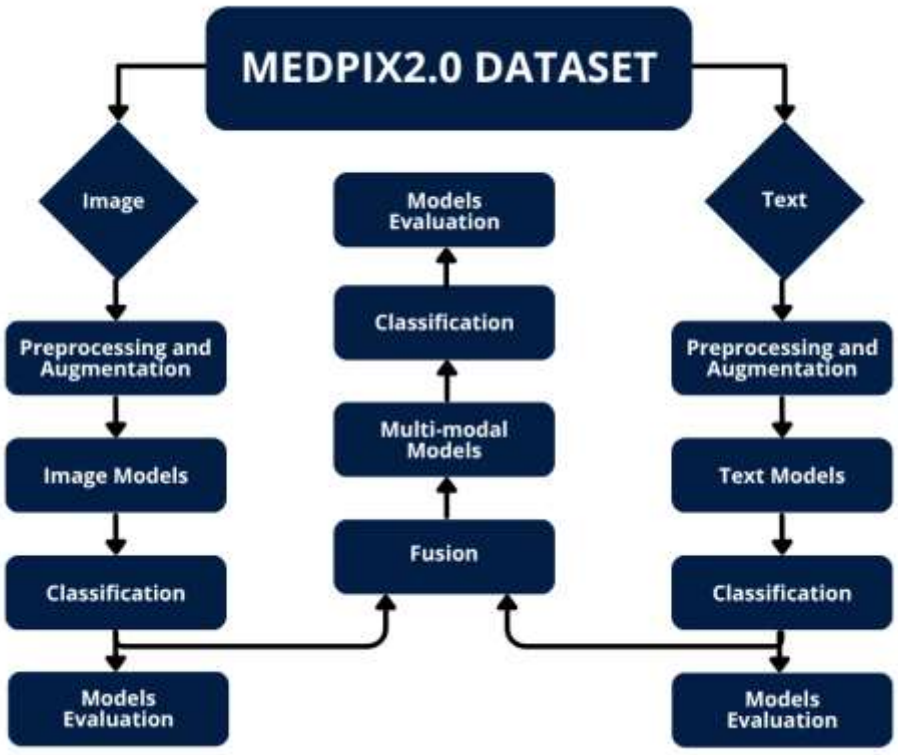

Figure 1: Design Overview

1.7 Document Organization

The document is divided into five main chapters, each addressing a specific aspect of the project.

Chapter 1: Introduction

This chapter goes over the key motivation, main objectives, and core design requirements of the project. It highlights the importance of multi-modal AI in the healthcare industry and the integration of various data modalities to increase diagnosis accuracy. The chapter details the constraints, design requirements, engineering standards, the system overview, and how the tasks will be divided.

Chapter 2: Background and Literature Review

This chapter reviews a wide range of recent studies and methodologies, highlighting the current capabilities, limitations, and trends in the healthcare field. It presents key terminologies and the basics of multi-modal systems, including how to integrate different data sources such as medical images (CT scans, MRI, X-rays) and medical text to improve diagnostic accuracy.

Chapter 3: Design

This chapter outlines the overall design choices and methodologies implemented. The Medpix 2.0 dataset is described, detailing its multi-modal structure compromising both medical images and medical text. The image and text augmentation techniques used to enhance the quality and diversity of the data are discussed. The architecture of the various models is covered, explaining each design choice and fusion technique utilized.

Chapter 4: Results

This chapter presents the performance results of all implemented models, evaluated on the task of disease diagnosis using multimodal data. Each model's output is assessed using a range of standard classification metrics, including accuracy, precision, recall, F1-score, specificity, Area Under the Curve (AUC), and False Negative Rate. A detailed comparison of the models is provided to highlight the strengths and weaknesses of different design choices and fusion strategies.

Chapter 5: Conclusion and Future Work

This chapter summarizes the key findings and conclusions drawn from the project, reflecting on the performance of the implemented models and the effectiveness of integrating image and text data to improve diagnostic accuracy**.** The chapter concludes with future enhancements, including integrating the existing models into a real-world setting.

2 Background and Literature Review

This chapter is a summarization of previous research and models done in the field of medical diagnoses and multi-modal models in general. This work will be compared to our work.

2.1 Literature Review

Many studies in the literature have explored various types of information and data sources to support the diagnosis of medical conditions. For example, in [2], a multi-modal approach was used to predict the severity of COVID-19 in patients. The used dataset included CT scans and clinical features such as patient demographics, history, and lab examination results. This dataset included 733 patients with varying COVID-19 severity levels. In the data preprocessing stage, the CT images were converted to Greyscale for noise reduction and computational efficiency. Then cropped to fit the lung region and eliminate background interference using OpenCV.

To evaluate the patient's severity, Attention-based Convolutional Neural Network (CNN) was the chosen model used to process the CT scan images along with pre-trained parameters. This is due to the relatively small dataset, so using a pre-trained network parameter was necessary to reduce overfitting. For the clinical data, a High-order Factorization Network (HoFN) is used. HoFN exploits high-order interactions of features explicitly. For the Fusion technique, IMSLoss was used to help the model learn modality-shared features.

As for the evaluation metrics, the results obtained were compared to 10 other models, such as Linear Regression, Support Vector Machines, Decision Trees, Random Forests, etc. The multi-modal approach achieved an accuracy of 96% with IMSLoss and 93% without the IMSLoss. Single model approaches accuracies in the range of 70%-75%.

In [3], deep learning and multi-modal fusion were used in order to detect cardiovascular diseases. Electrocardiogram (ECG) and phonocardiogram (PCG) are widely used in CVD for early prevention and detection. Most researchers treat each signal separately and as a reference for the other. Very few explore the interrelationship between them. The dataset was imbalanced and there were a few samples only.

Data augmentation was applied, and each signal was split into 8s with sliding window segmentation. The length of the sliding window of the normal and abnormal samples were 3s and 8s, respectively. Additionally, all samples were then resampled to 500Hz.

Dual-scale deep residual network (DDR-Net) along with a Bi-LSTM layer was used to extract raw features from the ECG and PCG signals. Then, DDR-Net is separately used on the resampled ECG and PCG signals. This will allow the model to capture multi-scale features found in the data. Recursive Feature Selection (RFS) was used until all features were traversed and it automatically deletes the worst features. SVM was used as the evaluator for RFS. Cross-validation was applied to reduce the bias and improve the generalization performance.

The proposed multi-modal automatically selected 25 features, 12 from the ECG and 13 from the PCG, from a total of 128 features. This multi-modal approach which used both ECG and PCG signals, yielded an f1-score of 0.915, an accuracy of 91.6%, which was better than using any of them separately. Indicating the complementary role of ECG and PCG signals in CVD detection.

The authors of [4] propose a machine learning model for diagnosing coronary artery disease (CAD), a common cardiovascular disease (CVD), using Raman Spectroscopy (RS). RS is an optical detection technique that quickly captures unique molecular fingerprints with just a tiny blood sample, making it efficient and minimally invasive. To process the spectral data, the study utilizes the Gramian Angular Field (GAF) algorithm, which transforms RS data into images (referred to as Raman Gramian Angular Field images), enabling image-based analysis. These images are then fed into a multiscale convolutional neural network (CNN) feature extractor, which mines critical features necessary for classifying CAD and its subtypes.

For improved accuracy in subtype classification, the model incorporates a fusion strategy combining RS data with patients' medical histories. Specifically, the authors create a probability matrix that quantifies the influence of various medical conditions—such as percutaneous coronary intervention (PCI), essential hypertension (EH), diabetes mellitus (DM), acute cerebral infarct (ACI), and smoking—on each CVD subtype. Alongside this, a custom-designed weight matrix balances preliminary diagnostic predictions from RS images with medical history information, allowing for a nuanced approach that considers both image-based and historical data in diagnosis.

To validate their model, the authors compare its performance against thirteen other algorithms, including conventional machine learning techniques (like PCA-LDA, PCA-RF, and PCA-SVM) and deep learning methods (like ANN, CNN, and LSTM). These baseline methods

are implemented using public code from the sci-kit-learn library or reproduced from GitHub. Evaluation metrics for the test include accuracy, precision, recall, specificity, and F1 score, which collectively confirm the efficacy of the proposed model in diagnosing CAD with high reliability and accuracy.

In [5], a deep learning model is used to classify lung cancer. Multi-modal Fusion Deep Neural Network (MFDNN) architecture design effectively integrates information from different modalities (e.g., medical imaging, genomics, clinical data) to enhance lung cancer diagnostic accuracy, achieves an accuracy rate of 92.5 %, precision: 87.4 % accuracy in predicting cancer cases, recall: 86.4 % of actual cancerous cases and F1-score: 86.2. The performance is compared with established methods like CNN, DNN, and ResNet.

The used datasets were the Cancer Imaging Archive (TCIA) and The Cancer Genome Atlas (TCGA). For data preparation, missing values are imputed using techniques like mean imputation or removing samples with significant missing information. Outliers are removed using statistical methods. For medical images, it's essential to normalize pixel values to a common scale (e.g., 0 to 1) to ensure consistency across different imaging modalities. Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) are used to reduce noise and redundancy. Categorical variables like cancer type or clinical stage are encoded using techniques like one-hot encoding.

After extracting features from lung cancer images at different scales using convolutional neural networks (CNNs), the proposed network combines these features through attention layers, gating mechanisms, or simple concatenation. The model's weights are updated through backpropagation and optimization techniques (e.g., stochastic gradient descent) to minimize the loss function. At the end of the network, there is a classification layer, typically a fully connected layer with a softmax activation function. (The class with the highest probability is selected)
MFDNN (92.5 %) surpasses all contenders, edging ahead of CNN (91.2 %), DNN (90 %), and ResNet (89.8%)

The authors of [6] focused on a novel method for Alzheimer's disease (AD) diagnosis that leverages the power of multi-modal feature fusion. Recognizing that clinicians rely on various sources of information when diagnosing AD, the authors address the limitations of existing computer-aided diagnostic models that primarily focus on neuroimaging data. The study introduces a model that incorporates neuroimaging data from MRI scans, gene sequence data

(APOE), personal attributes (age, gender, weight), and clinical mental state scale data (MMSE and CDR). This multi-modal approach aims to provide a more comprehensive representation of the patient's condition, enabling more accurate and informed diagnostic decisions.

A key innovation of the proposed model lies in its handling of the inherent dimensional differences between various data sources. The authors introduce a feature transformation method based on geometric algebra to bridge the gap between low-dimensional clinical and biological data and high-dimensional neuroimaging features. This method extends feature vectors from Euclidean space to a higher-dimensional geometric, algebraic space, capturing more complex relationships within the data. After extending feature dimensions, a feature filtration algorithm based on statistical hypothesis testing identifies and removes redundant or insignificant features, ensuring that only relevant information contributes to the diagnosis. The model utilizes a Convolutional Neural Network (CNN) to extract features from MRI images and an Artificial Neural Network (ANN) to fuse these features with the transformed clinical and biological data. The model was evaluated on the ADNI dataset, achieving impressive results: 96.2% accuracy for AD vs. NC and 87.4% for MCI vs. MCI. The authors note that incorporating multi-modal features also leads to faster model convergence. The study highlights the potential of multi-modal feature fusion to significantly improve the accuracy and efficiency of AD diagnosis.

In [7], the authors explore the use of deep learning and multi-modal machine learning (ML) for medical image analysis and clinical decision support systems. Traditional machine learning models in biomedicine mostly focused on unimodal data (e.g., medical images, genetic data, or clinical records). Still, recent advancements aim to integrate multiple types of data to improve predictive performance and diagnostics. Multi-modal data fusion holds the promise of a more holistic approach to patient diagnosis, capturing relationships that might be missed when each modality is treated separately.

The paper identifies five primary challenges in multi-modal ML: representation, fusion, translation, alignment, and co-learning. Representation refers to how data from different modalities (e.g., MRI, CT scans, EHR data) are mathematically encoded for machine learning models. In this regard, deep learning techniques, such as convolutional neural networks (CNNs), have been applied to create rich representations of medical images. Fusion deals with combining these representations from multiple modalities into a unified model, ensuring that the system can handle the inherent differences between modalities. For example, data from imaging and lab tests

may be combined to improve diagnostic accuracy for conditions like cancer or cardiovascular disease.

A key method to enhance data fusion is the use of model-based approaches, where custom networks are designed to process different types of data (e.g., images and EHRs) before combining them for decision-making. In contrast, model-agnostic methods involve post-processing fusion, where outputs from various unimodal models are combined. Recursive feature selection (RFS) and cross-validation techniques are often used to reduce model bias and improve generalizability by selecting the most informative features and ensuring robust performance across different datasets.

The paper also emphasizes the challenge of data alignment, particularly when data from different modalities (e.g., temporal data from EHRs and spatial data from MRIs) need to be synchronized for more effective predictions. Methods such as transformers and attention mechanisms have been successfully applied to this issue, allowing models to learn how to focus on the most relevant features across modalities.

One application of this multi-modal approach discussed is in clinical decision support systems (CDSS), where AI models integrate patient images, medical records, and diagnostic data to assist healthcare providers in making informed decisions. By combining both imaging and non-imaging data, these systems can provide more accurate and personalized treatment recommendations.

The review concludes with the need for continuous research in multi-modal ML, especially addressing challenges like data bias, scarcity of large datasets, and the integration of multi-modal data in real-world clinical environments. By refining these methods and ensuring practical implementation, multi-modal ML is expected to play a transformative role in clinical biomedicine, improving diagnostic accuracy and patient outcomes.

Paper [8] talks about how integrating different data modalities enhances the precision and reliability of disease diagnosis. The dataset used contained chest X-rays and clinical data for 65000 patients; however, the paper only took a sample of 289 patients to evaluate the multi-modal approach. The clinical data included blood tests, urine, fluids, and other chemical tests. Each patient was either labeled as a healthy patient or a sick patient.

Since the X-rays were performed on different equipment, they had different resolutions, coloring, and image sizes. Images were standardized and resized into 225x225 and 64*64 dimensions. The 64*64 dimensions were used as the 225x225 turned out to be quite expensive.

Data augmentation was applied using the following methods. Rotating the images to the side(right to left), adding 20% noise, and converting the images to monochrome. For the clinical data, the number of features was reduced from 1630 to 521 by removing features that only existed in a few patients. Next, augmentation was applied to generate the same number of X-rays using WGAN, MDCLone, and other noise autoencoders.

Two main architectures were proposed: Late fusion and Intermediate Fusion. Late Fusion involves training 2 separate models and then taking the results of both and combining them using techniques such as a soft-max function. The drawback is that in late fusion, it doesn't learn the correlation between different modalities. Intermediate fusion transforms the different input data into higher-level representations, and this allows the model to learn from mappings between the data.

For the late fusion, 3 pre-trained Neural Networks were used on the images: DenseNet121, DenseNet169, and ResNet50. For the clinical data, Long Short-Term Memory (LSTM) and Attention models were used. For the intermediate fusion, 5 layers were used to process the images, and SVM's were used to align the features of different modalities. Adam's optimizer was used to implement the learning rate. In both architectures, adaptive batch sizes were implemented.

The results yielded without the adaptive batch sizes showed that intermediate fusion achieved the highest F1-Score of 94.24% and a lower loss of 1.52 compared to Late Fusion of 91.07% and 3.71, respectively. The results were similar when the adaptive batch size was used with intermediate fusion, obtaining better results with an F1-Score of 93.11% compared to 91.67% and a higher accuracy of 93.15% compared to 89.99%.

The work in [9] proposes a disease classification model based on multi-modal fusion, using both chest X-ray images along text descriptions to enhance performance. The dataset used is a subset of the Chest X-ray dataset from OpenI, and it contains originally 3999 samples, which was then reduced to 800 samples after cleaning the data. To balance the dataset Synthetic Minority Oversampling Technique(SMOTE) was applied to provide balance to the dataset and increased the sample size to 3754.

Image preprocessing included normalization of the image data and cleaning the textual data by removing unclear text and placeholders. Dimensionality reduction was applied via Principal Component Analysis (PCA). Text features were extracted using the BERT- BERT-based model,

capturing important information, while for the images, features were obtained by using the ResNet-50 convolutional neural network.

The architecture used integrates a multi-modal attention mechanism for feature fusion, which calculates correlation weights between feature vectors and adaptively fuses them based on their relevance. This is superior to traditional vector concatenation methods. Then, a multi-layer perceptron (MLP) served as the classifier and achieved better results compared to other alternatives such as logistic regression, decision trees, and random forests.

The results showed that the proposed architecture achieved a maximum accuracy of 97.72% on the validation set. The multi-modal approach improved accuracy by 0.55% compared to image-only models and 2.69% compared to text models. These results highlight the effectiveness of multi-modal feature integration and the importance of balancing the dataset to improve results.

In this paper [10], a comprehensive survey of multi-modal machine learning (ML) is presented, emphasizing the integration of data from diverse modalities such as language, vision, and audio. The authors propose a taxonomy for the key technical challenges in MMML, including representation, translation, alignment, fusion, and co-learning. The dataset specifics are broad and span various multi-modal applications, such as audio-visual speech recognition, multimedia retrieval, and emotion recognition.

In the data preprocessing stage, multi-modal data often undergo feature extraction tailored to each modality, followed by normalization to reduce modality-specific noise. For example, image features may be extracted using pre-trained Convolutional Neural Networks (CNNs), while text features may rely on word embeddings such as Word2Vec or GloVe.

To address the heterogeneity of modalities, representation learning was identified as a foundational challenge, focusing on either joint or coordinated representations. Additionally, translation tasks, such as image captioning and text-to-image generation, were categorized into generative and example-based approaches. Alignment techniques, vital for establishing correspondences between modalities, included supervised and unsupervised methods.

The fusion process combines multi-modal features for downstream tasks like classification or regression, leveraging early, late, or hybrid strategies. To enhance learning in scenarios with limited data, co-learning was highlighted as a promising method, exploiting shared knowledge across modalities.

Evaluation metrics compared MMML approaches across multiple domains, demonstrating significant improvements in tasks where multi-modal data was effectively utilized. The survey underscores the potential of MMML in bridging gaps between modalities and advancing fields like healthcare, multimedia retrieval, and affective computing.

This study [11] investigates the predictive capabilities of integrating multi-modal data—radiology, pathology, and genomics—for determining response to PD-(L)1 blockade in patients with advanced non-small cell lung cancer (NSCLC). The cohort comprised 247 patients who underwent baseline assessments, including CT imaging, PD-L1 immunohistochemistry, and genomic sequencing. This multi-modal data was processed using machine-learning workflows to develop a deep attention-based multiple-instance learning model with masking (DyAM), which computes patient-specific risk scores.

In the data preprocessing stage, CT images underwent segmentation by expert thoracic radiologists to isolate lesions, and features were extracted using radionics. Similarly, PD-L1 IHC slides were digitized, and textural features of protein expression patterns were quantified. Genomic data were obtained using the MSK-IMPACT sequencing platform, targeting alterations associated with NSCLC outcomes.

Modeling efforts demonstrated that the multi-modal DyAM model significantly outperformed unimodal methods in predicting patient response to immunotherapy, with an AUC of 0.80 for multi-modal data compared to 0.61-0.73 for individual modalities (e.g., CT imaging, PD-L1 IHC scores). The model's attention mechanism allowed for modality-specific weight adjustments, making it robust to missing data.

The study underscores the power of combining multi-modal diagnostic data to improve precision oncology practices, providing a rationale for using machine-learning-driven integration to enhance response prediction. However, limitations such as the single-center cohort and the need for external validation were acknowledged, pointing to the potential for further research with larger, multi-institutional datasets.

This paper [12] presents a multi-modal deep learning model that integrates chest CT-scan and X-ray for diagnosing COVID-19 Pneumonia, which impacts the respiratory system. Previous research has primarily focused on single modalities, often achieving limited diagnostic accuracy. This study aims to utilize multi-modal learning by combining CT-scan and X-ray data through transfer learning models, thus harnessing complementary features for improved diagnosis.

The study used open-source datasets comprising 2,500 CT-scan images and 2,500 X-ray images (balanced between COVID-19 pneumonia and normal cases). Images were resized to 150x150 pixels, normalized, and divided into training and validation sets.

Transfer learning models (e.g., ResNet50, VGG16, DenseNet121, MobileNet, Xception, InceptionV3) pre-trained on ImageNet were employed. A multi-modal network was constructed by concatenating features extracted from CT-scan and X-ray models. Separate single-modality models were also trained for comparison. Accuracy, sensitivity, and specificity were used to evaluate performance. Computational efficiency was analyzed across different network combinations.

Multi-modal Performance: The concatenated DenseNet121-MobileNet and ResNet50-VGG16 models achieved the highest accuracy (99.87%), sensitivity (99.74%), and specificity (100%). Multi-modal models outperformed single-modality approaches in all metrics. Limitations of this study are the use of relatively small datasets, limiting generalizability, also reliance on only two imaging modalities without exploring additional biomarkers like clinical data or laboratory results, which is better in our project, which integrates text data with the 2 images.

The paper [13] talks about a multi-modal deep learning model (ResBioBERT) used for early diagnosis of Heart Failure (HF). It was created to enhance heart failure diagnosis by integrating clinical texts and chest X-ray (CXR) imaging. The dataset used is the MIMIC-CXR dataset comprising 377,110 CXR images and 227,835 radiology reports, which were filtered and divided into 12,896 paired samples for training, 1,612 pairs for testing, and 1,612 for validation.

The model consists of two models: ResNet-152 for image feature extraction and BioBERT for clinical text features derived from radiology reports and patient medical history. ResNet was adapted to extract 2048-dimensional features from CXR images, while BioBERT fine-tuned biomedical text and encoded clinical text into 768-dimensional feature vectors. Dense multi-modal features were mapped into a shared embedding space using a single-stream network, and the fused features were classified using a fully connected layer and a SoftMax activation function.

The ResBioBERT model was compared to uni-modal (CXR-only or text-only) and competing multi-modal models (like MultiCOVID). It outperformed both by achieving 0.9598 AUROC, 0.9438 AUPRC, 89.02% Accuracy, and 88.113% F1 score. Uni-modal models like ResNet-50 and Longformer achieved accuracies of 84.44% and 85.59%, respectively, highlighting

their limitations. The MultiCOVID multi-modal model achieved 0.913 AUROC, while ResBioBERT significantly improved this to 0.9598.

The limitations of this model are that it relies on specific datasets, necessitating validation of external datasets for generalizability. Our proposed model helps in generalizability by including a second type of image, which can help in validation.

2.2 Summary

Table [2] below summarizes the results of all of the above studies and compares them with our results.

Table 2: Comparison between the reviewed papers and our methodology

Paper	Task	Modalities Used	Key Models	Results
[2]	Predict COVID-19 severity	CT scans, Clinical data	Attention CNN, HoFN, and IMSLoss for fusion	96% accuracy (with IMSLoss); 93% (without)
[3]	Detect cardiovascular diseases	ECG, PCG	DDR-Net, Bi-LSTM, RFS	91.6% accuracy; F1-score 0.915
[4]	Diagnose coronary artery disease (CAD)	Raman Spectroscopy, Medical history	GAF algorithm, Multiscale CNN, Weighted fusion	High reliability and accuracy (various metrics)
[5]	Classify lung cancer	Medical imaging, Genomics, Clinical data	MFDNN, PCA, t-SNE	92.5% accuracy; Precision 87.4%, Recall 86.4%, F1-score 86.2%
[6]	Diagnose Alzheimer's disease (AD)	MRI, Gene sequences, Clinical data	Geometric algebra, Feature filtration, CNN-ANN fusion	96.2% accuracy (AD vs NC); 87.4% (sMCI vs MCI)
[7]	Medical image analysis and decision support	Various (images, EHRs, etc.)	Recursive Feature Selection, Transformers, Attention mechanisms	Improved diagnostic accuracy and generalizability
[8]	Enhance disease diagnosis	Chest X-rays, Clinical data	Late and Intermediate fusion, Pre-trained models	F1-score 94.24% (Intermediate fusion); 91.07% (Late fusion)
[9]	Disease classification	Chest X-rays, Text descriptions	BERT, ResNet-50, Multi-modal attention mechanism	97.72% accuracy; Improved over single modalities
[10]	Survey of ML applications	Language, Vision, Audio	Representation, Fusion, Alignment	Significant improvements across various tasks
[11]	Predict response to immunotherapy in NSCLC	CT imaging, PD-L1, Genomics	DyAM model, Radiomics, Attention mechanism	AUC 0.80 (multi-modal); 0.61-0.73 (unimodal)

[12]	Diagnose COVID-19 pneumonia	CT scans, X-rays	Transfer learning, Multi-modal networks	99.87% accuracy; Sensitivity 99.74%, Specificity 100%
[13]	Early diagnosis of Heart Failure (HF)	Chest X-rays, Clinical texts	ResNet-152, BioBERT, Shared embedding space	89.02% accuracy; F1-score 88.11%; AUROC 0.9598

3 Design

This chapter talks about the dataset we will be using to create our multi-modal model; it contains the description, shape, and distribution of the data. Also, data pre-processing techniques like data cleaning and augmentation are discussed in depth.

3.1 Analysis of Design Requirements

1- A minimum of two ML Algorithms shall be implemented and rigorously evaluated on performance.

2- The system shall utilize a dataset that includes both medical imaging data and corresponding text-based clinical reports such as (such as patient symptoms, patient history, and lab results.

3- The system shall process medical images from various sources, such as MRI, CT, and X-ray images.

4- The system shall evaluate performance using a minimum of four metrics

5- The system shall ensure all data is handled according to privacy regulations, such as The Health Insurance Portability and Accountability Act (HIPAA) for medical data protection, ensuring patient confidentiality.

6- The system shall support various medical image formats (e.g., DICOM, PNG, JPG) and accommodate variations in lighting, resolution, and overall image quality during the preprocessing stage.

7- Data augmentation shall be used to improve the balance of the dataset and increase the model's robustness.

The system requires the implementation of at least two machine learning algorithms, which ensures a solid comparison and improves the reliability of the results. This can involve using different models for text and images or testing variations of the same type, such as different CNNs for imaging tasks. Evaluating these models with multiple metrics like accuracy, precision, F1-

score, Specificity, and AUC helps ensure that performance is measured fairly and comprehensively.

The use of a multimodal dataset that includes both medical images and clinical text enhances diagnostic accuracy by combining visual and contextual information. The system's ability to handle images from various sources, MRI, CT, and medical text, demands careful preprocessing due to differences in format and quality.

Supporting multiple image formats like DICOM, PNG, and JPG is essential for real-world applications. Preprocessing must also handle variations in lighting and resolution to ensure consistent input quality. Additionally, data augmentation will play a key role in improving model robustness and addressing class imbalance.

Finally, handling all data in compliance with regulations like HIPAA is critical. This involves anonymizing patient information and ensuring secure storage and processing practices to protect privacy and meet ethical standards in healthcare AI applications.

3.2 Analysis of Design Constraints

The only constraint is that the system cost shouldn't exceed 200 JDs.

3.3 Discussion of Engineering Standards

The inclusion of engineering standards like support for multiple medical image formats (DICOM, PNG, JPEG) ensures the framework can operate in diverse clinical environments. This flexibility is important for interoperability with hospital systems and diagnostic tools. Additionally, enforcing HIPAA compliance guarantees that sensitive patient data is handled securely and ethically, aligning the system with legal and professional standards for medical data privacy.

Adopting ISO/IEC TR 24029-1:2021 further strengthens the framework by embedding principles of fairness, transparency, and robustness into the AI model's design. This is particularly critical in healthcare, where decisions based on AI must be reliable and unbiased. Following this

standard helps build trust in the system's outputs and ensures it can perform consistently across different patient groups and clinical scenarios.

3.4 Dataset

The data used was the Medpix 2.0 dataset from [1]. It consists of 671 patients whose name is anonymous. Each case is documented to include important patient information along with the medical images. Each patient has a unique ID (U_Id) ensuring traceability within the dataset. It also includes 2 types of medical imaging, MRIs and CT scans, where each patient has at least 1 type of image. The dataset is made up of 3 main files: Case Topic, Descriptions, and the images file.

The case topic file includes all the information about the patient's case. It is split into 2 parts: the Case and the topic parts. The case part includes all the information relevant to the case, such as title, history, findings, exam, case diagnosis, and diagnosis. The topic part includes general information about the disease itself, such as title, disease discussion, ACR Code, and category.

The description file includes a detailed description of every single image in the medical images file. It includes the U_Id, which is the patient identifier, the type of the image (CT or MRI), the image name, the location (micro location), location category (general location), age, caption, figure part, Modality scanning, the plane of the image.

It includes text about the patient's history, clinical findings, ACR code, and the category fields. It also includes the names of the medical images belonging to that patient in the medical images file. Figure 2 below provides a summary of the contents of each file.

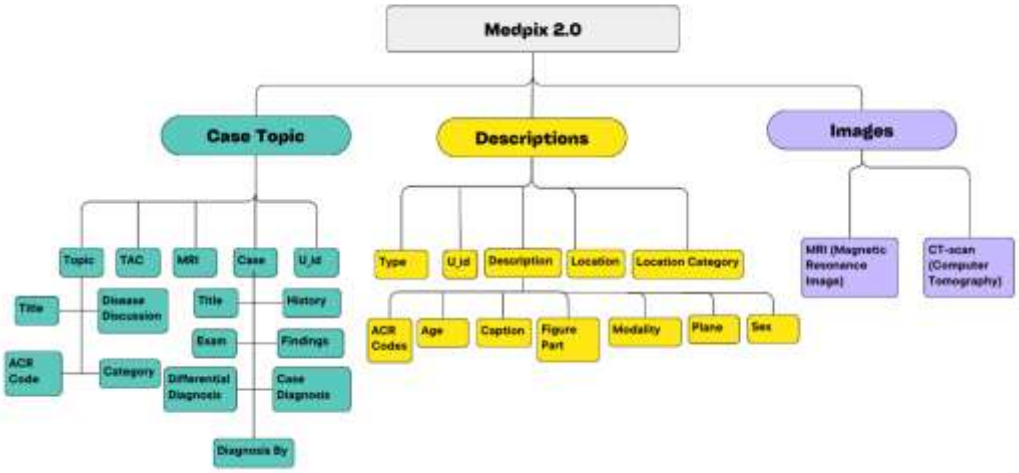

Figure 2: Dataset Shape

3.5 Data Preprocessing

The original dataset did not contain explicit diagnostic labels, making it necessary to assign labels based on existing metadata. We identified the "category" field as the most reliable source of diagnostic context and used it to group and label the data into 12 distinct diagnostic categories. To ensure clarity and consistency in labeling, we leveraged ChatGPT to assist in interpreting ambiguous or unclear category names. Once finalized, these labels were encoded for use in downstream machine-learning tasks.

Table 3: The final categories, along with the number of records

Category	Number of Records	Encoding
Congenital & Genetic	88	0
Trauma & Physical Injuries	59	1
Vascular & Circulatory	45	2
Infections	41	3
Neoplasm - Benign & Sarcoma	57	4
Neoplasm - Carcinoma	36	5
Neoplasm - Other Malignant	101	6
Inflammatory & Autoimmune	24	7
Metabolic & Endocrine	17	8
Cysts & Degenerative Conditions	16	9
Obstruction & Structural Abnormalities	15	10
Miscellaneous Conditions	172	11

Each record in the dataset was assigned to one of these categories based on its diagnosis label. This grouping facilitated targeted analysis and enabled the implementation of augmentation strategies for underrepresented classes.

An initial analysis revealed significant disparities in the number of records across the categories, ranging from as few as 15 to over 170 records per category, as shown in Table [3]. So, based on the number of records, several balancing techniques were implemented using both text and image augmentations. The number of augmentations to be done on each category is shown in Tables [4] and [5]. Rather than augmenting all classes uniformly, augmentation was applied based on the severity of the imbalance. For the image augmentation, classes with less than 40 records were augmented twice, while the classes with more than 40 were only augmented once. The goal is to have approximately 200 records per class after the image and text augmentations.

For each augmented image, the corresponding metadata (textual descriptions) was duplicated without modification to maintain the consistency of multimodal records, and the same for the augmented text in which the corresponding images were duplicated. Additionally, each new augmented sample was uniquely labeled to avoid duplication conflicts during training.

Table 4: Image augmentations

Category	Records Before Augmentation	Number of Augmentations	Final Number of Records
Congenital & Genetic	88	0	88
Trauma & Physical Injuries	59	1	118
Infections	45	1	82
Vascular & Circulatory	41	1	90
Neoplasm - Benign & Sarcoma	57	1	114
Neoplasm - Carcinoma	36	2	72
Neoplasm - Other Malignant	101	1	202
Inflammatory & Autoimmune	24	2	72
Metabolic & Endocrine	17	2	51
Cysts & Degenerative Conditions	16	2	48
Obstruction & Structural Abnormalities	15	2	45
Miscellaneous Conditions	172	0	172

The number of image augmentations in Table [4] was determined based on the number of records in each class. To balance the dataset, we aimed for approximately 200 samples per

category. Some classes already had sufficient data and did not require any augmentation, while others required one or two augmented texts per original sample to reach the target count. This class-wise augmentation strategy helped ensure a more balanced and representative dataset for model training.

Table 5: Text augmentations

Category	Records Before Augmentation	Number of Augmentations	Final Number of Records
Congenital & Genetic	88	1	176
Trauma & Physical Injuries	118	1	236
Infections	82	1	180
Vascular & Circulatory	90	1	164
Neoplasm - Benign & Sarcoma	114	1	228
Neoplasm - Carcinoma	72	2	216
Neoplasm - Other Malignant	202	0	202
Inflammatory & Autoimmune	72	2	216
Metabolic & Endocrine	51	3	204
Cysts & Degenerative Conditions	48	3	192
Obstruction & Structural Abnormalities	45	3	180
Miscellaneous Conditions	172	0	172

The number of text augmentations in Table [5] was determined based on the number of records after image augmentation, such that the number of patients for each class after balancing was in the range 164-236, which is almost balanced.

Text augmentation

The following are the 3 techniques that were employed:

- Synonym Replacements: The first text technique used was Synonym replacement, where we changed words with their synonyms to provide some balance to the dataset. Some libraries, such as TextAttack and NLPAug, were used but didn't offer promising results as they changed the meaning behind the sentences drastically. Instead, we manually picked widely used words and phrases found in the dataset and mapped them to their corresponding synonym to perform the balancing. Words such as "pain" were replaced

with "ache," "operation," "surgery," "swelling," "inflammation," and so on. Additionally, the patient's age was changed by adding or removing 1-2 years. This technique was able to provide meaningful variations of data while preserving the context.

- Sentence Re-ordering: This technique involves changing the sequence of sentences within a text while maintaining the original meaning and logical flow. Fields that only included 1 sentence meant that re-ordering couldn't be used, so another technique was used, which was Active-Passive voice conversion. These sentences were rephrased to ensure the structure of it was changed.
- Back Translation: This is a data augmentation technique used to improve diversity. This involved translating the text into French and then translating it back to English. This offered a different variance of the original text while maintaining the medical meaning behind it. "Googletrans" was used, which is a Python library that provides an interface to the Google Translate API.

These methods were applied selectively to the underrepresented categories, increasing their sample sizes while retaining the original topic data and the images.

Image Augmentations:

The following are the two techniques that were implemented to balance the data.

- Contrast Adjustment: The first image augmentation is implemented using the PIL.ImageEnhance module. It increases the image's contrast by a factor of 1.5 (50%), making dark areas darker and bright areas brighter. This is particularly useful for enhancing the visibility of subtle features in medical images, making certain visual features more prominent, where differences in tissue density or pathology can be highlighted with better contrast. Applying contrast variation during training improves the model's robustness to images captured under varying lighting conditions or device settings, ensuring it doesn't overly rely on consistent illumination to make predictions.
- Grid Distortion: it is applied using albumentations' GridDistortion class. This method divides the image into a grid and applies random distortions to each section, simulating non-linear warping. Such distortions mimic real-world imaging artifacts, patient

movement, or tissue deformations, encouraging the model to generalize better to spatially inconsistent inputs. By introducing controlled geometric variability, this augmentation helps the model become less sensitive to exact spatial arrangements, promoting a focus on structural patterns rather than precise pixel locations.

To further enhance the dataset and ensure robust model training, image augmentation was applied to the CT/MRI images associated with the dataset, as shown in Figures 2-6. The following transformations were implemented:

- Clockwise rotation: Images were rotated 30 degrees to the right.
- Anti-clockwise rotation: Images were rotated 30 degrees to the left

The following shows a sample image and the augmented versions of it:

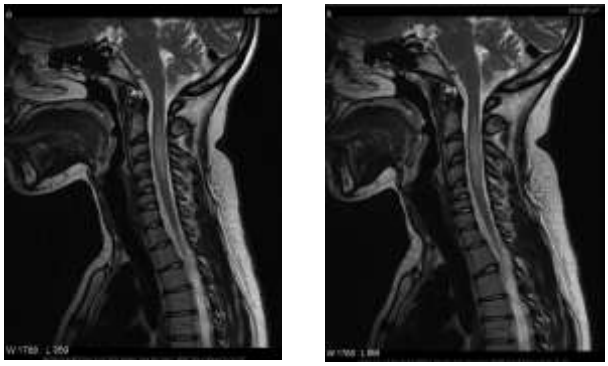

Figure 3: Original image vs grid-distorted

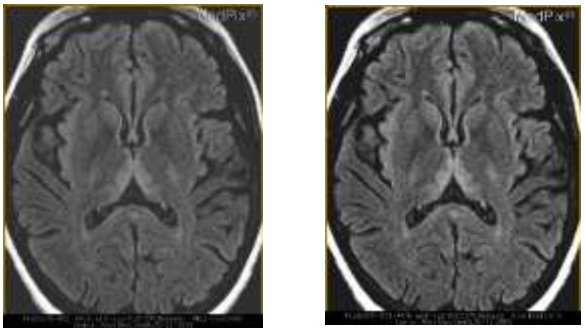

Figure 4: Original image vs contrast-adjusted

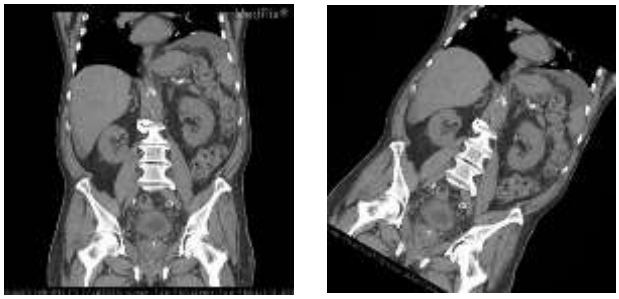

Figure 5: Original image vs rotated 30 degrees clockwise

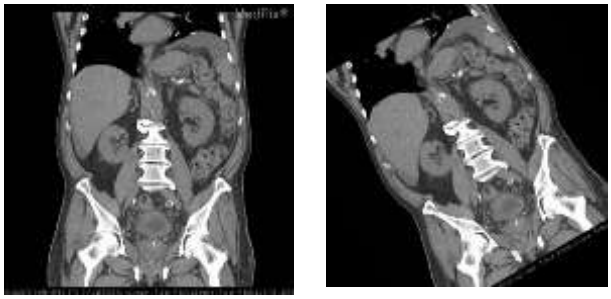

Figure 6: Original image vs rotated 30 degrees anti-clockwise

3.6 Final Data Composition

The final dataset was enriched with augmented text and image data. The balancing efforts ensured a more equitable distribution of records across all categories, while the image augmentations added variability to the dataset. The original topic data was preserved as it did not require augmentation.

The preprocessing phase transformed the MedPix 2.0 dataset into a more balanced and diverse dataset suitable for machine learning. These enhancements addressed the challenges of data imbalance and limited sample size, laying the foundation for the subsequent modeling phase. Table [6] blow shows the final number of records for each class.

Table 6: Final data distribution

Category	Final Number of Records
Congenital & Genetic	528
Trauma & Physical Injuries	708
Infections	492
Vascular & Circulatory	540
Neoplasm - Benign & Sarcoma	684
Neoplasm - Carcinoma	648
Neoplasm - Other Malignant	606
Inflammatory & Autoimmune	648

Metabolic & Endocrine	612
Cysts & Degenerative Conditions	576
Obstruction & Structural Abnormalities	540
Miscellaneous Conditions	516
Total	**7098**

3.7 Design Choices

The design and implementation of the multi-modal disease classification framework involved several critical design decisions across the dataset, model architecture, fusion strategies, and evaluation methodologies. The following summarizes the main options considered and the rationale behind the selected approaches

1. Choice of Dataset

Choosing the dataset proved to be very important in building a reliable multi-modal diagnostic system. Many researchers in medical ML depend on datasets like MIMIC-CXR, OpenI, and CheXpert, but they are exclusively about X-rays and lack adequate clinical background information. MedPix 2.0 is a contrasting choice that provides MRI and CT image pairs and in-depth explanations of each case, including its history, findings, exam, and diagnosis. Additionally, MedPix offers examples from various diseases aside from chest-related conditions, which helps our model be applied in many medical domains. Despite being smaller, MedPix provides more details, different kinds of images, and detailed text, which is why it shines in training a comprehensive multimodal framework. Table [7] below compares different medical datasets.

Table 7: Dataset Comparison

Dataset	Modalities	Text Detail	Image Diversity	Chosen	Reason
MIMIC-CXR	X-ray + reports	Limited	Low (mostly CXR)	No	Focused only on lungs
OpenI	X-ray + abstracts	Medium	Low	No	Fewer paired cases
MedPix 2.0	**CT, MRI + full reports**	**High**	**High**	**Yes**	**Multiple diseases**

2. Image Models

We reviewed popular Convolutional Neural Networks (CNNs) for feature extraction from images and selected ResNet50, DenseNet121, and EfficientNet-B0 because they performed strongly and efficiently. ResNet50 gives a secure starting point that includes residual learning, DenseNet121 cuts down on parameters by reusing features, and EfficientNet-B0 performs at the highest level with the least computing demand. While other models like VGG16/VGG19 are easy to understand and have a history in medical research, they did not make the cut because they have too many parameters and work more slowly than the recent models. The models we selected are modern and fit our needs for resources and performance. Table [8] below compares different image models.

Table 8: Image Models Comparison

Model	Params (M)	Accuracy	Speed	Chosen	Reason
VGG16/VGG19	138/148	High	Very slow	No	Too heavy
ResNet50	25.6	High	Moderate	Yes	Strong baseline
DenseNet121	8	High	Fast	Yes	Efficient, stable gradients
EfficientNet-B0	**5.3**	**Highest**	**Fastest**	**Yes**	**Best accuracy & efficiency**

3. Text Models

The part of our system dealing with text relies on models able to understand medical narratives. Both classic and deep learning were picked to cover a wide range of outputs. Among the classifiers, Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF) were chosen mainly because of their understandable nature and reasonable speed working on TF-IDF vectorized text. Thanks to these classical models, quick testing became possible. The options of DistilBERT and BiomedBERT were considered for deep learning. Even though DistilBERT was efficient, BiomedBERT, which was trained in biomedical information, was better at handling technical terms in medical topics. Because GPT or XLNet is more expensive to train and use, we chose to stay with simpler models. With these two types of AI models, we could analyze many options for text processing. Table [9] below compares different text models.

Table 9: Text Models Comparison

Model	Type	Domain Specific	Performance	Chosen	Reason
Logistic Regression	Classical	No	Moderate	Yes	Lightweight baseline
SVM	Classical	No	High	Yes	Strong on high-dim data
Random Forest	Classical	No	Very High	**Yes**	Best classical performer
DistilBERT	Transformer	Partial	High	No	Lightweight but weaker
GPT/XLNet	Transformer	No	Very High	No	Too large and slow
BiomedBERT	**Transformer**	**Yes**	**Highest**	**Yes**	**Best on clinical text**

4. Fusion Strategies

We looked into each of the three main Fusion ways: Early Fusion, Late Fusion, and Intermediate Fusion. Concatenation of image and text features at the beginning of fusion in the model allowed it to find how images and text work together as soon as possible. Independent models could be used to make predictions, which were then assembled using either averaging or weighted probabilities—this arrangement made the entire approach more manageable and easier to repair. Adding cross-modal attention to the model helped it highlight relevant sections of the text that were influenced by images. Due to increased difficulty and little additional benefit with our limited data, we did not apply hybrid fusion (enlisting strategies from both early and late or multi-stage fusion). More importantly, fusion approaches were judged by how easily they can be understood, how well they work with larger datasets, and whether they are applicable to many tasks. Table [10] below compares different fusion strategies.

Table 10: Fusion Strategy Comparison

Strategy	Fusion Level	Cross-Modality Learning	Complexity	Chosen	Notes
Early Fusion	Feature-level	High	Moderate	Yes	Effective but prone to overfitting
Hybrid Fusion	Multi-stage	Very High	Very High	No	Too complex for dataset size
Intermediate Fusion	Attention	Very High	High	Yes	Best for interaction learning
Late Fusion	**Output-level**	**Low**	**Low**	**Yes**	**Interpretable and modular**

5. Data Augmentation Techniques

Our data augmentation strategy was tailored to address the needs of working with medical data. Clinically safe methods like adjusting contrast, distorting the images, and rotation were used for the images. Transformations like cropping, flipping, zooming, or adding artistic noise (e.g., style transfer or color jitter) are usual in natural image tasks; however, they may blur disease features and lead to bad diagnoses in medical contexts. For clinical text, we used synonym replacement, back translation, and sentence reordering—strategies all to maintain the meaning of the clinical text. To avoid affecting the medical meaning, we didn't use word deletion, random swapping, or sentence generation from language models. We decided on our augmentation techniques to keep the accuracy of diagnoses but also introduce variety to the data. Table [11] below compares augmentation techniques.

Table 11: Augmentation Comparison

Modality	Technique Type	Safe for Medical?	Used	Reason
Image	**Rotation, Contrast, Grid Distortion**	**Yes**	**Yes**	**Preserves anatomy and enhances features**

Image	Zoom, Flip, Crop, Color Jitter	No	No	Can distort critical medical features
Text	**Synonym Replacement, Back Translation, Reordering**	**Yes**	**Yes**	**Maintains clinical meaning**
Text	Word Deletion, Random Swapping, LM-Generated Sentences	No	No	Risk of hallucination or losing intent

6. Fusion Partners for Final Model

After evaluating the performance of individual image and text models, we selected the best-performing fusion pairs. EfficientNet-B0, as the best image model, was combined with BiomedBERT, the best text model, as well as Random Forest, the next best text model. This allowed us to compare fusion strategies under both a deep-learning and classical NLP setting. Interestingly, the combination of EfficientNet and Random Forest using weighted late fusion achieved the best overall performance. This outcome demonstrated that cross-modality pairings between deep and classical models can outperform deep-only architectures, especially when paired with careful fusion strategies. Table [12] below compares text-image paring models used.

Table 12: Fusion Partners Comparison

Image Model	Text Model	Fusion Strategy	Accuracy	Chosen
EfficientNet-B0	BiomedBERT	Early / Intermediate	94–95%	Yes
EfficientNet-B0	**Random Forest**	**Weighted Late Fusion**	**96.62%**	**Yes**

7. Evaluation Metrics

We assessed our models widely by looking at Accuracy, Precision, Recall, F1-Score, AUC-ROC, and False Negative Rate (FNR). Although accuracy is normally intuitive, it may not accurately reflect the outcome for minority classes in the data, so the measures of precision and recall were introduced to examine the performance of these classes. The F1-score gave a measure that balanced the two aspects. AUC-ROC helped evaluate results based on all possible thresholds, and in medicine, missing a case of disease can be very serious, so FNR was essential to highlight. Thus, the wide range of metrics assured that models could be compared and we could tell exactly how their performances differed. Table [13] below showcases the different evaluation metrics used.

Table 13: Evaluation Metrics Used

Metric	Purpose	Used	Reason
Accuracy	Overall correctness	Yes	Basic performance measure
Precision	Minimize false positives	Yes	Relevant for trustworthy results
Recall	Minimize false negatives	Yes	Critical in medical diagnosis
F1-score	Balance between precision and recall	Yes	Handles imbalanced datasets
AUC-ROC	Class separability across thresholds	Yes	Evaluates model robustness
False Negative Rate	Critical error detection	Yes	Essential for detecting missed diagnoses

3.8 Design Development

The design for the models outlines the practical implementation and architecture of the proposed approach, detailing how the system was built to achieve the desired results. It includes the data preprocessing techniques, the models employed for each modality, and the strategy used to integrate text and image features. Particular attention was given to handling multimodal data efficiently, ensuring that both textual and visual information were appropriately processed and encoded before fusion.

3.8.1 Developed Design for Resnet50 and DistilBert model

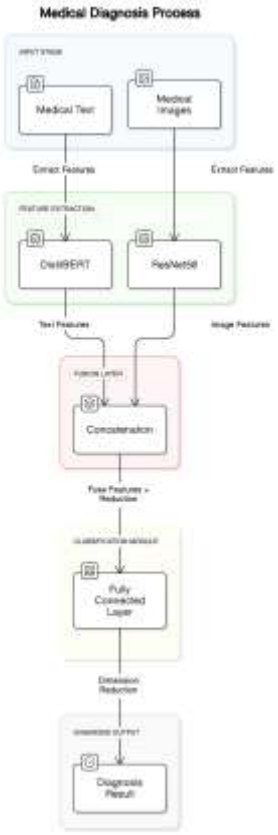

Figure 7: Resnet50 and DistilBert design

The proposed design, as seen in Figure [7], involves integrating features from the medical text along with the medical images. The medical images go into the ResNet50, which is a pre-trained Convolutional Neural Network designed for visual feature extraction. It processes the images (CT or MRI) and outputs a 2048-dimensional feature vector, which is then reduced to 256. At the same time, the medical text utilized the DistilBERT model, which is a transformer-based NLP model, to process text data such as the patient history and type of examinations. The DistilBERT extracts a 768-dimensional feature vector by using the CLS token representation.

After the feature extraction stage, the concatenation fusion layer combines the 256 and 768-dimensional layers from both models into a unified 1024-feature vector and reduces it down to 512. This large vector is then passed into a fully connected layer to learn the patterns and similarities between the 2 modalities. The layer reduces the dimensionality down to 128. Finally, the output layer maps the features to the target classes.

3.8.2 Developed Design for Logistic Regression Model

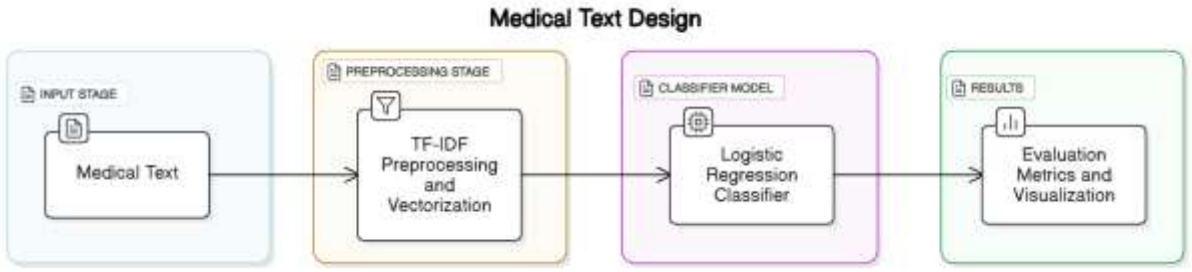

Figure 8: Logistic Regression design

Logistic Regression is a widely used linear model for classification tasks that works by estimating the probability that a given input belongs to a particular class. In the multiclass setting of this project, operates using a one-vs-rest strategy, training separate classifiers for each disease category.

The model pipeline begins by loading and preprocessing clinical data stored in JSON format. For each record, key fields such as the case title, history, findings, location, and differential diagnosis are concatenated to form a comprehensive text input. This text is then transformed using the TF-IDF(Term Frequency-Inverse Document Frequency) vectorizer, which captures the importance of each word across the dataset while reducing the impact of commonly occurring, less informative terms. A limited vocabulary of the top 1000 features is used to balance expressiveness and model efficiency.

3.8.3 Developed Design for Random Forests

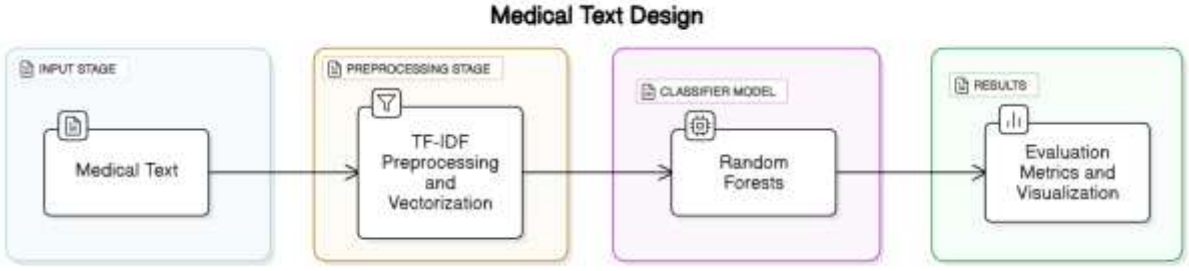

Figure 9: Random Forest design

Random Forests(RF) is a powerful ensemble learning algorithm that builds a collection of decision trees and aggregates their predictions to make a final decision. It is widely used for classification tasks due to its robustness, ability to handle high-dimensional data, and resistance to overfitting. Figure [9] shows the design for the RF model.

3.8.4 Developed Design for Support Vector Machines

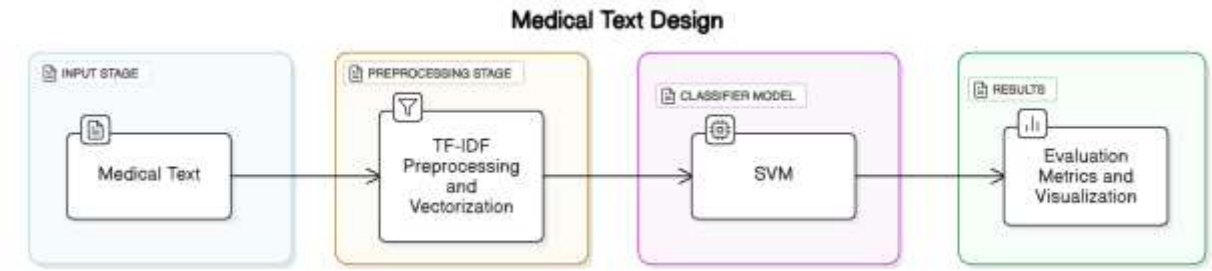

Figure 10: SVM design

The classification pipeline shown in Figure [10] begins with text preprocessing, where relevant fields Title, History, Differential Diagnosis, and Location are combined into a single text input. These texts are then transformed into numerical features using TF-IDF vectorization with parameters optimized for medical text (max_features=1000, ngram_range=(1,1), and English stop words removal).

The SVM classifier is configured with a linear kernel, which is particularly effective for text classification tasks due to its ability to handle high-dimensional sparse data. Key hyperparameters include a regularization parameter C=1 (which controls the trade-off between maximizing the margin and minimizing classification error) and probability=True to enable probability estimates for each class.

3.8.5 Developed Design for BiomedBERT

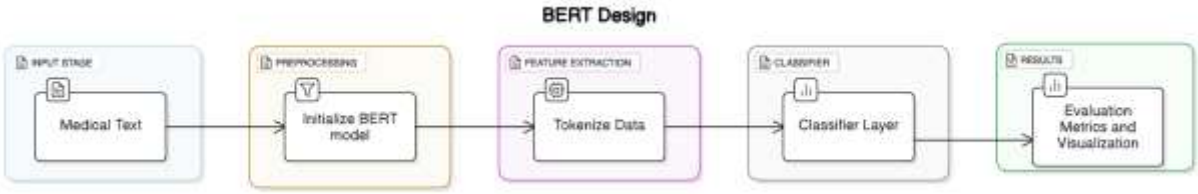

Figure 11: BiomedBERT design

The BiomedBERT model is loaded for sequence classification, as shown in Figure [11], with the number of labels set to the distinct classes in the dataset. The model is then trained using the Trainer class which allows the evaluation of multiple metrics, including accuracy, precision, recall, F1 score, and others, calculating these against the model's performance on the validation dataset.

3.8.6 Developed Design for Resnet50

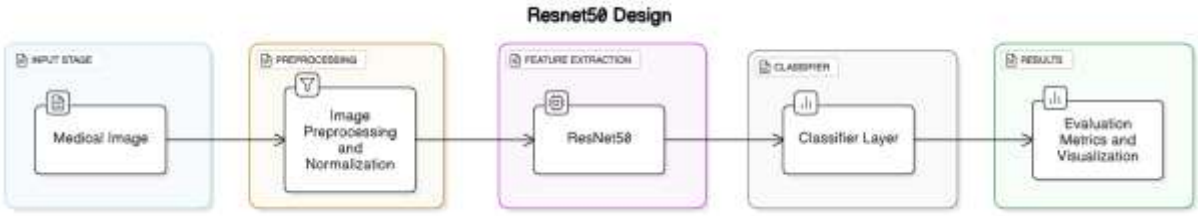

Figure 12: Resnet50 design

The proposed architecture in Figure [12] utilizes Resnet50, which is a pre-trained Convolutional Neural Network (CNN) with 50 layers designed to handle deep architectures using residual connections to mitigate vanishing gradient issues. In this setup, the final classification layer is removed, retaining the convolutional backbone to extract features from medical images.

Images are preprocessed (resized to 224x224, converted to RGB, normalized) and passed through ResNet50, producing a feature map of shape [B, 2048, 1, 1], where B is the batch size. This is flattened to a 2048-dimensional feature vector per image. The feature vector is fed into a linear classifier to predict one of the 12 medical condition classes.

3.8.7 Developed Design for DenseNet121

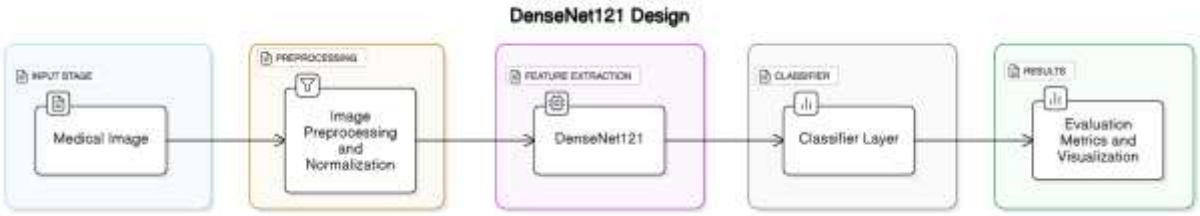

Figure 13: DenseNet121 design

The proposed design, in Figure [13], leverages DenseNet121 as the image feature extractor. DenseNet121 is a pre-trained Convolutional Neural Network that employs dense connections between layers to encourage feature reuse and improve gradient flow. In this implementation, the model retains only the convolutional backbone (densenet.features) and removes the final classification layer to focus solely on feature extraction.

The medical images are passed through the DenseNet121 feature extractor, producing a feature map of shape [B, 1024, 7, 7], where B is the batch size. This output is then passed through an adaptive average pooling layer to reduce the dimensions to [B, 1024, 1, 1]. The result is flattened into a 1024-dimensional vector that captures high-level visual representations of the input image. Finally, this 1024-dimensional feature vector is fed into a fully connected classification layer, which maps it to one of the 12 target classes.

3.8.8 Developed Design for EfficientNet

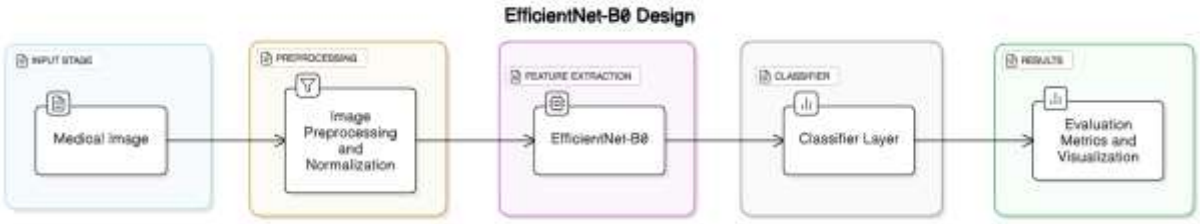

Figure 14: EfficientNet design

The design in Figure 14 utilizes EfficientNet-B0 as the image feature extractor. EfficientNet is a family of pre-trained Convolutional Neural Networks designed to achieve state-of-the-art

performance with high computational efficiency through a compound scaling method that balances network depth, width, and resolution. In this setup, EfficientNet-B0, the smallest variant, is used to process medical images.

Medical images are passed through EfficientNet-B0 after being transformed (resized to 224x224, converted to RGB, and normalized), producing a feature map that captures high-level visual patterns. The output is pooled and flattened into a feature vector, which is then used for classification and maps into one of the classes.

3.8.9 Developed Design for EfficientNet+BiomedBert Early Fusion

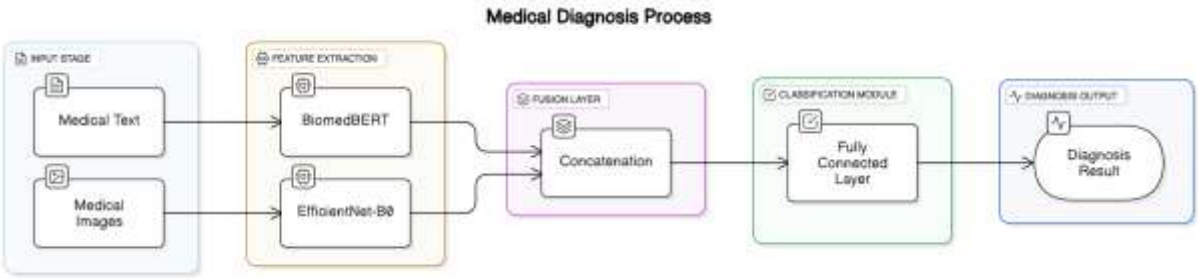

Figure 15: EfficientNet+BiomedBert Early Fusion design

The proposed design, as seen in Figure [14], involves integrating features from the medical text along with the medical images. The medical images go into the EfficientNet- B0, which is a pre-trained Convolutional Neural Network designed for visual feature extraction. It processes the images (CT or MRI) and outputs a 1280-dimensional feature vector, which is then reduced to 256. At the same time, the medical text utilized the BiomedBERT model, which is specialized for biomedical text understanding, to process text data such as the patient history and type of examinations. The BiomedBERT extracts a 768-dimensional feature vector by using the CLS token representation.

After the feature extraction stage, the concatenation fusion layer combines the 256 and 768-dimensional layers from both models into a unified 1024-feature vector and reduces it down to 512. This large vector is then passed into a fully connected layer to learn the patterns and similarities between the 2 modalities. The layer reduces the dimensionality down to 128. Finally, the output layer maps the features to the target classes.

3.8.10 Developed Design for EfficientNet+ BiomedBert Late Fusion

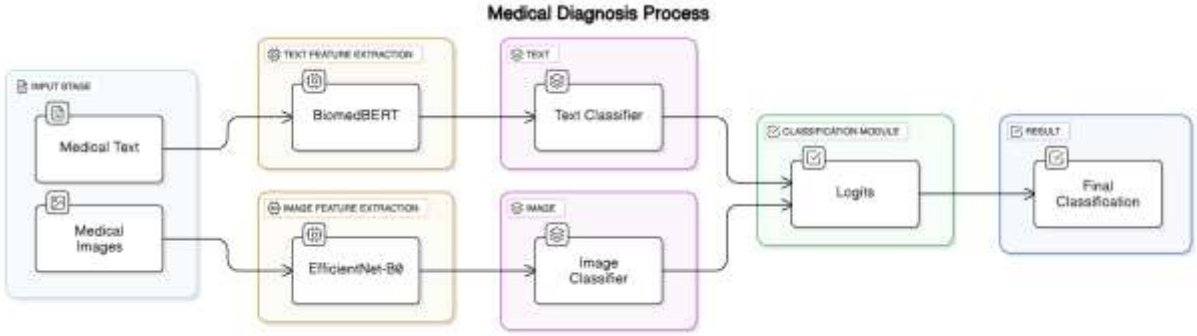

Figure 16: EfficientNet+BiomedBert Late Fusion design

The proposed design, as seen in Figure [16], involves integrating features from the medical text along with the medical images. The medical images go into the EfficientNet- B0, which is a pre-trained Convolutional Neural Network designed for visual feature extraction. We modify EfficientNet-B0 by removing its final fully connected layer, allowing us to extract a rich image feature embedding of size 1280. At the same time, the medical text utilized the BiomedBERT model, which is specialized for biomedical text understanding, to process text data such as the patient history and type of examinations. The BiomedBERT extracts a 768-dimensional feature vector by using the CLS token representation. Both feature vectors are extracted directly and then passed separately to their respective classification heads.

Each branch has its classification layer, which maps the 768-dimensional text features and the 1280-dimensional image features independently into 12 classes. After obtaining the logits from both classifiers, we apply softmax activation to convert them into probability distributions. The final prediction is then obtained by averaging/weighting the probability outputs from the text and image classifiers, ensuring that both modalities contribute equally to the final diagnosis.

3.8.11 Developed Design for EfficientNet Encoder+TF-IDF Vectorizer Intermediate Fusion with Cross-Modal Attention

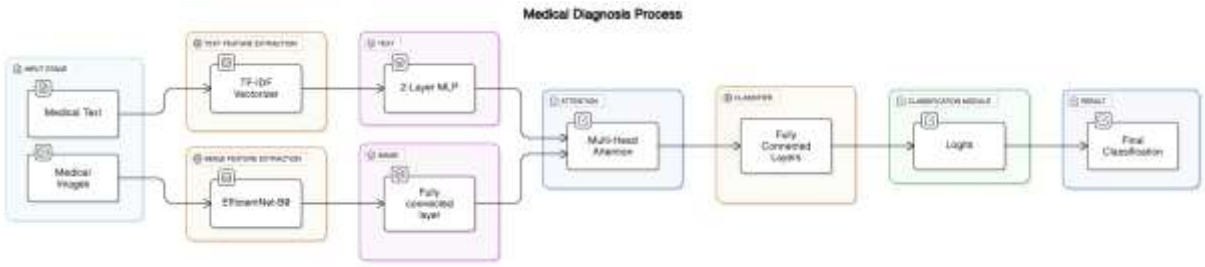

Figure 17: EfficientNet Encoder+TF-IDF Vectorizer Intermediate Fusion with Cross-Modal Attention Design

This multimodal classifier is designed to analyze medical data by integrating both image and text modalities for improved disease classification. The system in Figure [17] uses an EfficientNet-B0 model as the image encoder to extract visual features, while a simple feedforward network processes TF-IDF vectorized text data from clinical notes, including patient history, differential diagnosis, and case details. The model employs a cross-modal attention mechanism to dynamically fuse information between visual and textual features, allowing it to focus on relevant patterns across modalities. The image features are used as queries. The text features are used as keys and values. Attention weights determine how the model focuses on text, given visual context. With an embedding dimension of 128, the architecture includes intermediate fusion through a 4-head multi-head attention layer before final classification into one of 12 medical categories via a 256-unit ReLU network with dropout regularization. The training uses Adam optimization (lr=1e-3) with a cross-entropy loss on 224x224 normalized images and 3000-dimensional text features.

The implementation demonstrates several key specifications: image preprocessing with resizing to 224x224, batch processing (size=32), and stratified 80-20 train-test splitting to handle class imbalance. The text pipeline uses max_features=3000 TF-IDF vectorization with English stop word removal, while the image branch leverages pre-trained EfficientNet weights. The cross-attention module enables modality interaction at the feature level, and the classifier includes dropout (0.3) for regularization.

3.8.12 Developed Design for EfficientNet+Random Forest (Average Late Fusion)

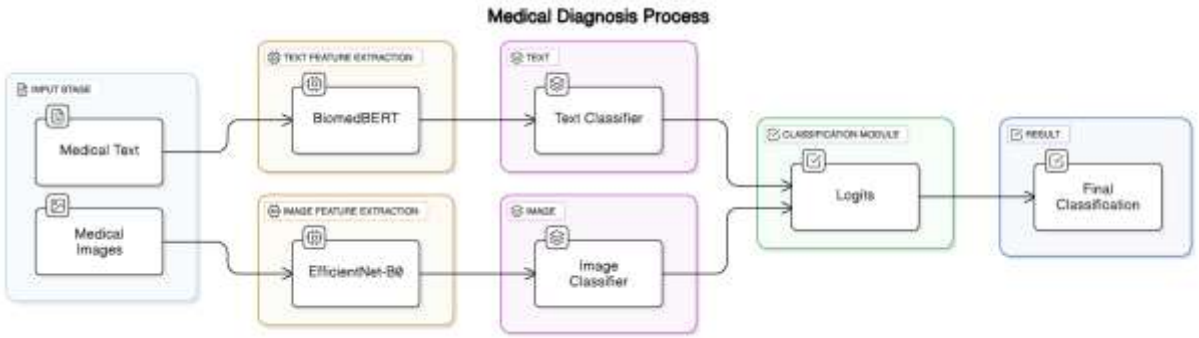

Figure 18: EfficientNet+RandomForest (Average Late Fusion) design

This multimodal classifier in Figure [18] combines deep learning for medical image analysis (EfficientNet-B0) with a traditional machine learning approach (Random Forest) for text processing. The imaging branch uses a pre-trained EfficientNet-B0 model, fine-tuned with a dropout-equipped classifier head (0.3 dropout rate), to predict one of 12 disease categories from 224x224 RGB images (resized and tensor-normalized). The text branch processes clinical notes (concatenating titles, histories, diagnoses, and locations) into 3000-dimensional TF-IDF vectors, which a Random Forest classifier (200 trees, max depth 15) uses to generate probabilistic predictions. The two modalities are fused via averaging (alpha = 0.5) of their output probabilities, balancing contributions from visual and textual features without complex cross-modal attention.

Key specifications include EfficientNet-B0 (pre-trained on ImageNet, Adam optimizer, 1e-3 learning rate, 5 epochs) for image encoding, Random Forest (n_estimators=200, stratified train-test split 80:20) for text, and late fusion via softmax probabilities. The pipeline handles single-modality fallbacks (using the first available MRI/CT image per case) and leverages CUDA for GPU acceleration.

3.8.13 EfficientNet+Random Forest(Average Late Fusion Without Miscellaneous Conditions)

This model is the same as the previous one, but it was trained on the data after removing the last class, which is the “Miscellaneous Conditions.” Since this class had the least number of correct

classifications among all models, this implementation was done to test if the model accuracy would change after removing the class.

3.9 Design Accomplishments

The final system design successfully meets all specified requirements. It integrates and evaluates two distinct machine-learning algorithms on both medical images and corresponding clinical text data. The system processes images from multiple modalities, including MRI CT, and supports various formats such as DICOM, PNG, and JPG. Performance is rigorously assessed using three key evaluation metrics. Additionally, robust preprocessing techniques and data augmentation are applied to enhance model performance and address quality variations. Most importantly, all data handling procedures strictly adhere to HIPAA regulations, ensuring patient privacy and compliance with healthcare data standards.

4 Results

4.1 Prototype Setup

The main objective of the design was to develop a multi-modal classification model that is capable of accurately diagnosing medical conditions using CT/MRI scans as well as the associated text. The environment in which the design was implemented was Google Colab and the A100 as our GPU. For the code section, PyTorch was utilized for the ResNet implementation and the fully connected layers, while for the text the DistilBERT transformer was used for text encoding.

As mentioned in the design, the model architecture is made from 3 main components: Image processing, Text Processing, and the fusion layer. ResNet50 was used to extract features from the CT/MRI images after normalizing them. For the text, a pre-trained BERT model (DistilBERT) was used to tokenize the text into vector representations. An early fusion approach was implemented by concatenating the feature vectors from both the text and images into 1 vector, which was then passed through fully connected layers and a softmax classifier for the final prediction.

4.2 Testing

Each case with its corresponding images was treated as a single data point, and the dataset was split 80/20. From the text data, only 4 fields were used: the history, examination, location of the injury, and the label. The labels are already numbered 0-11, representing the medical cases. A single patient may have multiple images from different modalities (CT, MRI), so the code combines all the images into a tensor and then averages it out.

The training was configured with a batch size of 16 and optimized using the Adam optimizer along with a learning rate of 0.001. The loss function utilized was the Cross-Entropy, which is suitable for a multiclass classification. The model was trained for 5 epochs and periodic checkpoints ensured that training could be resumed if interrupted, and the model state would be saved.

The model was evaluated using several performance metrics: the precision, recall, and the F1-Score for each class. A confusion matrix was also generated to visually assess the model's

performance and identify any misclassification trends. These metrics provide a comprehensive understanding of the model's ability to deal with multi-modal data.

4.3 Results Discussion

4.3.1 Resnet50+DistilBert

The multi-modal built using the proposed design shown in Figure [7] achieved accurate results, which demonstrate its ability to integrate both medical images and text data effectively. The test set accuracy was 94.46%. This reflects the model's ability to leverage information from both modalities effectively.

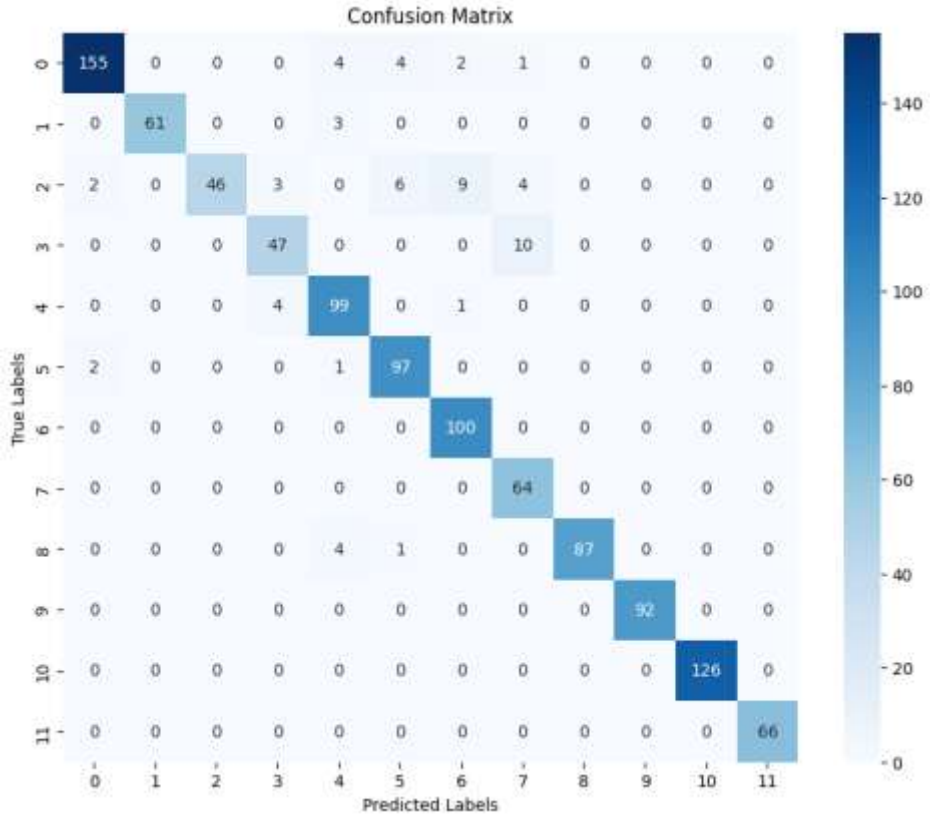

Figure 19: Resnet50+DistilBert confusion matrix

Table [14] below presents the precision, recall, and F1-Score for each class. The model demonstrates excellent classification results, with several classes achieving perfect scores. For these classes, the model was able to successfully identify all true cases without any false positives

or false negatives. The model performs overall well with the rest of the classes with the majority achieving F1-Scores of 0.9 and greater.

Table 14: Resnet50+DistilBert design Precision, Recall, F1-Score, and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.97	0.93	0.95	0.07
Class 1	1.00	0.95	0.98	0.05
Class 2	1.00	0.66	0.79	0.34
Class 3	0.87	0.82	0.85	0.18
Class 4	0.89	0.95	0.92	0.05
Class 5	0.9	0.97	0.93	0.03
Class 6	0.89	1.00	0.94	0
Class 7	0.81	1.00	0.9	0
Class 8	1.00	0.95	0.97	0.05
Class 9	1.00	1.00	1.00	0
Class 10	1.00	1.00	1.00	0
Class 11	1.00	1.00	1.00	0

4.3.2 Logistic Regression

The model achieved 88.82% accuracy with the confusion matrix shown below:

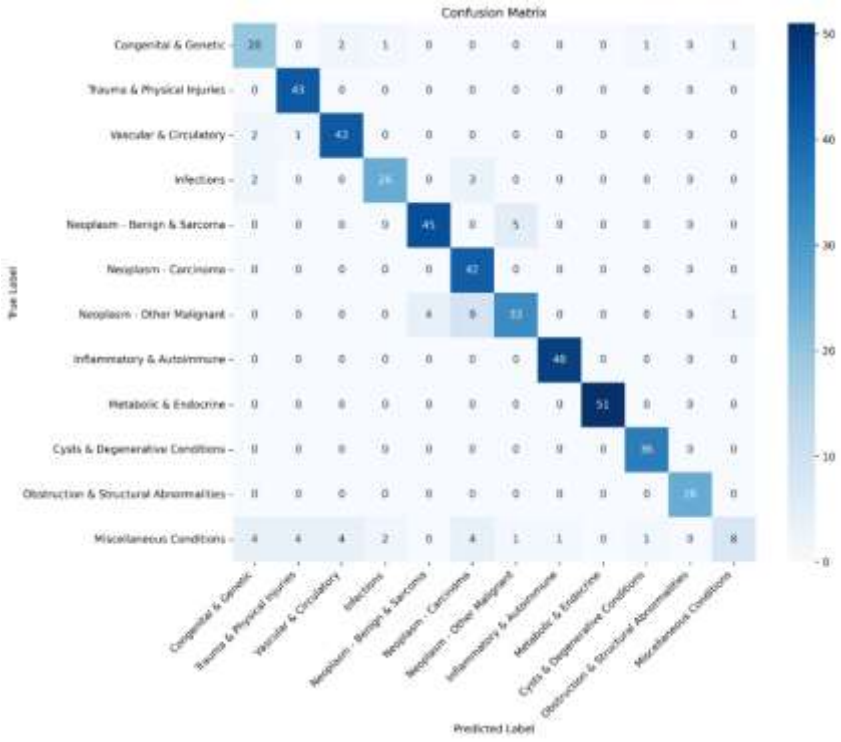

Figure 20: Logistic Regression confusion matrix

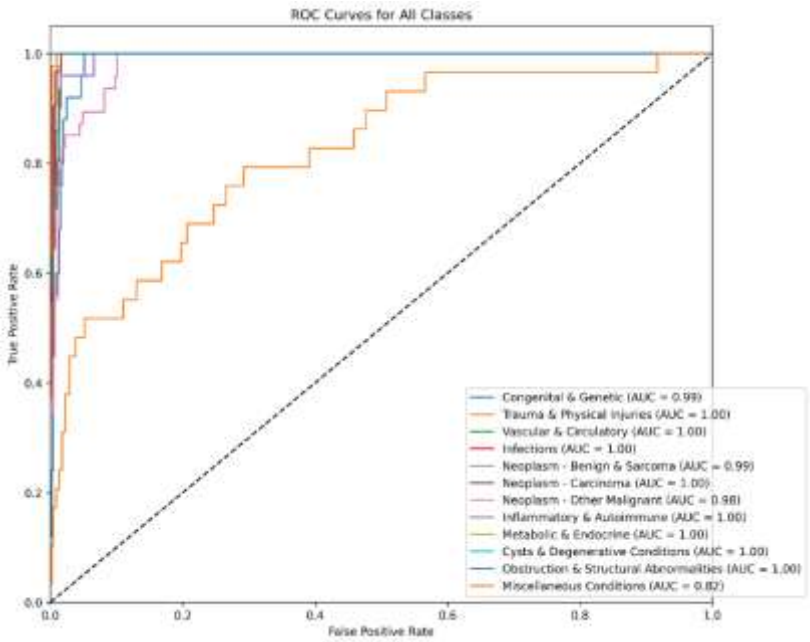

Figure 21: Logistic Regression AUC-ROC curve

Figure [21] above shows the AUC-ROC curve for all classes, where the AUC value of most classes is closer to 1, indicating a high ability to distinguish between classes, apart from the miscellaneous conditions class.

Table 15: Logistic Regression Precision, Recall, F1-Score, and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.71	0.80	0.75	0.2
Class 1	0.9	1.00	0.95	0
Class 2	0.88	0.93	0.91	0.07
Class 3	0.90	0.84	0.87	0.16
Class 4	0.92	0.90	0.91	0.1
Class 5	0.72	1.00	0.84	0
Class 6	0.85	0.70	0.77	0.3
Class 7	0.98	1.00	0.99	0
Class 8	1.00	1.00	1.00	0
Class 9	0.95	1.00	0.97	0
Class 10	1.00	1.00	1.00	0
Class 11	0.80	0.28	0.41	0.72

Table [15] presents the precision, recall, and F1-Score for each class. The classifier demonstrates strong overall performance, with several classes such as Class 1, 2, 4, 8, 9, and 10

achieving perfect or near-perfect scores across all metrics, indicating highly confident and consistent predictions for those categories.

4.3.3 Random Forest

The RF model was trained with 500 estimators, a max depth of 15, and 5 min samples split. It achieved an accuracy of 89.66% with the confusion matrix below:

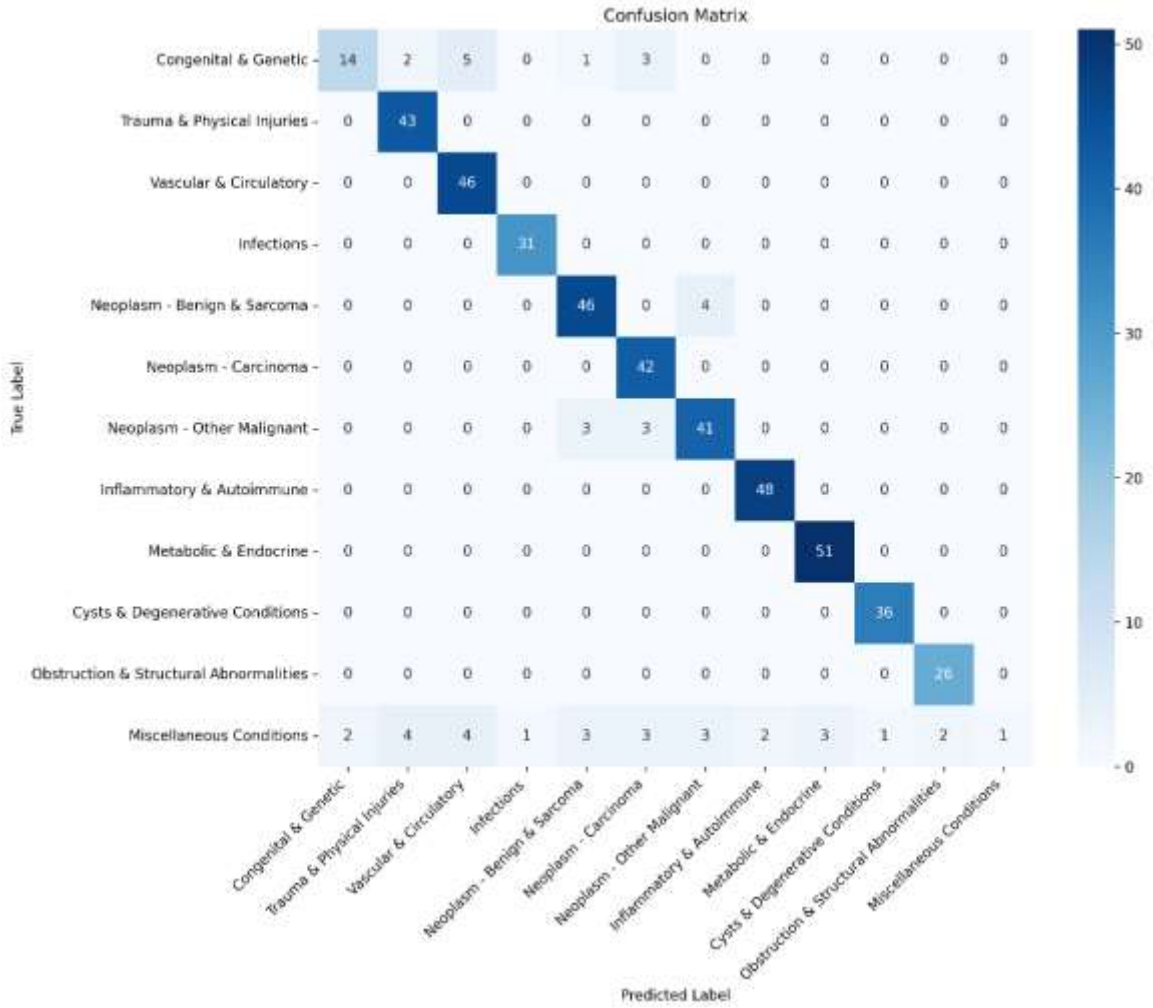

Figure 22: Random Forest confusion matrix

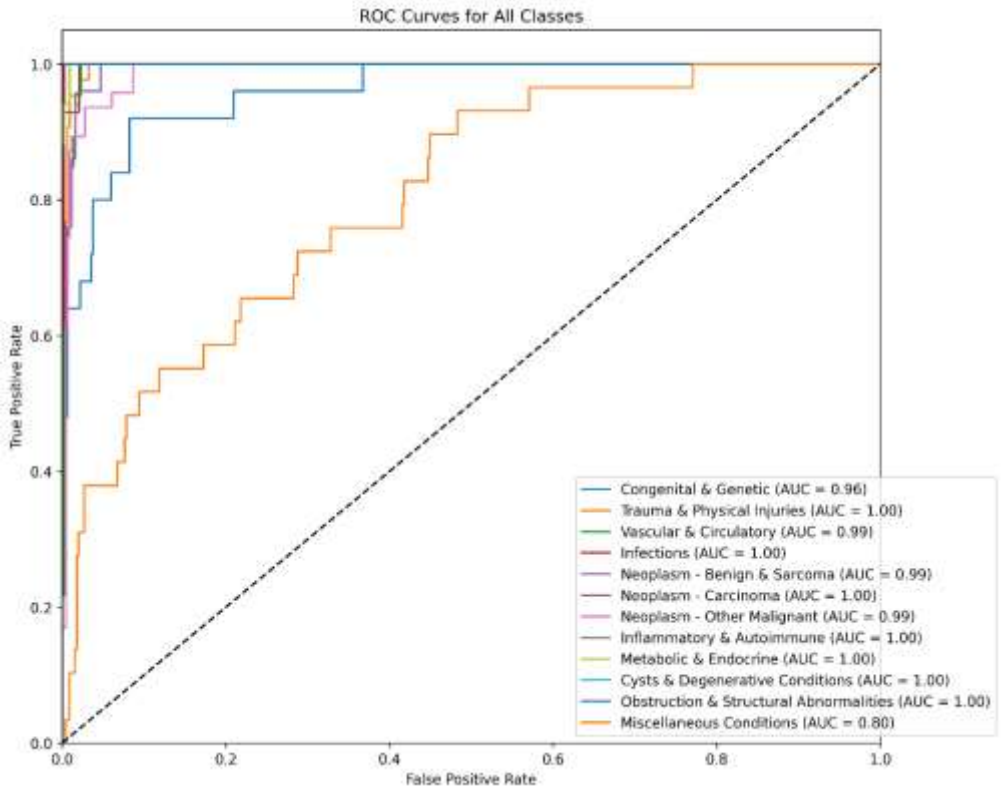

Figure 23: Radom Forest AUC-ROC curve

Figure [23] above shows the AUC-ROC curve for all classes, where the AUC value of most classes is closer to 1, indicating a high ability to distinguish between classes.

Table 16: Random Forest Precision, Recall, F1-Score, and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.88	0.56	0.68	0.44
Class 1	0.88	1.00	0.93	0
Class 2	0.84	1.00	0.91	0
Class 3	0.97	1.00	0.98	0
Class 4	0.87	0.92	0.89	0.08
Class 5	0.82	1.00	0.90	0
Class 6	0.85	0.87	0.86	0.13
Class 7	0.96	1.00	0.98	0
Class 8	0.95	1.00	0.97	0
Class 9	0.97	1.00	0.99	0
Class 10	0.93	1.00	0.96	0
Class 11	1.00	0.03	0.07	0.97

Table [16] presents the precision, recall, and F1-Score for each class. The classifier demonstrates strong overall performance, with several classes, such as Class 2, 4,7, 8, 9, and 10, achieving perfect or near-perfect scores across all metrics, indicating highly confident and consistent predictions for those categories.

4.3.4 Support Vector Machine

It achieved an accuracy of 89.51% with the confusion matrix below:

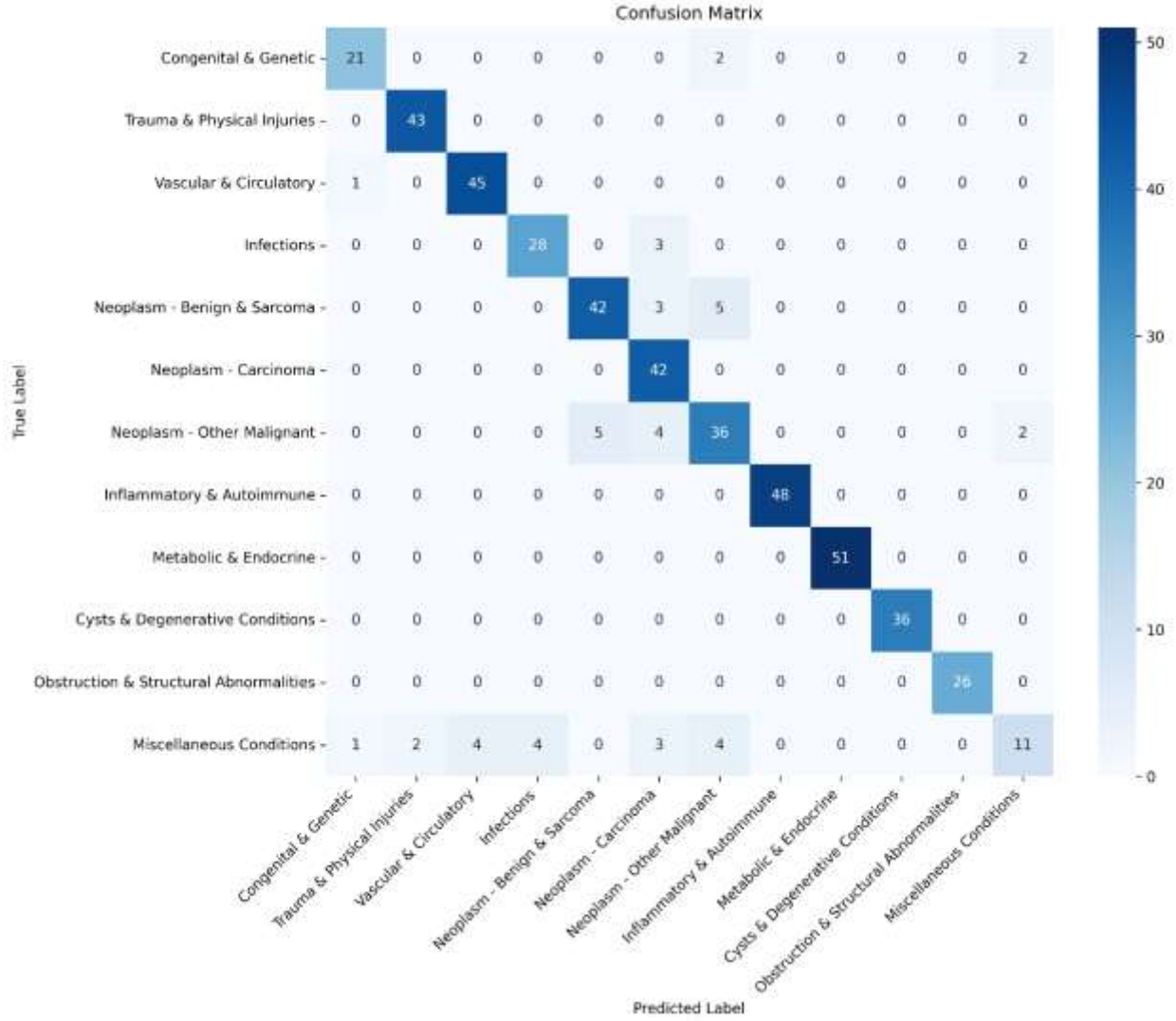

Figure 24: SVM confusion matrix

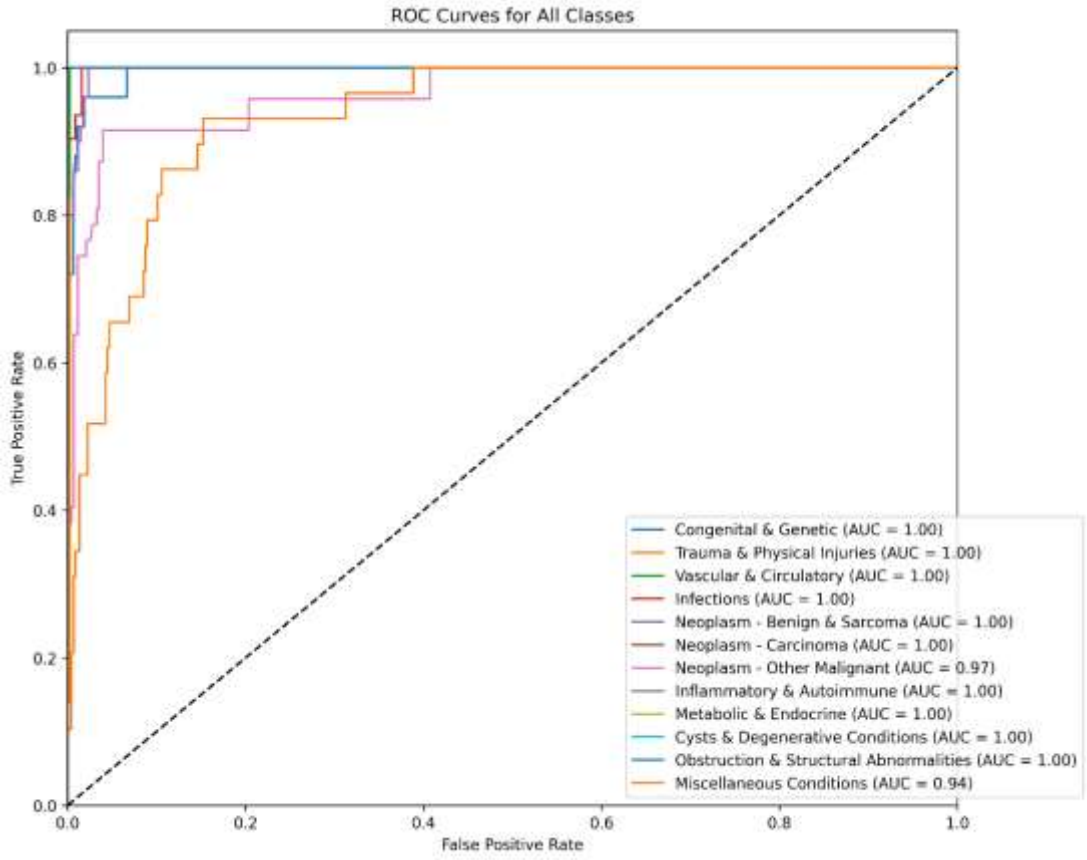

Figure 25: SVM AUC-ROC curve

The figure above shows the AUC-ROC curve for all classes, where the AUC value of most classes is closer to 1, indicating a high ability to distinguish between classes.

Table 17: SVM Precision, Recall, F1-Score, and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.91	0.84	0.88	0.16
Class 1	0.96	1.00	0.98	0
Class 2	0.92	0.98	0.95	0.02
Class 3	0.88	0.90	0.89	0.1
Class 4	0.89	0.84	0.87	0.16
Class 5	0.76	1.00	0.87	0
Class 6	0.77	0.77	0.77	0.23
Class 7	1.00	1.00	1.00	0
Class 8	1.00	1.00	1.00	0
Class 9	1.00	1.00	1.00	0
Class 10	1.00	1.00	1.00	0
Class 11	0.73	0.38	0.50	0.62

Table [17] presents the precision, recall, and F1-Score for each class. The classifier demonstrates strong overall performance, with several classes, such as Class 7, 8, 9, and 10, achieving perfect or near-perfect scores across all metrics, indicating highly confident and consistent predictions for those categories.

4.3.5 BiomedBert

It achieved an accuracy of 93.25% with the confusion matrix below

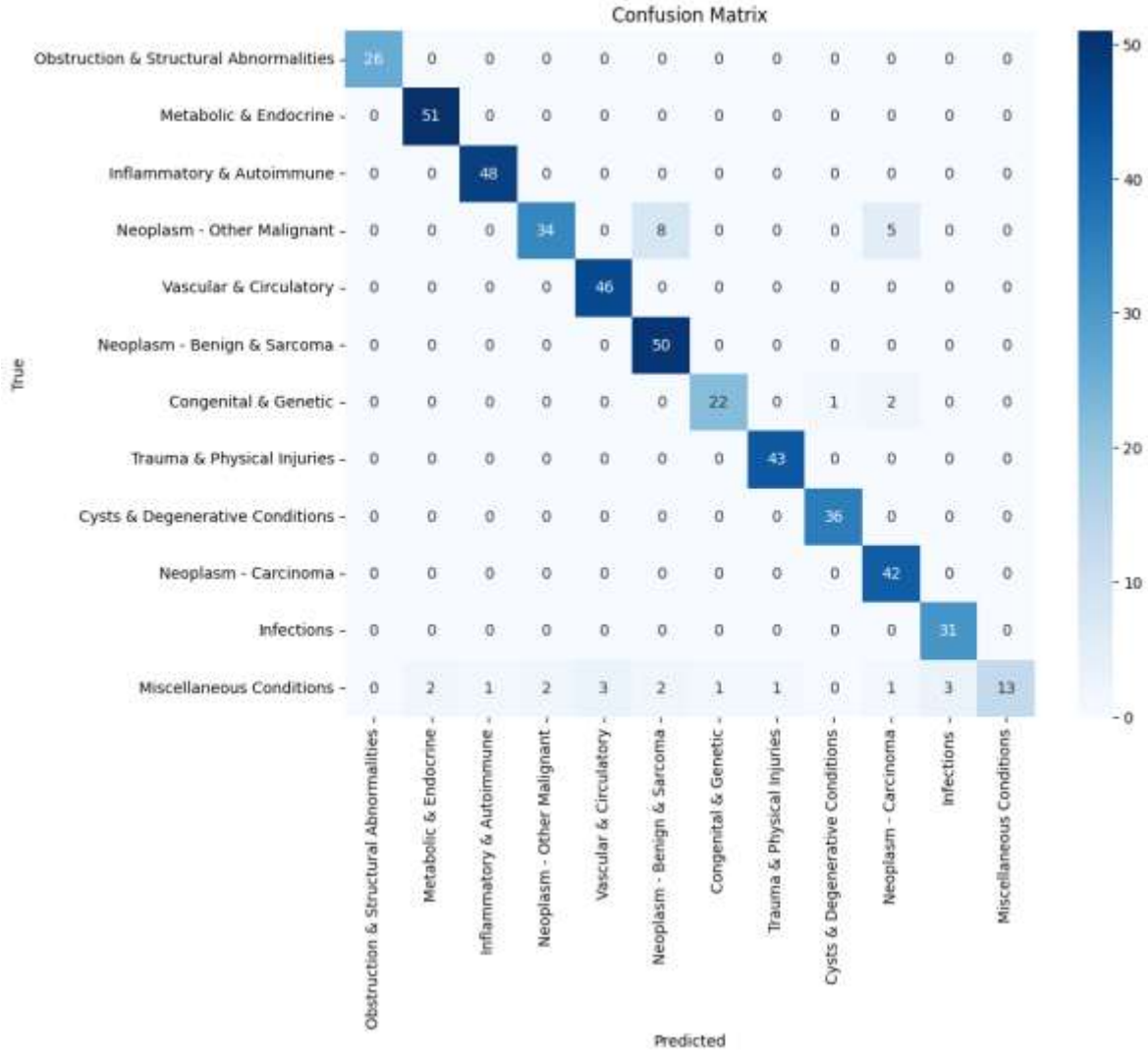

Figure 26: BiomedBert confusion matrix

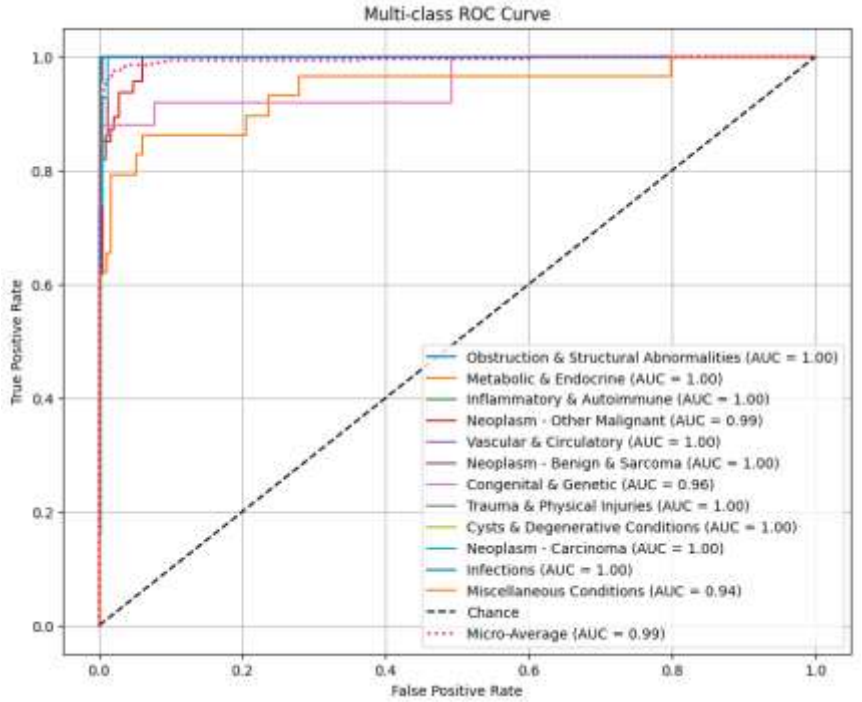

Figure 27: BiomedBert AUC-ROC curve

Figure [27] above shows the AUC-ROC curve for all classes, where the AUC value of most classes is closer to 1, indicating a high ability to distinguish between classes.

Table 18: BiomedBert Precision, Recall, F1-Score, and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	1.00	1.00	1.00	0
Class 1	0.96	1.00	0.98	0
Class 2	0.98	1.00	0.99	0
Class 3	0.94	0.72	0.82	0.28
Class 4	0.94	1.00	0.97	0
Class 5	0.83	1.00	0.91	0
Class 6	0.96	0.88	0.92	0.12
Class 7	0.98	1.00	0.99	0
Class 8	0.97	1.00	0.99	0
Class 9	0.84	1.00	0.91	0
Class 10	0.91	1.00	0.95	0
Class 11	1.00	0.45	0.62	0.55

Table [18] presents the precision, recall, and F1-Score for each class. The classifier demonstrates strong overall performance, with several classes, such as Class 0, 2, 4, and 5,

achieving perfect or near-perfect scores across all metrics, indicating highly confident and consistent predictions for those categories.

Table [19] below summarizes and compares text models.

Table 19: Text models summary tables

Model Type	Accuracy	Macro F1-Score	Macro AUC-ROC	Macro FN rate	Training Time(minutes)
Logistic Regression	88.82%	0.86	0.9811	0.1292	9
Random Forests	89.66%	0.84	0.9781	0.1350	11
SVM	89.51%	0.89	0.9925	0.1075	10
BiomedBert	93.25%	0.92	0.9908	0.0792	16

4.3.6 ResNet50

The ResNet50 model achieved an accuracy of 91.83%, and the confusion matrix is shown below.

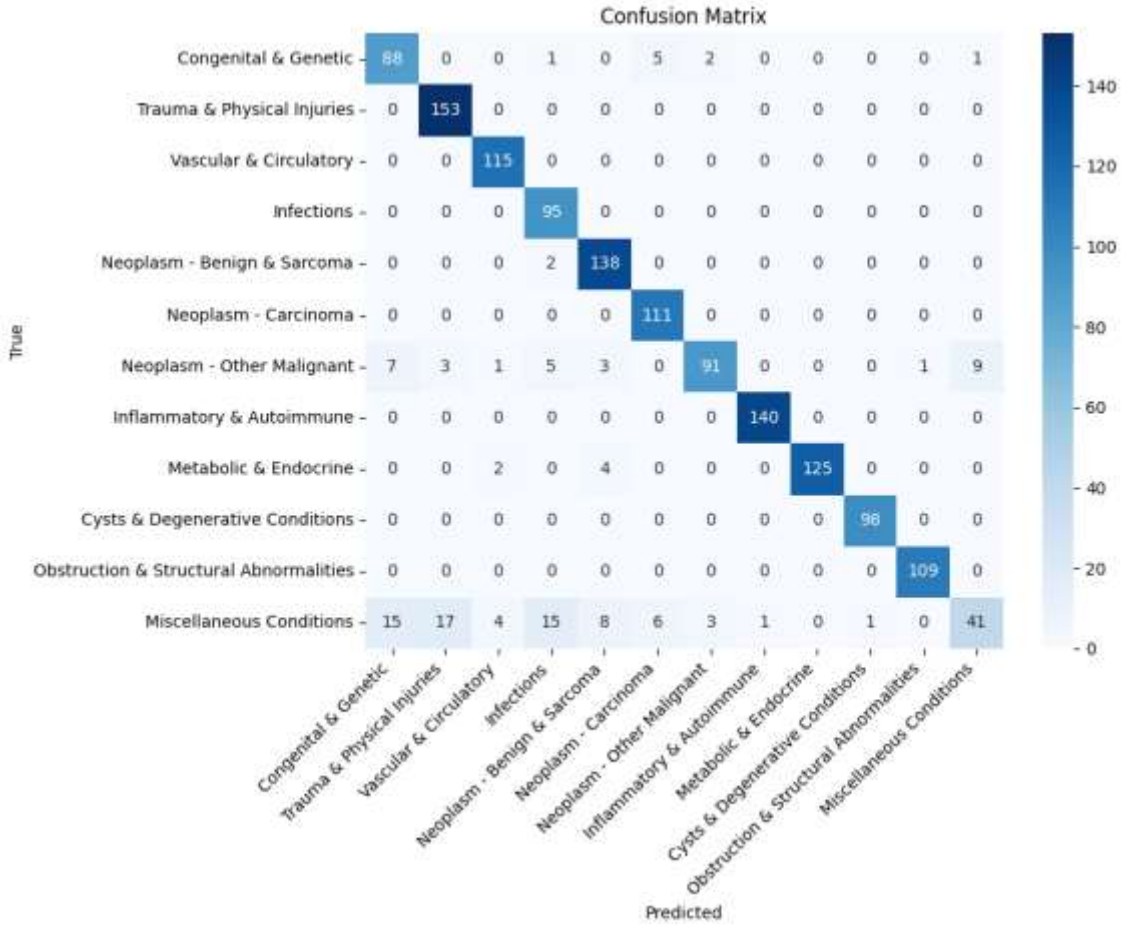

Figure 28: Resnet50 confusion matrix

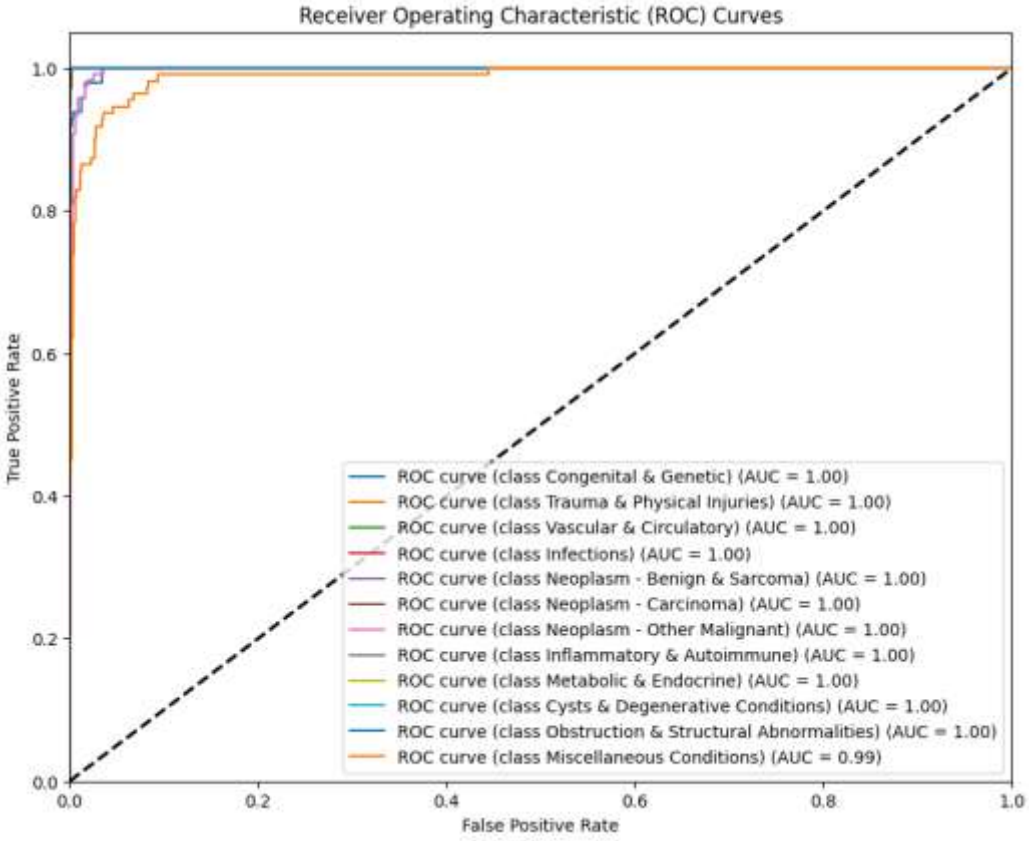

Figure 29: ResNet50 AUC-ROC curve

Table 20: ResNet50 Precision, Recall, F1-Score, and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.93	0.96	0.89	0.04
Class 1	0.98	1.00	0.99	0
Class 2	1.00	1.00	1.00	0
Class 3	0.94	1.00	0.97	0
Class 4	0.98	1.00	0.99	0
Class 5	0.97	1.00	0.98	0
Class 6	0.93	0.93	0.93	0.07
Class 7	1.00	1.00	1.00	0
Class 8	1.00	1.00	1.00	0
Class 9	1.00	1.00	1.00	0
Class 10	1.00	1.00	1.00	0
Class 11	0.94	0.67	0.78	0.33

Table [20] presents the precision, recall, and F1-Score for each class. Classes 1, 2, 4,7, 8, 9, and 10 stand out with exceptionally high scores, reflecting the model's strong ability to accurately and confidently identify these cases with minimal false positives or negatives.

4.3.7 DenseNet121

The DenseNet121 model achieved an accuracy of 95.21%, and the confusion matrix is shown below.

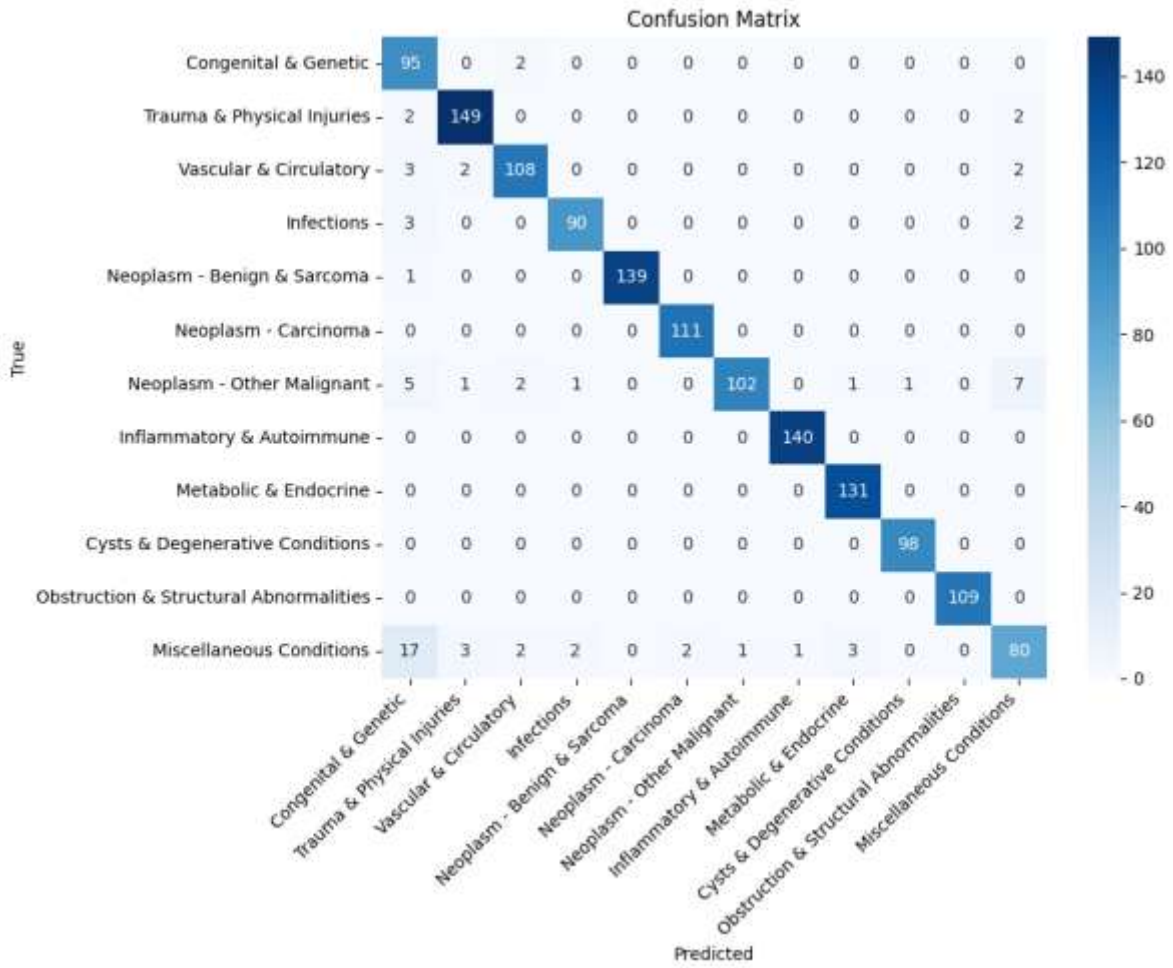

Figure 30: DenseNet121 confusion matrix

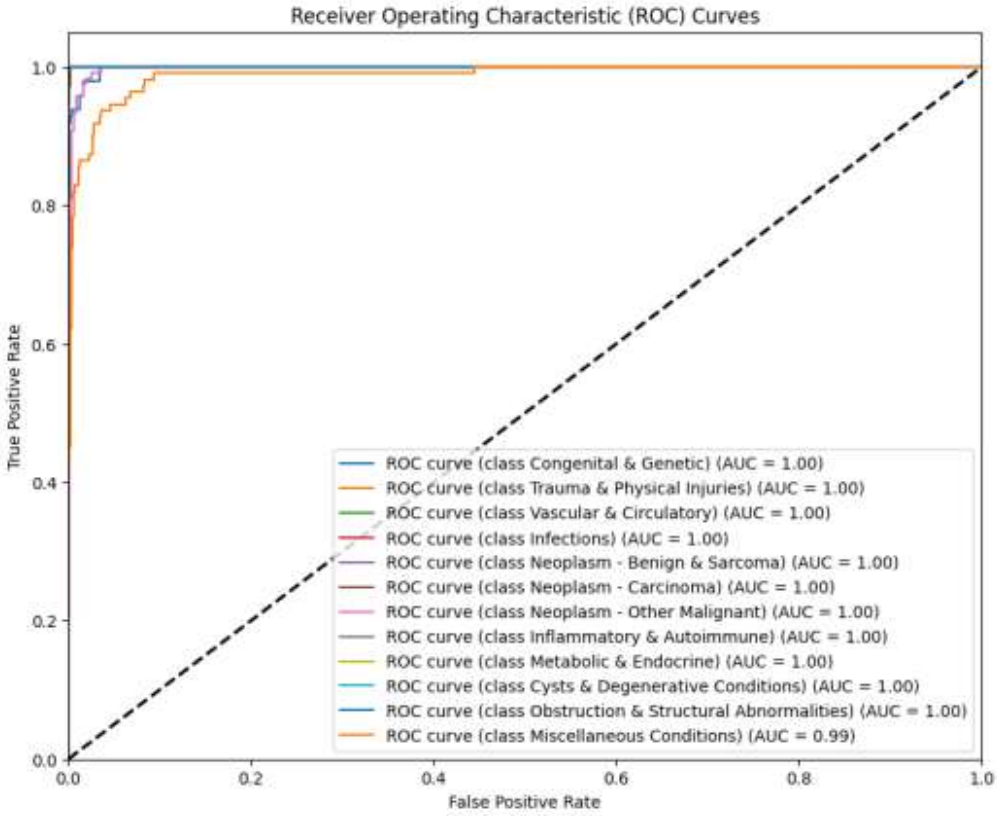

Figure 31: DenseNet121 AUC-ROC curve

Table 21: DenseNet121 Precision, Recall, F1-Score, and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.83	0.96	0.89	0.04
Class 1	0.98	1.00	0.99	0
Class 2	1.00	1.00	1.00	0
Class 3	0.94	1.00	0.97	0
Class 4	0.98	1.00	0.99	0
Class 5	0.97	1.00	0.98	0
Class 6	0.93	0.93	0.93	0.07
Class 7	1.00	1.00	1.00	0
Class 8	1.00	1.00	1.00	0
Class 9	1.00	1.00	1.00	0
Class 10	1.00	1.00	1.00	0
Class 11	0.94	0.67	0.78	0.33

Table [21] presents the precision, recall, and F1-Score for each class. The model exhibits robust classification performance, particularly for Classes 1, 2, 3, 4, 8, 9, and 10, where both

precision and recall are consistently high. This suggests the model is able to identify these conditions with high confidence and minimal error.

4.3.8 EfficeintNet-B0

The EfficientNet-B0 model achieved 96.06% with the confusion matrix shown below:

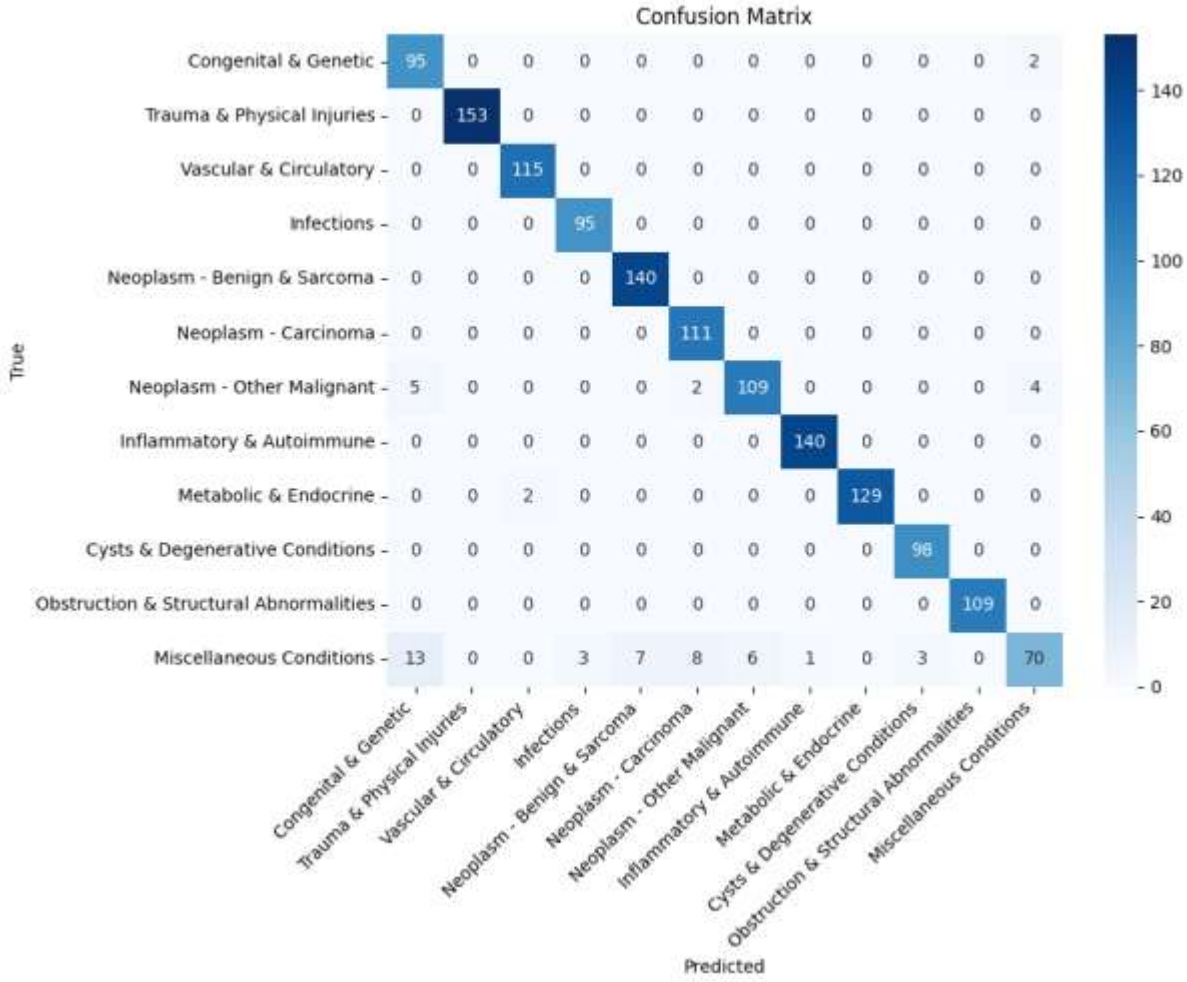

Figure 32: EfficientNet-B0 confusion matrix

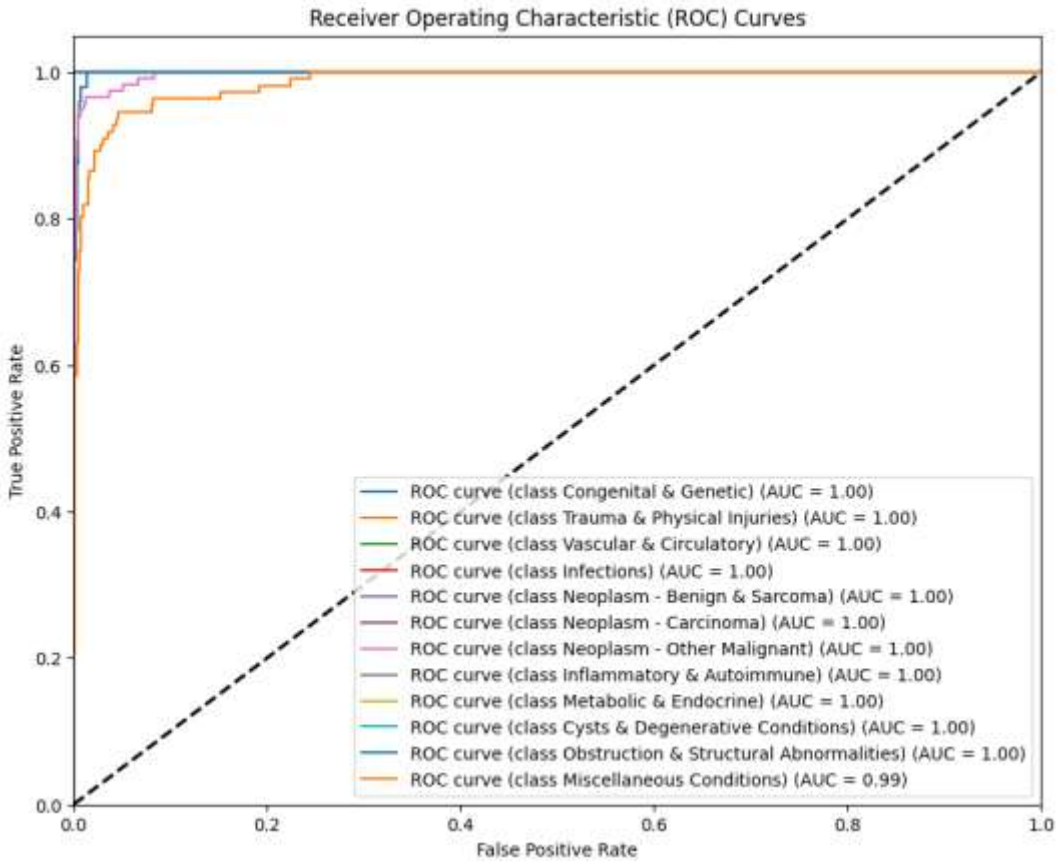

Figure 33: EfficientNet-B0 AUC-ROC curve

Table 22: EfficientNet-B0 Precision, Recall, F1-Score, and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.84	0.98	0.90	0.02
Class 1	1.00	1.00	1.00	0
Class 2	0.98	1.00	0.99	0
Class 3	0.97	1.00	0.98	0
Class 4	0.95	1.00	0.98	0
Class 5	0.92	1.00	0.96	0
Class 6	0.95	0.91	0.93	0.09
Class 7	0.99	1.00	1.00	0
Class 8	1.00	0.98	0.99	0.02
Class 9	0.97	1.00	0.98	0
Class 10	1.00	1.00	1.00	0
Class 11	0.92	0.63	0.75	0.37

The Precision, Recall, and F1-Score for each of the classes are shown in Table [22] above. The model demonstrates good classification results where most of the F1-Scores are 0.9 and above.

Table [23] below summarizes and compares text models.

Table 23: Image models summary tables

Model Type	Accuracy	Macro F1-Score	Macro AUC-ROC	Macro FN rate	Training Time (minutes)
ResNet50	91.83%	0.9608	0.9992	0.0367	62
DenseNet121	95.21%	0.9608	0.9992	0.0367	67
EfficientNet-B0	96.06%	0.9550	0.9992	0.0417	54

4.3.9 EfficientNet+BiomedBert(Early Fusion)

After Text Augmentation:

The model accuracy after text augmentation was 81.43%, with the confusion matrix shown below in Figure 33.

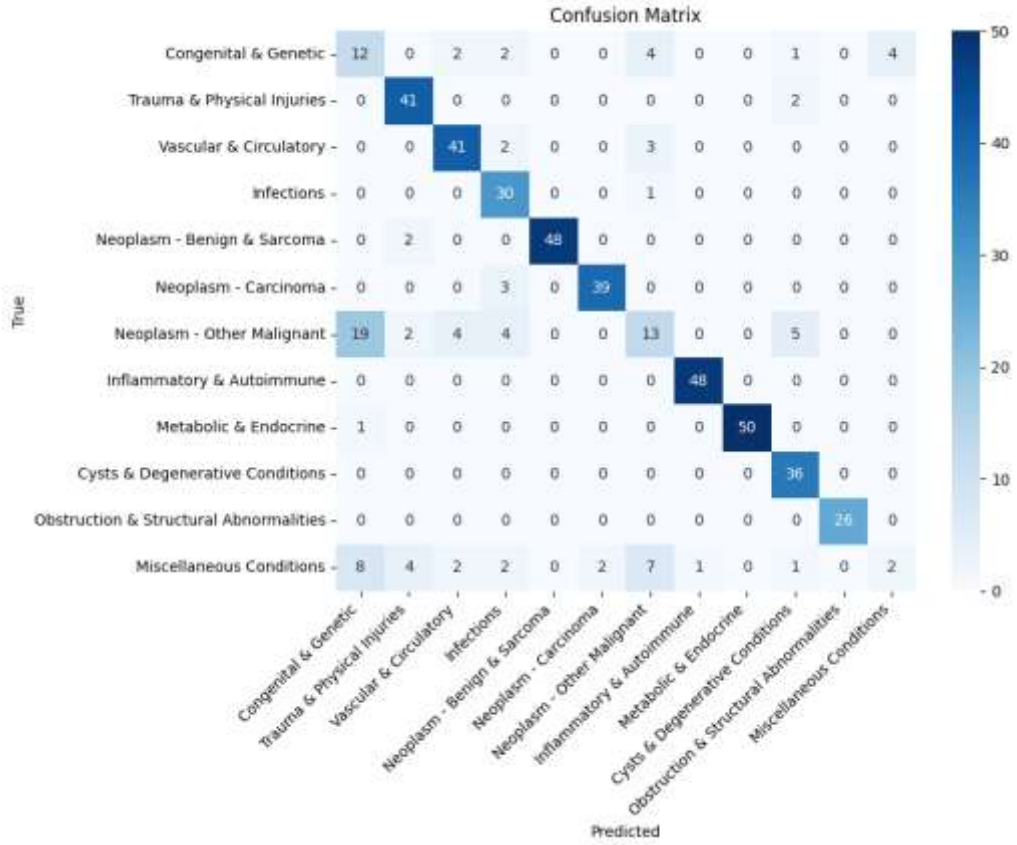

Figure 34: EfficientNet+BiomedBert Version1 confusion matrix

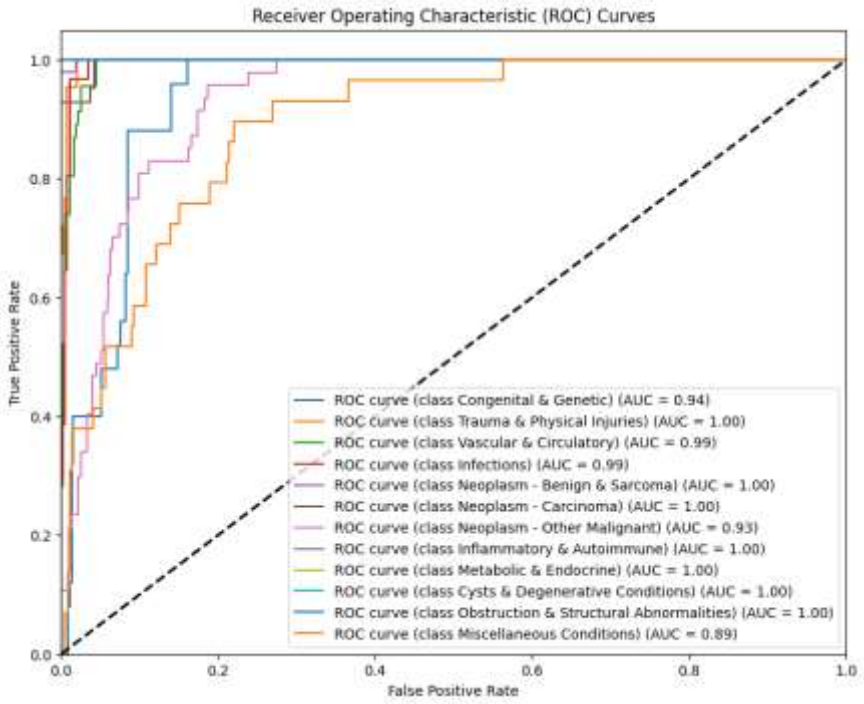

Figure 35: EfficientNet+BiomedBert Version1 AUC-ROC curve

Table 24: EfficientNet+BiomedBert Version1 Precision, Recall, F1-Score and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.30	0.48	0.37	0.52
Class 1	0.84	0.95	0.89	0.05
Class 2	0.84	0.89	0.86	0.11
Class 3	0.7	0.97	0.81	0.03
Class 4	1.00	0.96	0.98	0.04
Class 5	0.95	0.93	0.94	0.07
Class 6	0.46	0.28	0.35	0.72
Class 7	0.98	1.00	0.99	0
Class 8	1.00	0.98	0.99	0.02
Class 9	0.8	1.00	0.89	0
Class 10	1.00	1.00	1.00	0
Class 11	0.33	0.07	0.11	0.93

The Precision, Recall, and F1-Score for each of the classes are shown in Table [24] above. The model demonstrates good classification results where most of the F1-Scores are 0.8 and above, but the model struggles with some classes, such as class 0 and class 11.

After 1 Image Augmentation:

The model accuracy after 2-image augmentation was 92.46%, with the confusion matrix shown below in Figure 33.

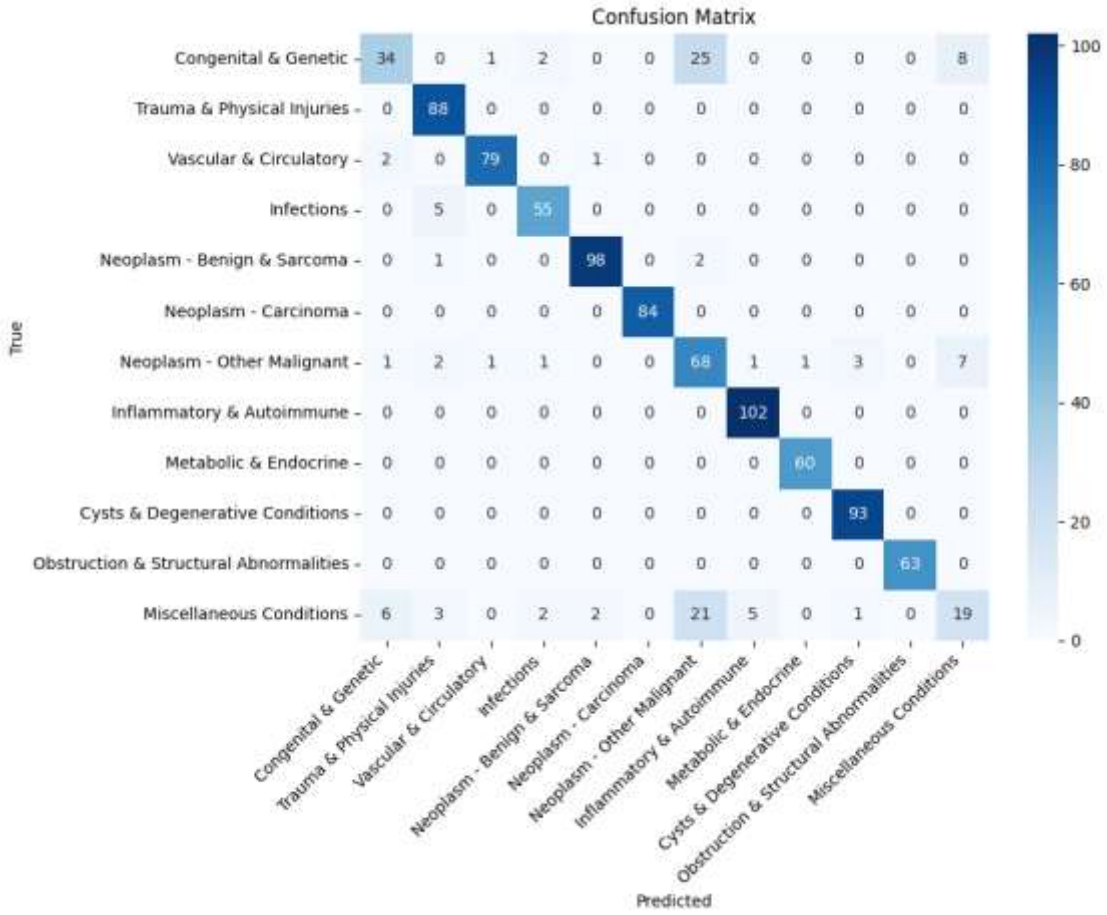

Figure 36: EfficientNet+BiomedBert Version2 confusion matrix

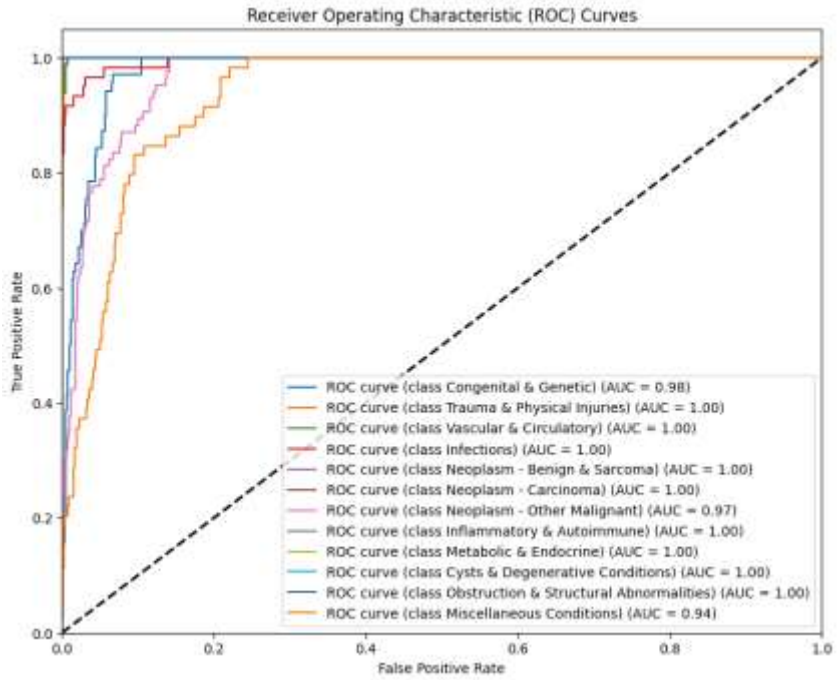

Figure 37: EfficientNet+BiomedBert Version2 AUC-ROC curve

Figure [37] above shows the AUC-ROC curve for all classes, where the AUC value of most classes is closer to 1, indicating a high ability to distinguish between classes.

Table 25: EfficientNet+BiomedBert Version2 Precision, Recall, F1-Score and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.79	0.49	0.6	0.51
Class 1	0.89	1.00	0.94	0
Class 2	0.98	0.96	0.97	0.04
Class 3	0.92	0.92	0.92	0.08
Class 4	0.97	0.97	0.97	0.03
Class 5	1.00	1.00	1.00	0
Class 6	0.59	0.8	0.68	0.2
Class 7	0.94	1.00	0.97	0
Class 8	0.98	1.00	0.99	0
Class 9	0.96	1.00	0.98	0
Class 10	1.00	1.00	1.00	0
Class 11	0.56	0.32	0.41	0.68

The Precision, Recall, and F1-Score for each of the classes are shown in Table [25] above. The model demonstrates good classification results where most of the F1-Scores are 0.9 and above.

After 2 image augmentation:

The model accuracy after 2-image augmentation was 92.46%, with the confusion matrix shown below in Figure 37.

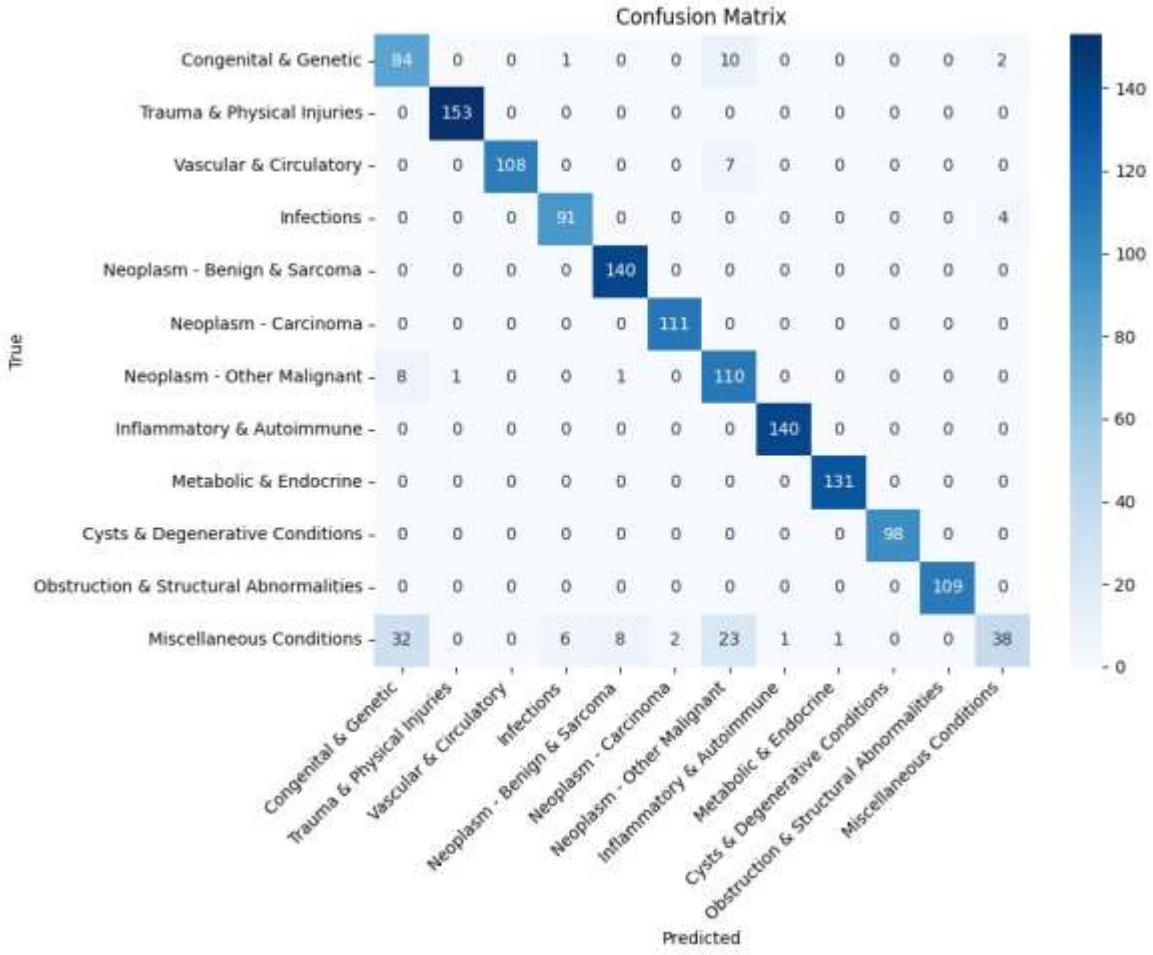

Figure 38: EfficientNet+BiomedBert Version3 confusion matrix

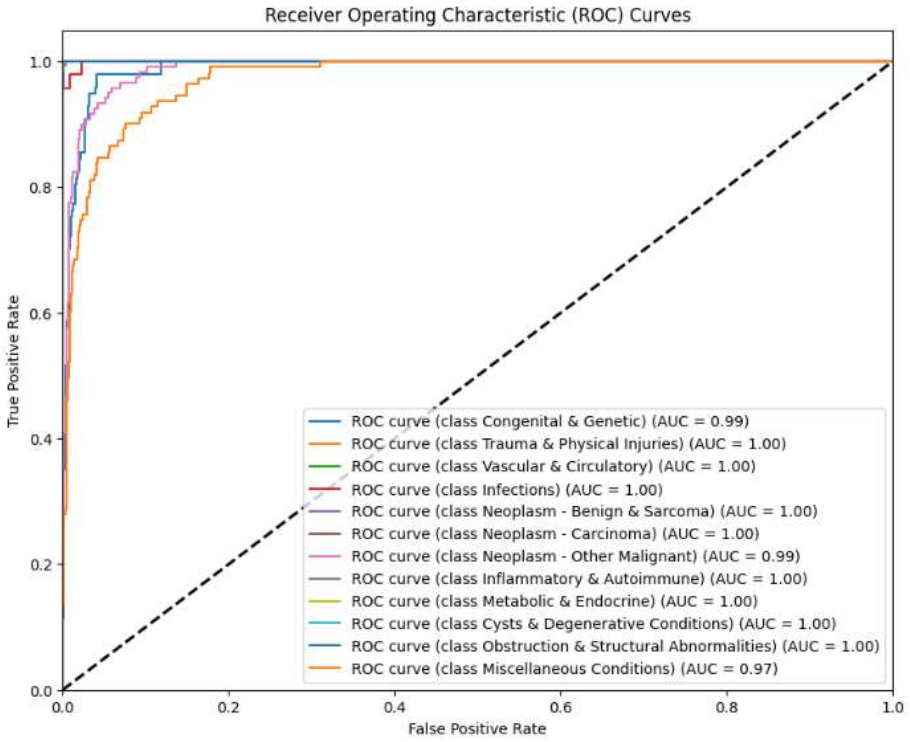

Figure 39: EfficientNet+BiomedBert Version3 AUC-ROC curve

Figure [39] above shows the AUC-ROC curve for all classes, where the AUC value of most classes is closer to 1, indicating a high ability to distinguish between classes.

Table 26: EfficientNet+BiomedBert Version3 Precision, Recall, F1-Score and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.68	0.87	0.76	0.13
Class 1	0.99	1.00	1.00	0
Class 2	1.00	0.94	0.97	0.06
Class 3	0.93	0.96	0.94	0.04
Class 4	0.94	1.00	0.97	0
Class 5	0.98	1.00	0.99	0
Class 6	0.73	0.92	0.81	0.08
Class 7	0.99	1.00	1.00	0
Class 8	0.99	1.00	1.00	0
Class 9	1.00	1.00	1.00	0
Class 10	1.00	1.00	1.00	0
Class 11	0.86	0.34	0.49	0.66

The Precision, Recall, and F1-Score for each of the classes are shown in Table [26] above. The model demonstrates good classification results where most of the F1-Scores are 0.95 and above

Summary of EfficientNet+BiomedBert Model:

Table 27: Summary Table for the EffiecientNet+BiomedBert versions

Augmentation Stage	Accuracy	Macro F1-Score	Macro AUC-ROC	Macro FN rate	Training Time (minutes)
After Text	81.43%	0.77	0.9789	0.2075	47
After 1 image augmentation	89.02%	0.88	0.9899	0.1283	52
After 2 image augmentation	92.46%	0.92	0.9959	0.0808	59

As shown in the Table [27] above, the best results are achieved after the 2-image augmentation due to having a larger amount of data and reducing overfitting by preventing the model from memorizing the data during training. The rest of the multimodal models are all using the data after 2 image augmentations.

4.3.10 EfficientNet+BiomedBert (Average Late Fusion)

The model achieved an accuracy of 96.55% with the confusion matrix shown below:

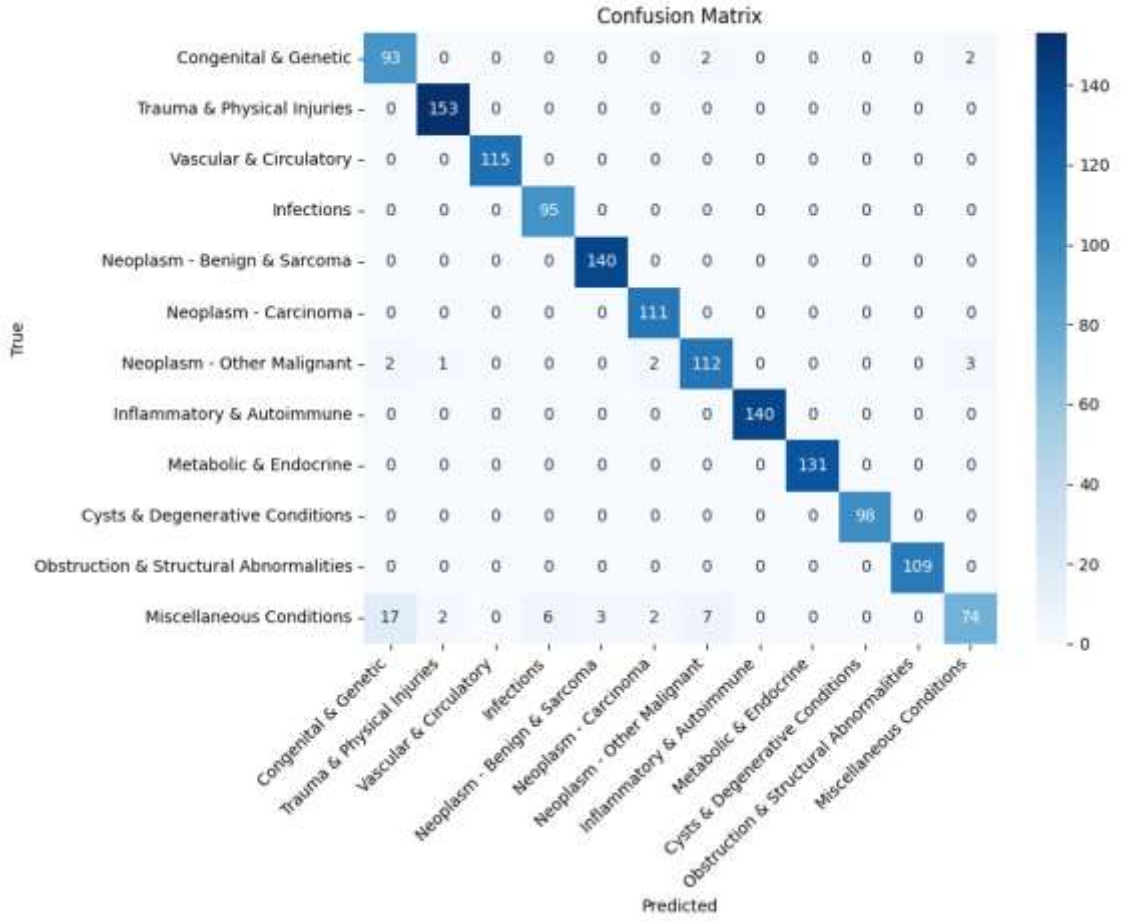

Figure 40: EfficientNet+BiomedBert Average Late Fusion confusion matrix

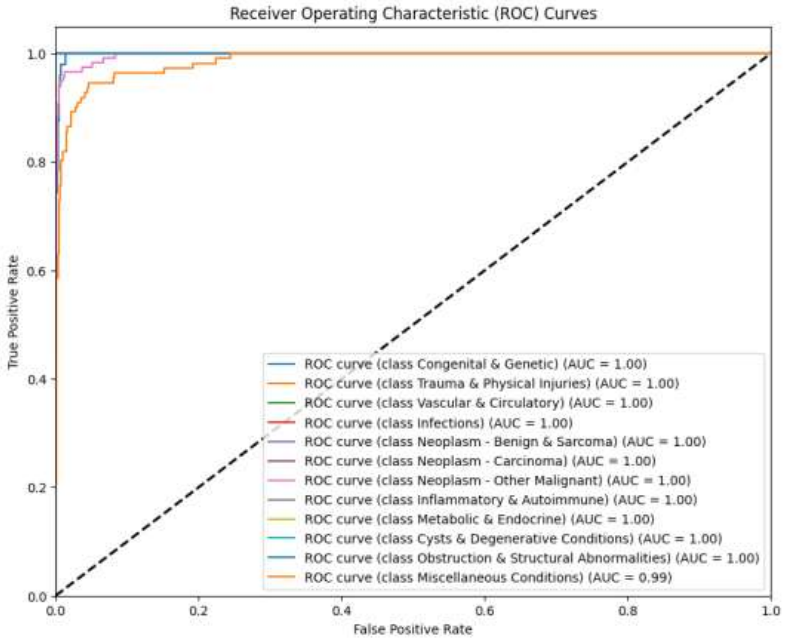

Figure 41: EfficientNet+BiomedBert Average Late Fusion AUC-ROC curve

Figure [41] above shows the AUC-ROC curve for all classes, where the AUC value of most classes is closer to 1, indicating a high ability to distinguish between classes.

Table 28: EfficientNet+BiomedBert Average Late Fusion Precision, Recall, F1-Score, and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.84	0.98	0.90	0.02
Class 1	1.00	1.00	1.00	0
Class 2	0.98	1.00	0.99	0
Class 3	0.97	1.00	0.98	0
Class 4	0.95	1.00	0.98	0
Class 5	0.92	1.00	0.96	0
Class 6	0.95	0.91	0.93	0.09
Class 7	0.99	1.00	1.00	0
Class 8	1.00	0.98	0.99	0.02
Class 9	0.97	1.00	0.98	0
Class 10	1.00	1.00	1.00	0
Class 11	0.92	0.63	0.75	0.37

Table [28] presents the precision, recall, and F1-Score for each class. The classifier demonstrates strong overall performance, with almost all of the classes achieving perfect or near-perfect scores across all metrics, indicating highly confident and consistent predictions for those categories.

4.3.11 EfficientNet+BiomedBert (Weighted Late Fusion)

The model achieved an accuracy of 96.20% with the confusion matrix shown below:

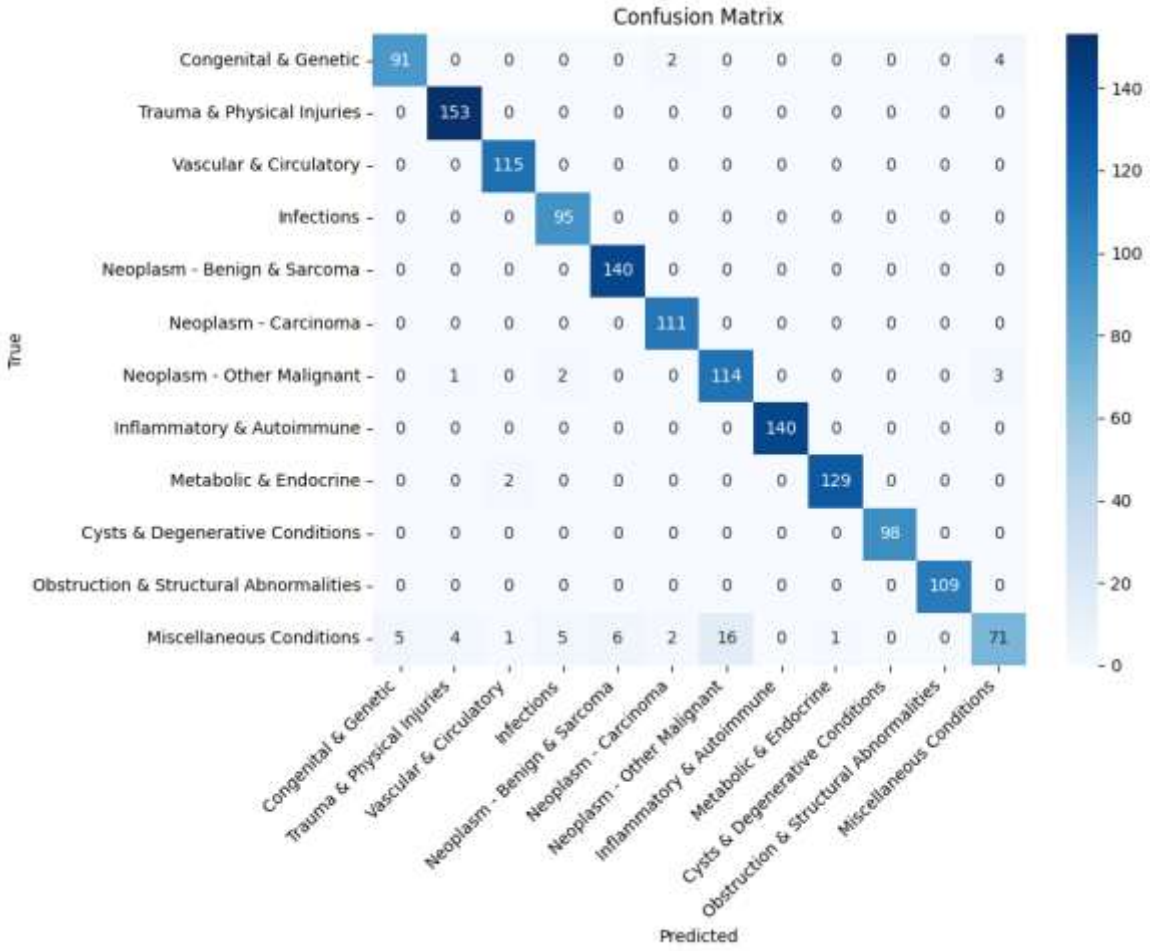

Figure 42: EfficientNet+BiomedBert Weighted Late Fusion confusion matrix

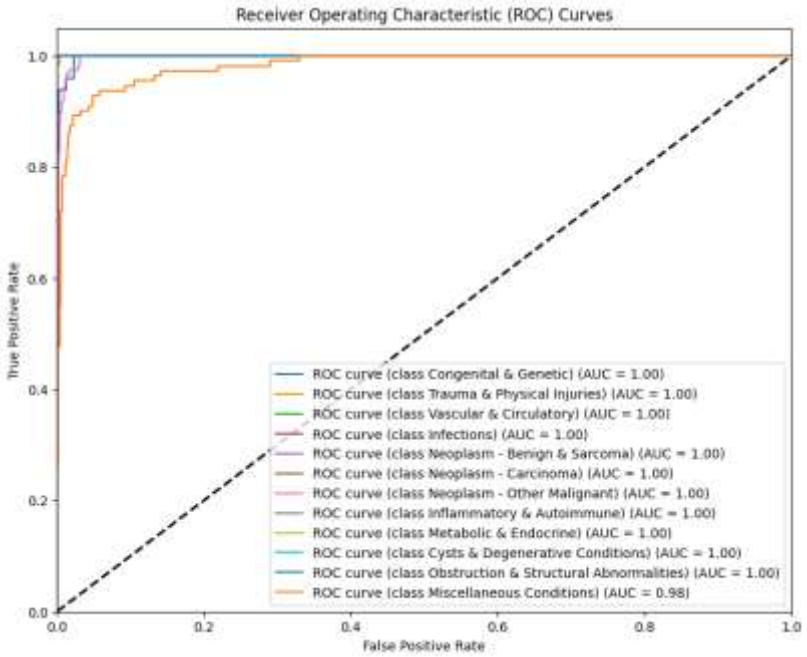

Figure 43: EfficientNet+BiomedBert Weighted Late Fusion AUC-ROC curve

Figure [43] above shows the AUC-ROC curve for all classes, where the AUC value of most classes is closer to 1, indicating a high ability to distinguish between classes.

Table 29: EfficientNet+BiomedBert Weighted Late Fusion Precision, Recall, F1-Score, and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.95	0.94	0.94	0.06
Class 1	0.97	1.00	0.98	0
Class 2	0.97	1.00	0.99	0
Class 3	0.93	1.00	0.96	0
Class 4	0.96	1.00	0.98	0
Class 5	0.97	1.00	0.98	0
Class 6	0.88	0.95	0.91	0.05
Class 7	1.00	1.00	1.00	0
Class 8	0.99	0.98	0.99	0.02
Class 9	1.00	1.00	1.00	0
Class 10	1.00	1.00	1.00	0
Class 11	0.91	0.64	0.75	0.36

Table [29] presents the precision, recall, and F1-Score for each class. Classes 2, 7, 8, 9, and 10 stand out with exceptionally high scores, reflecting the model's strong ability to accurately and confidently identify these cases with minimal false positives or negatives.

4.3.12 EfficientNet+Random Forest(Average Late Fusion)

The model achieved an accuracy of 96.62% with the confusion matrix shown below:

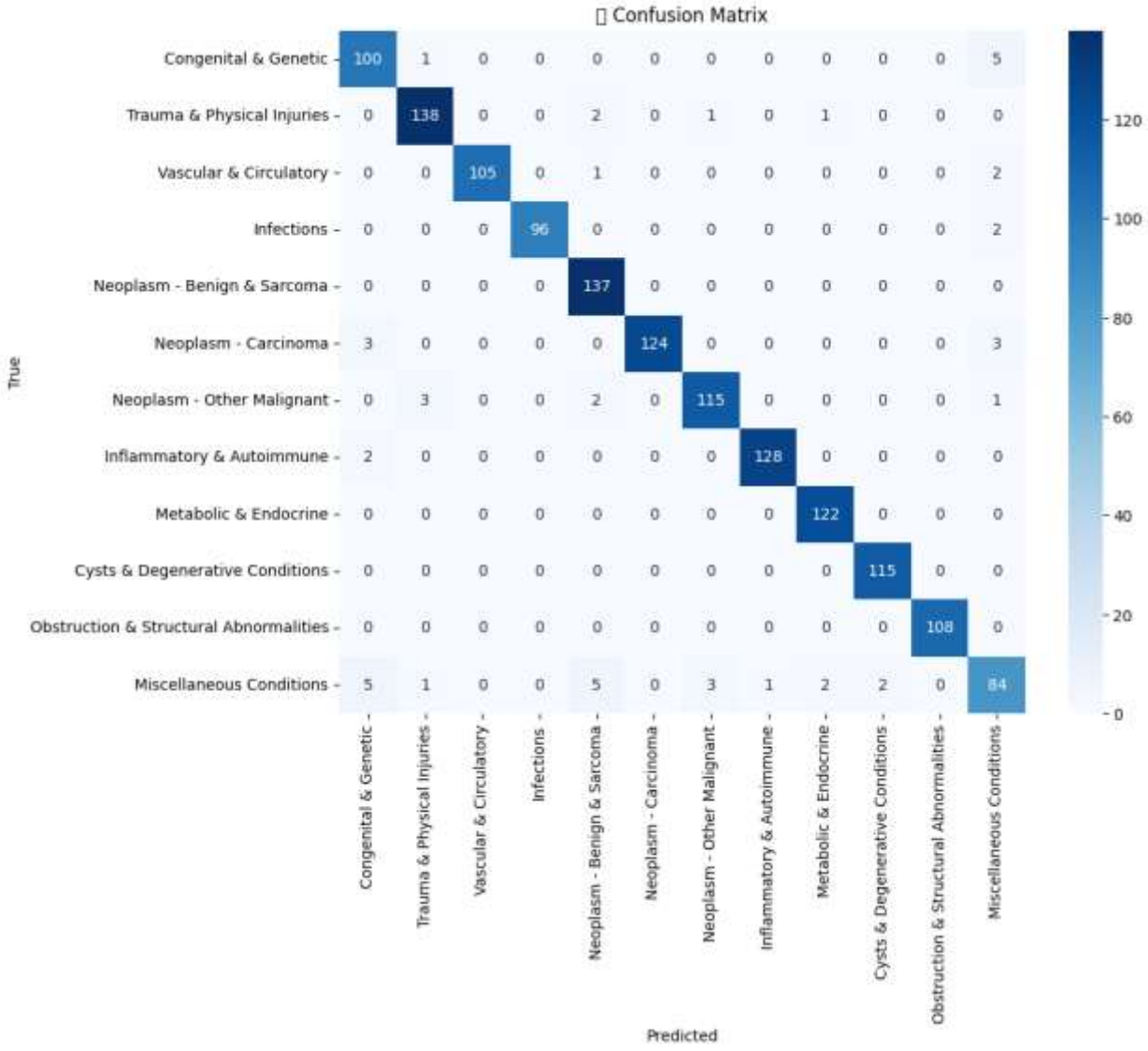

Figure 44: EfficientNet+Random Forest Average Late Fusion confusion matrix

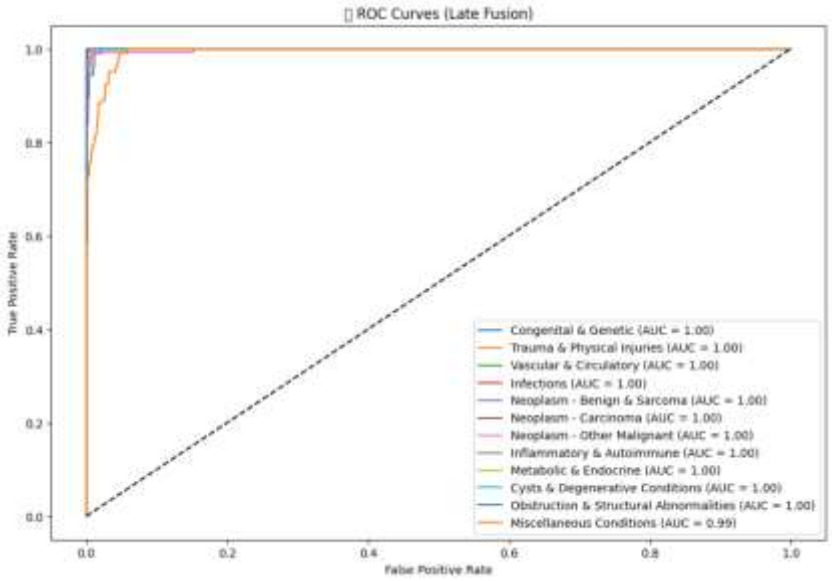

Figure 45: EfficientNet+Random Forest Average Late Fusion AUC-ROC curve

Figure [45] above shows the AUC-ROC curve for all classes, where the AUC value of most classes is closer to 1, indicating a high ability to distinguish between classes.

Table 30: EfficientNet+Random Forest Average Late Fusion Precision, Recall, F1-Score, and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.91	0.94	0.93	0.06
Class 1	0.97	0.97	0.97	0.03
Class 2	1.00	0.97	0.99	0.03
Class 3	1.00	0.98	0.99	0.02
Class 4	0.93	1.00	0.96	0
Class 5	1.00	0.95	0.98	0.05
Class 6	0.97	0.95	0.96	0.05
Class 7	0.99	0.98	0.99	0.02
Class 8	0.98	1.00	0.99	0
Class 9	0.98	1.00	0.99	0
Class 10	1.00	1.00	1.00	0
Class 11	0.87	0.82	0.84	0.18

Table [30] presents the precision, recall, and F1-Score for each class. The model exhibits robust classification performance for almost all of the classes where both precision and recall are consistently high. This suggests the model is able to identify these conditions with high confidence and minimal error.

4.3.13 EfficientNet+TF-IDF Vectorizer (Intermediate Fusion with Cross-Modal Attention)

The model achieved an accuracy of 93.38% with the confusion matrix shown below:

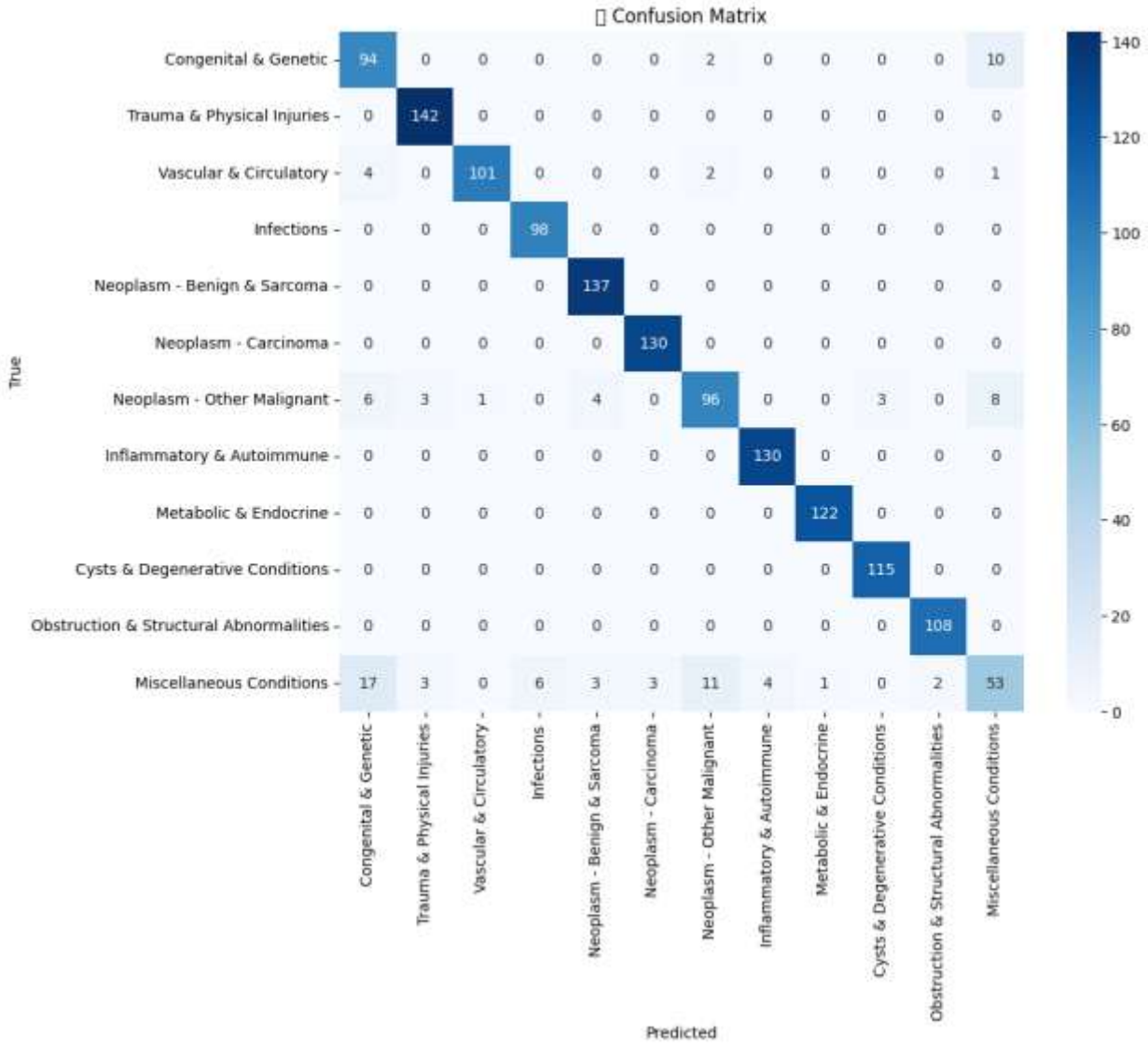

Figure 46: EfficientNet+TF-IDF Vectorizer Intermediate Fusion with Cross-Modal Attention Confusion Matrix

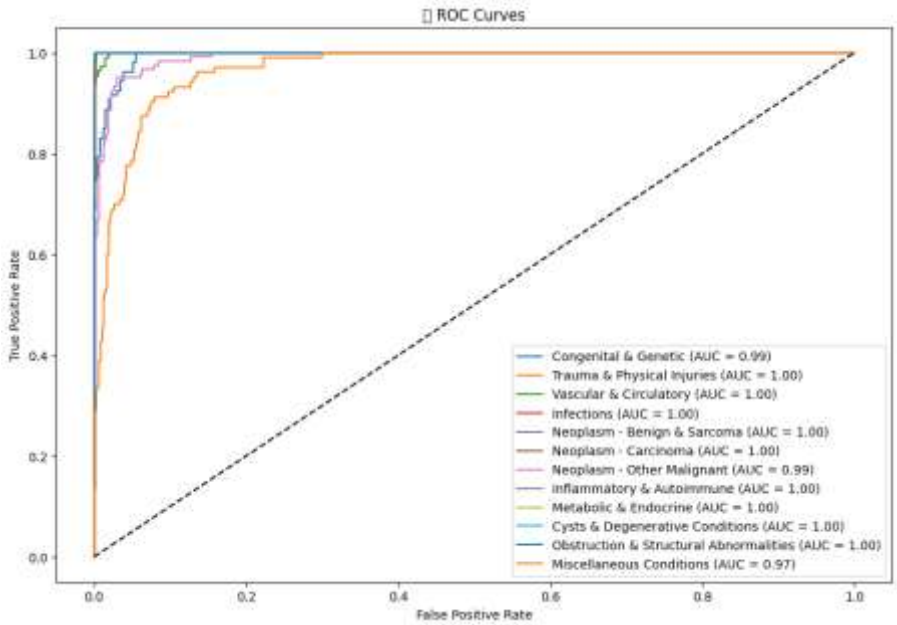

Figure 47: EfficientNet+TF-IDF Vectorizer Intermediate Fusion with Cross-Modal Attention AUC-ROC curve

Figure [47] above shows the AUC-ROC curve for all classes, where the AUC value of most classes is closer to 1, indicating a high ability to distinguish between classes.

Table 31: EfficientNet+TF-IDF Vectorizer Intermediate Fusion with Cross-Modal Attention Precision, Recall, F1-Score, and FN rate for each class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.78	0.89	0.83	0.11
Class 1	0.96	1.00	0.98	0
Class 2	0.99	0.94	0.96	0.06
Class 3	0.94	1.00	0.97	0
Class 4	0.95	1.00	0.98	0
Class 5	0.98	1.00	0.99	0
Class 6	0.86	0.79	0.83	0.21
Class 7	0.97	1.00	0.98	0
Class 8	0.99	1.00	1.00	0
Class 9	0.97	1.00	0.99	0
Class 10	0.98	1.00	0.99	0
Class 11	0.74	0.51	0.61	0.49

Table [31] presents the precision, recall, and F1-Score for each class. The model achieves good accuracy overall but struggles with minority classes and certain ambiguous categories.

4.3.14 EfficientNet+Random Forest(Average Late Fusion Without Miscellaneous Conditions)

The model achieved an accuracy of 96.51% with the confusion matrix shown below:

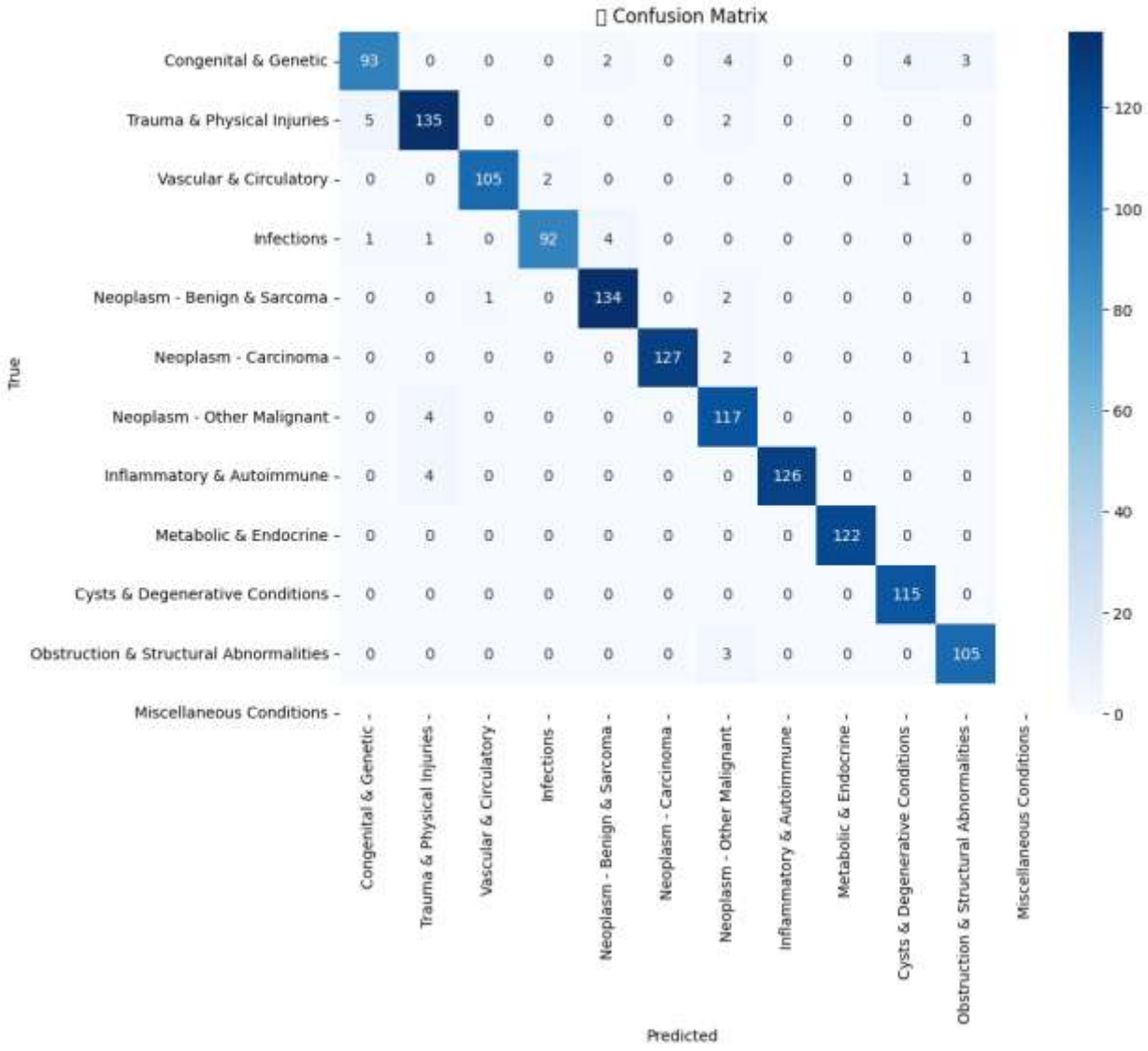

Figure 48: EfficientNet+Random Forest Average Late Fusion Version2 confusion matrix

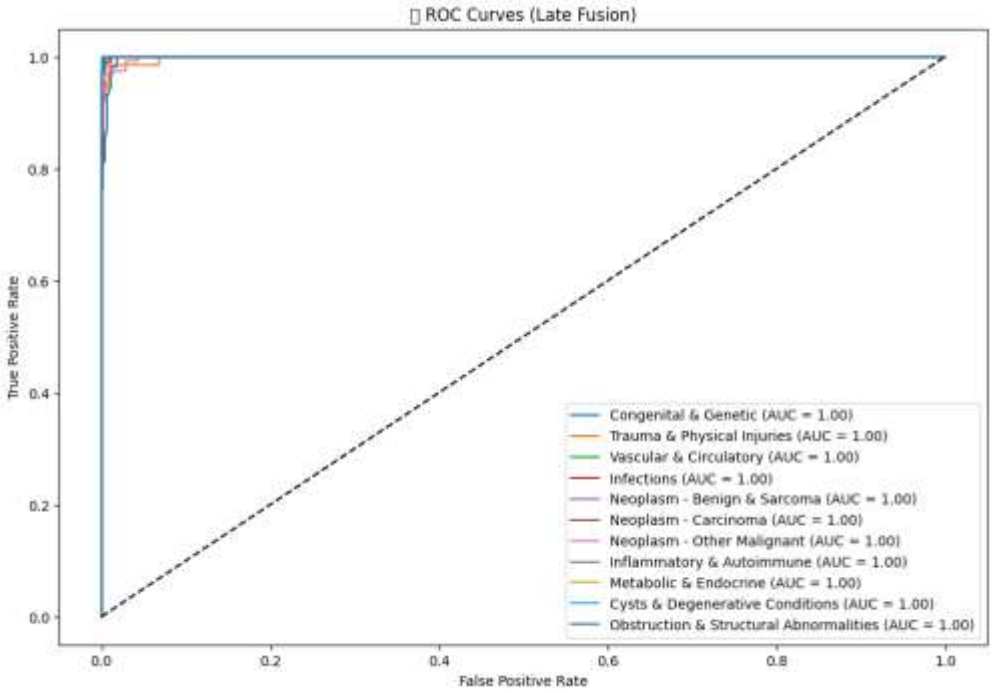

Figure 49: EfficientNet+Random Forest Average Late Fusion Version2 AUC-ROC curve

The new model without the "Miscellaneous Conditions" didn't improve the multimodal classifier since accuracy decreased slightly instead of increasing. Table [32] blow presents the precision, recall, and F1-Score for each class without class 11.

Table 32: EfficientNet+Random Forest Average Late Fusion Version2 Precision, Recall, F1-Score and FN rate per class

Classes	Precision	Recall	F1-Score	False Negative Rate
Class 0	0.94	0.88	0.91	0.12
Class 1	0.94	0.95	0.94	0.05
Class 2	0.99	0.97	0.98	0.03
Class 3	0.98	0.94	0.96	0.06
Class 4	0.96	0.98	0.97	0.02
Class 5	1.00	0.98	0.99	0.02
Class 6	0.90	0.97	0.93	0.03
Class 7	1.00	0.97	0.98	0.03
Class 8	1.00	1.00	1.00	0
Class 9	0.96	1.00	0.98	0
Class 10	0.96	0.97	0.97	0.03

Table [23] below summarizes and compares text models.

Table 33: Multi-modal models summary tables

Model Type	Accuracy	Macro F1-Score	Macro AUC-ROC	Macro FN rate	Training Time (minutes)
ResNet50+DistilBert	94.46%	0.9358	0.9817	0.0642	55
EfficientNet+BiomedBert(Early Fusion)	92.46%	0.9200	0.9959	0.0808	59
EfficientNet+Random Forest(Average Late Fusion)	96.62%	0.9648	0.9992	0.0375	65
EfficientNet+BiomedBert (Average Late Fusion)	96.55%	0.96	0.9987	0.0417	75
EfficientNet+BiomedBert (Weighted Late Fusion)	96.20%	0.96	0.9982	0.0408	75
EfficientNet+TF-IDF Vectorizer(Intermediate Fusion with Cross-Modal Attention)	93.38%	0.9246	0.9958	0.0725	72
EfficientNet+Random Forest(Average Late Fusion) Without Miscellaneous	96.51%	0.9645	1.0000	0.0325	60

Table [34] below summarizes and compares all implemented models; it shows that the models had great performance across both modalities in which accuracies ranged between 88.82% and 96.62%.

Table 34: Summary of all designed models

Modality	Model	Accuracy	Macro F1-Score	Macro AUC-ROC	Macro FN rate
Text	Logistic Regression	88.82%	0.86	0.9811	0.1292
	Random Forests	89.66%	0.84	0.9781	0.1350
	SVM	89.51%	0.89	0.9925	0.1075
	BiomedBert	93.25%	0.92	0.9908	0.0792
Image	EfficientNet-B0	96.06%	0.9550	0.9992	0.0417
	DenseNet121	95.21%	0.9608	0.9992	0.0367
	Resnet50	91.83%	0.9608	0.9992	0.0367
Multi-modal	ResNet50+DistilBert	94.46%	0.9358	0.9817	0.0642
	EfficientNet+BiomedBert(Early Fusion)	92.46%	0.9200	0.9959	0.0808
	EfficientNet+Random Forest(Average Late Fusion)	**96.62%**	**0.9648**	**0.9992**	**0.0375**
	EfficientNet+BiomedBert (Average Late Fusion)	96.55%	0.96	0.9987	0.0417
	EfficientNet+BiomedBert (Weighted Late Fusion)	96.20%	0.96	0.9982	0.0408
	EfficientNet+TF-IDF Vectorizer(Intermediate Fusion with Cross-Modal Attention)	93.38%	0.9246	0.9958	0.0725
	EfficientNet+Random Forest(Average Late Fusion) Without Miscellaneous	96.51%	0.9645	1.0000	0.0325

Table [35] below compares our best model overall (EfficientNet-B0 and RF Late fusion) with other work previously studied.

Table 35: Comparison with other work

Paper	Task	Modalities Used	Key Models	Results
[2]	Predict COVID-19 severity	CT scans, Clinical data	Attention CNN, HoFN, and IMSLoss for fusion	96% accuracy (with IMSLoss); 93% (without)
[5]	Classify lung cancer	Medical imaging, Genomics, Clinical data	MFDNN, PCA, t-SNE	92.5% accuracy; Precision 87.4%, Recall 86.4%, F1-score 86.2%

[6]	Diagnose Alzheimer's disease (AD)	MRI, Gene sequences, Clinical data	Geometric algebra, Feature filtration, CNN-ANN fusion	96.2% accuracy (AD vs NC); 87.4% (sMCI vs MCI)
[8]	Enhance disease diagnosis	Chest X-rays, Clinical data	Late and Intermediate fusion, Pre-trained models	F1-score 94.24% (Intermediate fusion); 91.07% (Late fusion)
[9]	Disease classification	Chest X-rays, Text descriptions	BERT, ResNet-50, Multi-modal attention mechanism	97.72% accuracy; Improved over single modalities
[11]	Predict response to immunotherapy in NSCLC	CT imaging, PD-L1, Genomics	DyAM model, Radiomics, Attention mechanism	AUC 0.80 (multi-modal); 0.61-0.73 (unimodal)
[12]	Diagnose COVID-19 pneumonia	CT scans, X-rays	Transfer learning, Multi-modal networks	99.87% accuracy; Sensitivity 99.74%, Specificity 100%
[13]	Early diagnosis of Heart Failure (HF)	Chest X-rays, Clinical texts	ResNet-152, BioBERT, Shared embedding space	89.02% accuracy; F1-score 88.11%; AUROC 0.9598
Our Proposed Methodology	**Medical Diagnosis Assistant**	**CT scans and MRIs**	**EfficientNet-B0 and RF Late fusion**	**96.62% accuracy, average F1-Score above 0.9**

Table [35] shows that the proposed methodology demonstrates competitive performance when compared to existing multimodal diagnostic systems. Achieving an accuracy of 96.62% and an average F1-score above 0.9, it ranks among the top-performing approaches in the field. While some models report slightly higher accuracy for specific tasks ([12] reports 99.87% for COVID-19 pneumonia diagnosis), our model stands out for its generality, targeting broader diagnostic support across conditions rather than a single disease. The use of late fusion ensures modularity, interpretability, and reduced complexity without compromising accuracy.

4.4 Web Application

To make the diagnostic system accessible and interactive, a web application was developed and deployed. The application integrates three underlying models to support flexible input options: a Random Forest model for text-only inputs, an EfficientNet model for image-only inputs, and a late fusion model combining both modalities when text and image are provided together. This dynamic model selection ensures that the system remains functional and accurate regardless of the type of input received from the user.

The backend was built using Flask and deployed on Render, handling all inference logic and model orchestration. The frontend was developed with React and TypeScript, providing a responsive and user-friendly interface, and is hosted on Vercel. This modular design allows seamless communication between the frontend and backend, enabling real-time predictions and a smooth user experience. The deployed application serves as a practical demonstration of the proposed multimodal diagnostic framework and its adaptability to real-world clinical scenarios.

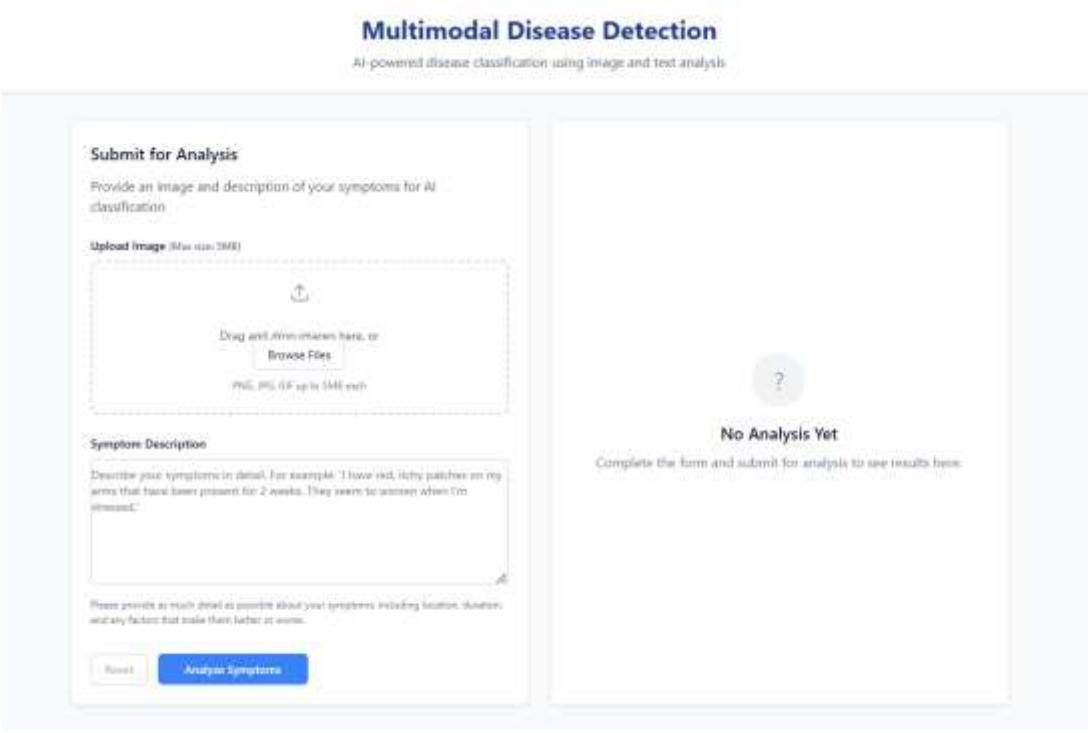

Figure 50: Web Application Page

As shown above, Figure [50] contains the main page of the application, where users can input the images and text and click “Analyze Symptoms” to get the results. Figure 50 shows a sample output of one of the cases found in the dataset.

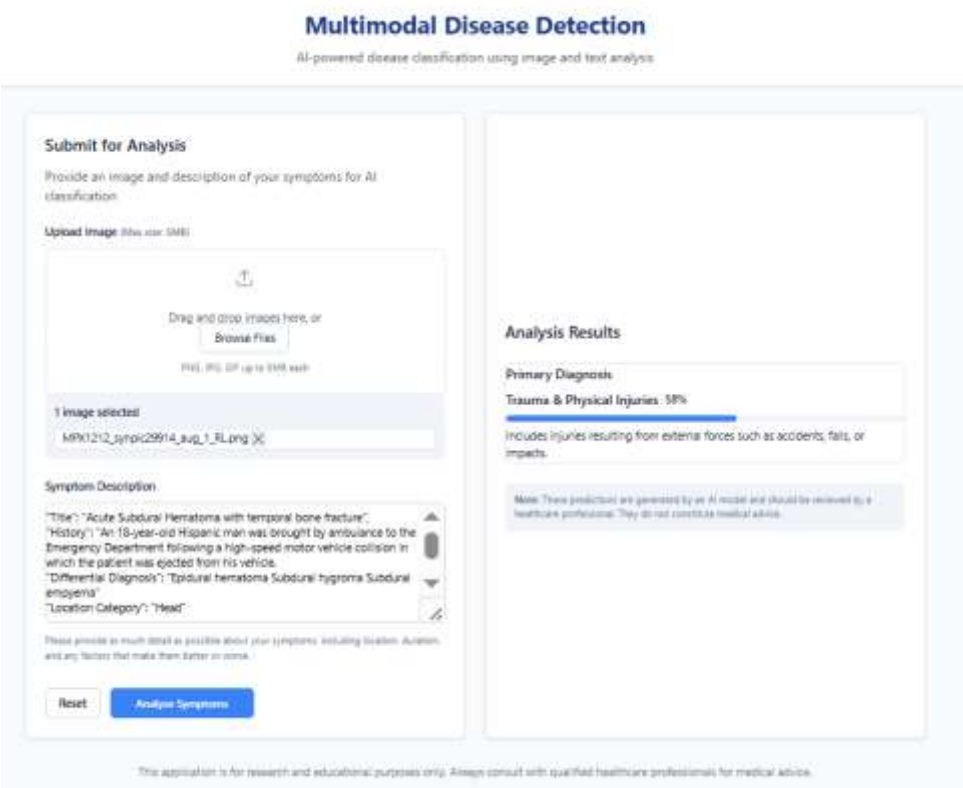

Figure 51: Web Application Analysis Results

4.5 Validation of Design Requirements within the Realistic Constraints

Before building our models, certain goals were set to ensure that the system met its desired requirements. These requirements were based on a set of rules, such as making sure the model was reliable, supporting different image types, augmenting the data to make sure enough samples were collected, and testing various models.

Table 36: Validation of Requirements and Design Constraints

Parameter	Achieved	Explanation
A minimum of 2 ML algorithms should be tested	✓	More than 10 ML/DL models were implemented
The system should support different modalities	✓	The systems utilize both medical images and text
The system shall process medical images from various sources, such as MRI, CT	✓	The system processes both MRI and CT scans
The system shall evaluate performance using a minimum of three metrics	✓	More than 3 evaluation metrics were used to evaluate the performance.
Data augmentation shall be used to improve the balance of the dataset and increase the robustness	✓	Various augmentation techniques were implemented to improve the dataset balance.
The whole system shall not cost more than 200 JDs	✓	The total system cost was 36.46 JDs

5 Conclusion and Future Work

5.1 Conclusion

The problem addressed in this project was a multi-modal machine learning framework for disease diagnosis across patient texts and imaging data, including CT scans and MRIs from the MedPix 2.0 dataset. With such a small dataset size (671 patient records), successful data balancing and augmentation strategies were used to improve model training. By integrating state-of-the-art deep learning models such as EfficientNet for images and Random Forest for text and employing a variety of fusion strategies, we demonstrated that multi-modal learning outperforms unimodal approaches. Data augmentation techniques for both images and text proved essential in addressing class imbalance and enhancing model robustness. Among the developed models, the best-performing configurations achieved over 96% classification accuracy and F1-scores above 0.95, indicating high diagnostic reliability.

The data demonstrate the value of combinatorial approaches to the complex case. Finally, the integration of textual and imaging data not only mimics the diagnostic process of clinicians but also illustrates the possibility of such systems adding to real-world decision-making. Nevertheless, considering the limits of dataset size and diversity, it is demonstrated that the model's generalizability needs to be validated further.

5.2 Future Work

To advance this work, the following general directions are proposed:

- Custom Model Development: Create a model tailored for MedPix2.0 to be even better than all generalized pre-trained models.
- Dataset Expansion: Combine MedPix2.0 with other publicly available datasets to increase the size and diversity of the dataset. This is an important step to enhance the robustness of the model and its applicability to many different populations of patients.
- Validation Across External Data: Running external validation on independent datasets to validate the generalizability and performance of the model in different healthcare settings.

- Incorporating Additional Modalities: You can extend the multi-modal approach by adding other types of data, e.g., lab reports, genetic data, or other imaging, which would bring more accuracy to diagnostic evaluation.
- Real-World Implementation: Pilot the system with healthcare providers and collaborate to gather feedback for further improvements and ensure practical utility.

By addressing these areas, this project has the potential to contribute significantly to the development of scalable, multi-modal diagnostic tools that align with the evolving demands of modern healthcare.

References

[1] Siragusa, I., Contino, S., Massimo, L. C., Alicata, R., & Pirrone, R. (2024, July 3). *MedPix 2.0: A Comprehensive Multimodal Biomedical Dataset for Advanced AI Applications*. arXiv.org. https://arxiv.org/abs/2407.02994

[2] Zhou J, Zhang X, Zhu Z, Lan X, Fu L, Wang H, Wen H. *Cohesive Multi-Modality Feature Learning and Fusion for COVID-19 Patient Severity Prediction*. IEEE Trans Circuits Syst Video Technol. 2021 Mar 4;32(5):2535-2549. doi: 10.1109/TCSVT.2021.3063952. PMID: 35937181; PMCID: PMC9280852.

[3] Jiayuan Zhu, Hui Liu, Xiaowei Liu, Chao Chen, Minglei Shu, *Cardiovascular disease detection based on deep learning and multi-modal data fusion*, Biomedical Signal Processing and Control, Volume 99,2025,106882, ISSN 1746-8094.

[4] Bo Yu, Hechang Chen, Chengyu Jia, Hongren Zhou, Lele Cong, Xiankai Li, Jianhui Zhuang, Xianling Cong, *Multi-modality multi-scale cardiovascular disease subtypes classification using Raman image and medical history*, Expert Systems with Applications, Volume 224, 2023, 119965, ISSN 0957-4174

[5] Sangeetha S.K.B, Sandeep Kumar Mathivanan, P Karthikeyan, Hariharan Rajadurai, Basu Dev Shivahare, Saurav Mallik, Hong Qin, *An enhanced multimodal fusion deep learning neural network for lung cancer classification*, Systems and Soft Computing, Volume 6, 2024, 200068, ISSN 2772-9419

[6] Yue Tu, Shukuan Lin, Jianzhong Qiao, Yilin Zhuang, Peng Zhang, *Alzheimer's disease diagnosis via multimodal feature fusion*, Computers in Biology and Medicine, Volume 148, 2022, 105901, ISSN 0010-4825,

[7] Elisa Warner, Joonsang Lee, William Hsu, Tanveer Syeda-Mahmood, Charles E. Kahn Jr., Olivier Gevaert, Arvind Rao, *Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects*, International Journal of Computer Vision, Volume 132, 2024, Pages 3753-3769, DOI: 10.1007/s11263-024-02032-8.

[8] Kumar, Sachin, et al. "*Deep-Learning-Enabled Multimodal Data Fusion for Lung Disease Classification*." Informatics in Medicine Unlocked, vol. 42, 1 Jan. 2023, pp. 101367–101367, https://doi.org/10.1016/j.imu.2023.101367.

[9] Wan, ZhengYu & Shao, XinHui. (2023). *Disease Classification Model Based on Multi-Modal Feature Fusion*. IEEE Access. PP. 1-1. 10.1109/ACCESS.2023.3252011.

[10] Baltrušaitis T, Ahuja C, Morency L-P. *Multimodal Machine Learning: A Survey and Taxonomy*. International Journal of Computer Vision. 2019 Nov;127(1):5-39. Doi: 10.1007/s11263-018-1129-8.

[11] Vanguri RS, Luo J, Aukerman AT, Egger JV, et al. *Multimodal integration of radiology, pathology, and genomics for predicting response to PD-(L)1 blockade in patients with non-small cell lung cancer*. Nature Cancer. 2022 Oct; 3(10):1151–1164. doi: 10.1038/s43018-022-00416-8.

[12] N. Hilmizen, A. Bustamam and D. Sarwinda, "*The Multimodal Deep Learning for Diagnosing COVID-19 Pneumonia from Chest CT-Scan and X-Ray Images*," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2020, pp. 26-31, doi: 10.1109/ISRITI51436.2020.9315478. keywords: {COVID-19;Lung;X-ray imaging;Biomarkers;Training;Feature extraction;Viruses (medical);Concatenate;COVID-19;CT-Scan;Multimodal;Pneumonia;Transfer Learning;X-Ray},

[13] Y. Lu, C. Zhang, and F. Tang, "ResBioBERT: *Deep learning combined with multimodal data for heart failure diagnosis*," 2024 16th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 2024, pp. 144-147, doi: 10.1109/IHMSC62065.2024.00040. keywords: {Fuses;Biological system modeling;Human-machine systems;Medical services;Radiology;Feature extraction;Cardiovascular diseases;History;X-ray imaging;Medical diagnostic imaging;heart failure;multimodal;chest X-ray;clinical text},

[14] *AI*. (n.d.). https://bepartofresearch.nihr.ac.uk/articles/artificial-intelligence/#:~:text=The%20use%20of%20artificial%20intelligence,UK%20to%20do%20this%20work.

[15] *ISO/IEC TR 24029-1:2021*. (n.d.). ISO. https://www.iso.org/standard/77609.html