Assessing the Reliability of AI-Generated Medical Images: A Comparative Study with Real X-ray, CT, and MRI Scans

By
Md Muhetul Islam
213-15-4448

FINAL YEAR DESIGN PROJECT REPORT

This Report is Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and Engineering, Daffodil International University

Co-Supervised by

Mr. Abdus Sattar
Assistant Professor
Department of Computer Science and Engineering, Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September 16, 2025

APPROVAL

This Project titled **"Assessing the Reliability of AI-Generated Medical Images: A Comparative Study with Real X-ray, CT, and MRI Scans"**, submitted by **Md Muhetul Islam**, ID No: **213-15-4448** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **16 September 2025**.

BOARD OF EXAMINERS

Dr. S.M Aminul Haque — **Chairman**
Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md Alamgir Kabir — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Md Assaduzzaman — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zulfiker Mahmud — **External Examiner**
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Sazzadur Ahamed, Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and
Engineering, Daffodil International University

Co-Supervised by:

Mr. Abdus Sattar
Assistant Professor
Department of Computer Science and
Engineering, Daffodil International University

Submitted by:

Muhet 4.9.25

Md Muhetul Islam
Student ID: 213-15-4448
Department of Computer Science and
Engineering, Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the Almighty for His divine blessing, making it possible for us to complete the **FINAL YEAR DESIGN PROJECT (FYDP)** successfully**.**

We are grateful and wish to express our profound indebtedness to **Md. Sazzadur Ahamed, Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Image classification**. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering for his kind help in finishing our project, and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire coursemates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

This paper presents a deep learning-based approach to differentiate between real medical images from AI-generated counterparts across X-ray, CT, and MRI images, addressing a critical challenge in healthcare diagnostics. Utilizing transfer learning with pre-trained models like InceptionV3, ResNet50, DenseNet121, VGG19, and MobileNetV2 on a 3,000 image dataset with 500 per class, the study achieves a peak accuracy of 0.82 with MobileNetV2, featuring an F1-score of 0.97 and near-perfect recall 1.00 for AI_MRI. InceptionV3 records 0.88, deeper models like ResNet50, DenseNet121, and VGG19 have accuracies below 0.60. It is relevant to engineering and medical standards to improve the reliability of the diagnosis. The framework follows IEEE 1012-2016 and DICOM standards. Limitations include the dataset's small size, leading to overfitting, like 0.9766 training vs. 0.8833 validation accuracy for InceptionV3. Future work proposes dataset expansion to 10,000 images, adversarial training, and edge device optimization to enhance diagnostic reliability and scalability in healthcare.

Table of Contents

Approval i

Declaration ii

Acknowledgements iii

Abstract iv

List of Figures viii

List of Tables ix

1 Introduction 1
1.1 Introduction 1
1.2 Motivation 2
1.3 Objectives 3
1.4 Methodology 4
1.5 Project Outcome 4
1.6 Organization of the Report 4

2 Background 6
2.1 Introduction 6
2.2 Literature Review 7
2.3 Gap Analysis 11
2.4 Summary 14

3 Research Methodology 16
3.1 Methodology 16
3.1.1 Overview 16
3.1.2 Proposed Methodology 17
3.1.2.1 Detailed Architecture……...19
3.1.2.2 Rationale for Selection……..20
3.1.3 Functional and Nonfunctional Requirements 21
3.1.3.1 Functional Requirements……21
3.1.3.2 Nonfunctional Requirements……..22
3.1.4 Data Flow Diagram 22
3.1.5 UI Design……..24
3.2 Detailed Methodology and Design……..24

3.2.1 Alternative Solutions Considered……24
3.2.2 Model Selection……..25
3.2.3 Implementation Details……26
3.2.3.1 Data Acquisition……..26
3.2.3.2 Preprocessing……..31
3.2.3.3 Training Process……..33
3.2.3.4 Model Evaluation……33
3.2.4 Comparative Analysis……35
3.3 Project Plan 35
3.3.1 Project Timeline and Phases……35
3.3.2 Risk Management and Mitigation……37
3.3.3 Tools and Technologies……..37
3.3.4 Deliverables……37
3.4 Task Allocation 38
3.5 Summary 38

4 Implementation and Results 40
4.1 Environment Setup 40
4.1.1 Hardware Configuration……40
4.1.2 Software and Libraries……...41
4.1.3 Dataset Configuration……41
4.1.4 Environment Initialization……42
4.1.5 Challenges and Mitigations……42
4.1.6 Validation of Setup……..43
4.2 Comparative Analysis 43
4.2.1 Models Architecture……...43
4.2.2 Training Hyperparameters……44
4.2.3 Evaluation Metrics……..44
4.2.4 Detailed Metrics by Model……45
4.3 Results and Discussion 59
4.3.1 Findings and Performance Overview……..59
4.3.2 Analysis of Accuracy Variations……...60
4.3.3 Modality-Specific Performance……60
4.3.4 Weaknesses and Misclassification Analysis……62
4.3.5 Overfitting and Underfitting Considerations……...62
4.3.6 Robustness Evaluation……..63
4.3.7 Statistical Significance……...63
4.3.8 Clinical Relevance and Future Directions……..……63
4.3.9 Computational Complexity and Runtime……..…….64
4.3.10 Comparative Results……..……...64
4.4 Summary 65

5 Engineering Standards and Design Challenges 66
5.1 Compliance with the Standards 66
5.1.1 Software Standards 66

5.1.2 Hardware Standards 67
5.1.3 Communication Standards 68
5.2 Impact on Society, Environment and Sustainability 69
5.2.1 Impact on Life 69
5.2.2 Impact on Society & Environment 69
5.2.3 Ethical Aspects 69
5.2.4 Sustainability Plan 70
5.3 Project Management and Financial Analysis 70
5.3.1 Project Management……..70
5.3.2 Cost Analysis……...70
5.3.3 Revenue Model……72
5.3.4 Financial Feasibility and Risks……..72
5.4 Complex Engineering Problem 72
5.4.1 Complex Problem Solving 72
5.4.2 Engineering Activities 74
5.5 Summary 74

6 Conclusion 76
6.1 Summary ..76
6.2 Limitation ..76
6.3 Future Work ..77

References 78

List of Figures

Figure 3.1 Proposed methodology…….. 18
Figure 3.2 MobileNetV2 Architecture……..19
Figure 3.3 Real Chest X-ray Images…….28
Figure 3.4 Real Brain MRI Images……....28
Figure 3.5: Real CT Scan Images……...29
Figure 3.6 AI-Generated X-Ray Images……30
Figure 3.7 AI-Generated CT Scan Images……30
Figure 3.8 AI-Generated MRI Images……31

Figure 4.1 Training vs. Validation Accuracy and Loss curves of MobileNetV2……45
Figure 4.2 Tabular Classification Report Results of MobileNetV2……46
Figure 4.3 Confusion Matrix visualization for six classes of MobileNetV2……..47
Figure 4.4 Training vs. Validation Accuracy and Loss curves for InceptionV3……48
Figure 4.5 Tabular Classification Report Results for InceptionV3……49
Figure 4.6 Confusion Matrix visualization for six classes for InceptionV3……..50
Figure 4.7 Training vs. Validation Accuracy and Loss curves for DenseNet121……..51
Figure 4.8 Tabular Classification Report Results for DenseNet121……..52
Figure 4.9 Confusion Matrix visualization for six classes for DenseNet121……....53
Figure 4.10 Training vs. Validation Accuracy and Loss curves for ResNet50…….54
Figure 4.11 Tabular Classification Report Results for ResNet50…….55
Figure 4.12 Confusion Matrix visualization for six classes for ResNet50……...56
Figure 4.13 Training vs. Validation Accuracy and Loss curves for VGG19…….57
Figure 4.14 Tabular Classification Report Results for VGG19…….58
Figure 4.15 Confusion Matrix visualization for six classes for VGG19……...59

List of Tables

2.1 Summary of Literature Reviewed……8
2.2 Feature Comparison Between Existing Systems and the Proposed System……13

3.1 Dataset Composition……31
3.2 Preprocessing Steps Summary……32
3.3 Comparison of Considered CNN Architectures……35

4.1 Model Performance Comparison……65

5.1 Cost Analysis of System Development……...71
5.1 Mapping with complex problem solving……....73
5.2 Mapping with a Knowledge Profile……....73
5.3 Mapping with complex engineering activities……...74

Chapter 1

Introduction

This chapter outlines the background, motivation, problem statement, objectives, and scope of the research, setting the foundation for understanding the significance of detecting AI-generated medical images.

1.1 Introduction

Artificial intelligence (AI) has become a transformative technology in medical imaging, reshaping how medical professionals diagnose, treat, and manage diseases in patients. At the center of this paradigm shift are deep generative models, like Generative Adversarial Networks (GANs) and diffusion models, which have shown the ability to create ultra-realistic synthetic medical images [1]–[3]. With their ability to generate synthesized datasets from X-rays, CT and MRI scans, these models can target pivotal issues that create barriers towards usable data, such as data unavailability and class imbalance [1]. The AI-generated images present an opportunity to enhance current datasets, thereby refining machine learning models for clinical research and applications. This advancement paves the way for innovative approaches to data-driven decision-making in the medical field.

However, there are now significant concerns over the integrity and validity of medical imaging datasets. As a result of these developments is important. Synthetic pictures pose a danger to the reliability of AI in clinical settings if they are not properly managed. This has sparked the justifiable worry that artificial intelligence-generated images can unintentionally be incorporated into real photographs, resulting in incorrect diagnosis and unsuitable clinical judgments. A scenario like this emphasizes how critical it is to protect data integrity and authenticity from ever-increasing dangers [3], [4], [7]. The inverse challenge of a detection process that separates synthetic images from real images in clinical datasets has not been as thoroughly studied as the forward problem of producing synthetic images to enhance models [4], [5].

AI generated images have attained such a high visual quality that they can nowadays most often not be distinguished from real images by the human eye, rendering manual inspection

impractical and unreliable. While this improvement in image quality carries enormous potential for enhancing professional and scientific work, it equally demands that increasingly sophisticated automated detection systems be developed to detect with reliability and efficiency such synthetically generated images. This system are essential to protect the quality of medical datasets, which contributes to increasing trust in AI-driven technologies in healthcare [5]–[7].

This study addresses the issue of real vs. AI-based medical imaging by developing an improved deep learning classification algorithm. This is capable of separating input data into real and fraudulent categories. This distinction is crucial for ensuring that health care machine learning models are trained on genuine data. For this, these are acceptable ethical procedures, as well as reducing the possibility of diagnosing errors and providing incorrect information [3]. In this study, we aim to address this pressing issue by subjecting a large sample of computed X-ray, CT, and MRI scans to a single-blind, comprehensive comparative investigation of actual versus AI-generated images. I employed advanced CNNs for binary classification, which have demonstrated broad capabilities in numerous domains of medical image analysis [7],[8].

My research addresses a significant deficiency in the existing literature and aligns with a larger goal of enhancing data transparency as well as promoting the ethical use of AI technologies in healthcare. This study will aid in the advancement of verifiable AI-based image processing solutions, making clinically relevant medical imaging technologies more applicable, reliable, and trustworthy, thereby facilitating their adoption in modern healthcare settings. Through this endeavor, we aim to establish a foundation for future advancements and enhancements in AI-supported diagnosis and treatment in clinical practice.

1.2 Motivation

Artificial intelligence (AI), which encompasses a variety of technologies, is most useful in the field of medical imaging. It is a constantly developing discipline that can enhance doctor's diagnostic abilities and enhance patient outcomes. Data imbalance and scarcity issues in medical datasets are prevalent in practice and can be concretely addressed by models such as diffusion models and Generative Adversarial Networks (GANs), which can produce synthetic medical images [1], [2], [6]. For the enhancement of diagnostic tools and procedures, these enable researchers to create more reliable machine learning-based models [1, 5, 6].

Despite its many benefits, creating images that are identical to the genuine thing is a difficult

computational operation. How artificial imagery may lead to incorrect clinical judgments and misdiagnosis. Synthetic pictures may threaten data integrity and lead to incorrect diagnoses and clinical choices if they are incorporated into clinical databases without being properly identified and explained to readers. Tools that can successfully differentiate between real and artificial intelligence-generated medical images are thus urgently needed [5]–[7]. If we wish to maintain the integrity and legitimacy of health care AI applications, we must fight this problem.

Finally, resolving this is not only a technological challenge. It is the first step toward enabling the safe, moral, and successful application of AI technologies in therapeutic settings. Our approach enables a deep learning-based classification model that can differentiate between real and AI images. This results in the creation of trustworthy medical datasets. By doing this, healthcare data ethics will be upheld while also enhancing the accuracy and functionality of AI-based diagnostic tools [3], [4], [6].

In my opinion, this issue needs to be resolved so that the scientific community and medical practitioners may use data in a more trustworthy and transparent way. This will improve patient care. Also, for this, results by strengthening the responsible incorporation of emerging technologies into clinical practice and fostering more trust in AI-based diagnostic tools. By enhancing the authenticity and legitimacy of medical AI solutions, my work is helping to create a more promising future for them.

1.3 Objectives

The rapid generation of these AI images for medical use cases, especially using GAN-based models for image Generation. This brought us to the idea of deepfake jamming to ensure the stability and authenticity of all Medical datasets. In response to this challenge this research aims to:

i. Train a deep learning model for classifying X-ray, CT and MRI images to authenticate whether they are real or AI-generated.

ii. Authenticate the dataset to help AI systems to prevent the abuse of synthetic medical images.

Pursuing these goals will help make AI for medical diagnostics more trustworthy and fair.

1.4 Methodology

In this work, a deep learning-based approach is used to distinguish real (medically acquired) images from AI (artificially-generated) images for three medical modalities: X-ray, CT, and MRI. We collected from public repositories a custom dataset consisting of 3,000 images, half of which are real scans, and the other half simulated scans. Specialized medical image formats (.nii, .gz) were converted to open-source images via preprocessing to standardized, which are read from two different folders, resizing and normalizing the samples and arranging them into directories grouped with their respective classes. We used the MobileNet convolutional neural network architecture due to its efficient setup with high accuracy in image classification tasks. For training and validation, Python libraries used on Google Colab and Kaggle. Accuracy, precision, recall, and F1-score were calculated to assess the model's classification performance. Data augmentation and regularization techniques were used to address several issues, like overfitting due to unusually high precision.

1.5 Project Outcome

This project's primary outcome is a robust classification algorithm that can distinguish between real and AI-generated medical photos. Also need good accuracy, protecting the AI Image Datasets from contamination. By preventing the unintentional or deliberate use of artificial data in clinical judgments, the system aims to provide a more trustworthy medical AI tool. Along with providing an overview of medical picture forensics, this study opens the door for future automated content authenticity verification in healthcare.

1.6 Organization of the Report

The report is divided into six chapters in order to present the research work in a systematic manner.

Chapter 1: Introduction

In this chapter, we introduce the research topic and the background information related to AI-generated medical imaging, its applications and challenges, and the aims of the study.

Chapter 2: Background

In this chapter, we will summarise the required background knowledge, from a thorough literature survey on the field of synthetic medical image generation and detection. Against it, it summarizes relevant research and closely related applications and draws out gaps that motivate this work.

Chapter 3: Research Methodology

In this section, the methodology is introduced in detail from collecting dataset up to the implementation of the model design. Planning for the projects, allocation of tasks, and other methods are explained.

Chapter 4: Implementation and Results

This section depicts the experimental setup, procedure, and classification model evaluation. This includes an analysis and performance measurements and a discussion of its results.

Chapter 5: Engineering Standards and Design Challenges

Mandatory compliance with sections of relevant software, hardware, and communication standards. This also includes the social impact, ethics, and sustainability of the project, project management, and the budget and financial analysis. It maps and rationalizes the solutions to complex engineering problems.

Chapter 6: Conclusion

The last chapter discusses the research results, addresses limitations and suggests future directions.

Chapter 2

Background

This chapter provides the necessary background knowledge and a comprehensive literature review to contextualize the research. It also presents an analysis of existing work and identifies the gaps that this study aims to address.

2.1 Introduction

Artificial intelligence (AI) technology in medical imaging is changing the healthcare diagnostic landscape. Recent deep generative models like GANs and diffusion models have enabled the synthesis of realistic medical images [1], [2], [6]. By offering well-grounded synthetic images to supplement current datasets. Also, enhance performance through improved training of machine learning models for use in clinical settings. As these advancements address major issues with data imbalance and sparseness [15], [22].

AI-generated images have a variety of applications in medical imaging. It is used in data augmentation, anonymization, and improving AI model training [1]. In these capacities, the use of synthetic images advances the development of trustworthy diagnostic tools that aid in the comprehension and management of various illnesses. However, despite all of these advancements, a significant problem with the legitimacy and validity of the medical picture collections still exists.

The integrity of the clinical data could be seriously jeopardized because synthetic data are indistinguishable from real data. Also, adding AI-generated photos to clinical databases can result in incorrect diagnoses. It causes a loss of trust in AI-powered medical applications [3], [4]. Therefore, methods for appropriately identifying and differentiating between synthetic and real images must be developed.

For data limitations, the literature has traditionally concentrated on image synthesis [4], [5]. On the other hand, the literature has only lately begun to emphasize how crucial it is to identify these images when they appear embedded in clinical statistics. Additionally, the visual quality of AI-generated images is improving than before. This makes manual detection efforts impractical. For this reason, automated detection methods are crucial [5], [16], [17].

The development of AI-based medical imaging, both historically and recently, will be covered in this chapter. This emphasizes how crucial it is to develop new methods for both detecting

and confirming the legitimacy of medical datasets used to create synthetic images. Through a thorough literature review, this chapter provides the background information needed to comprehend the significance of the research topic. Also, provide the suggested solutions to maintain data integrity from the use of AI-based medical applications.

2.2 Literature Review

The medical imaging industry has been a great location for quickly establishing the role of artificial intelligence (AI). Mostly in recent times due to recent developments in synthetic image-generating techniques. The findings of our survey demonstrate the predictive ability of generative models, like GANs, as well as the related implications for data ethics and quality. After that, I provide a summary and a discussion of the major studies that have shaped this field's research and practice.

Generative Adversarial Networks (GANs), introduced by Goodfellow. Goodfellow et al. [6] in 2014, represented a significant step forward in synthesizing images. Frid-Adar et al. [1] (2018) demonstrated the efficacy of GAN-based augmentation. It significantly enhances CNN performance in liver lesion classification. Also, Chen et al. [2] (2019) utilized the F&F GAN model to improve robustness and accuracy in the same domain. Perez et al. [15] (2017) further validated the role of data augmentation. Perez shows its substantial impact on classification accuracy. In recent times, El Kojok et al. [22] (2025) integrated VAEs, GANs, and transfer learning to augment spine CT scan datasets. It enhances spinal fracture detection. Finally, Ali et al. [23] (2022) explored GAN and AI applications for COVID-19 medical imaging through a scoping review.

Medical image processing has benefited greatly from architectural advancements. MobileNets, an effective CNN designed for mobile vision. It was proposed by Howard et al. [7] (2017). Sandler et al. [8] (2018) improved upon this with MobileNetV2's inverted residuals and linear bottlenecks. Now, both of which worked well for real-time applications. Dai and Gao [20] (2021) used transformers in a multi-modal method (TransMed) to improve classification. Besides, Huang et al. [19] (2022) improved breast cancer detection with a bilinear structure (BM-Net) based on MobileNetV3. The architectural landscape was expanded by Saquib et al. [24] (2025) with their introduction of a MedNet-MoBiL hybrid model. Finally, Xu et al. [9] (2014) with the enhancement of feature representation through multiple instance learning.

Goebel et al. [16] (2020) and McCloskey et al. [17] (2018) explored the detection and

localization of GAN-generated images, employing techniques like color cue analysis. Singh et al. [5] (2023) developed a CNN with an attention mechanism for high-accuracy detection, emphasizing reliability, while Joyce et al. [4] (2023) surveyed threats and countermeasures. Arshed et al. [21] (2024) addressed financial fraud by detecting counterfeit images, and Lee et al. [18] (2023) analyzed GAN artifacts' impact on mammogram simulation, highlighting detection challenges.

Critical datasets from Kaggle [10-12] (2018-2020) for pneumonia, brain tumors, and COVID-19 CT scans have supported model training, complemented by CSIRO's [14] (2023) synthetic data toolkit. Surveys by Satapathy et al. [13] (2020) and Rana et al. [3] (2023) reviewed deep learning techniques and ethical concerns, respectively, with the latter stressing the need for authentic data preservation.

The authors point out the limitations of the existing dataset. It includes its small size, vulnerability to intrusions, and moral dilemmas. This serves as the impetus for the current MobileNetV2 and transfer learning study. The literature has a solid framework that aids in the development of a repeatable diagnostic technique for medical imaging authenticity.

Table 2.1: Summary of Literature Reviewed

Author(s)	Year	Title	Methodology	Key Findings
H. Frid-Adar et al.	2018	GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance [1]	GAN augmentation, CNN classification	Showed significant improvement in CNN performance for liver lesion classification using GAN-augmented data.
J. Chen et al.	2019	Synthetic Medical Images Using F&F GAN for Improved Liver Lesion Classification [2]	F&F GAN model	Demonstrated enhanced robustness and accuracy in liver lesion classification through synthetic data.
R. A. Rana et al.	2023	AI-based Synthetic Medical Image Generation: Challenges and Future Directions [3]	Review	Addressed ethical concerns and emphasized the need for robust detection to preserve data authenticity.
A. R. Joyce et al.	2023	Deepfake Detection in Medical Imaging: Threats and	Survey	Suggested countermeasures for reliable detection of

		Countermeasures [4]		deepfake medical images.
K. K. Singh et al.	2023	Medical Image Deepfake Detection Using CNN with Attention Mechanism [5]	CNN with attention mechanism	Developed a high-accuracy detection model highlighting the importance of reliability in medical imaging.
I. Goodfellow et al.	2014	Generative Adversarial Nets [6]	Introduction of GANs	Introduced the foundational GAN framework critical for synthetic image generation.
A. Howard et al.	2017	MobileNets: Efficient CNNs for Mobile Vision Applications [7]	MobileNets architecture	Proposed an efficient architecture suitable for real-time medical image processing in mobile applications.
M. Sandler et al.	2018	MobileNetV2: Inverted Residuals and Linear Bottlenecks [8]	MobileNetV2 architecture	Improved classification performance in medical imaging through architectural advancements.
Y. Xu et al.	2014	Deep Learning of Feature Representation with Multiple Instance Learning [9]	Multiple instance learning	Enhanced feature representation to boost medical image classification.
Kaggle	2018	Chest X-ray Images (Pneumonia) [10]	Dataset	Provided a critical dataset supporting pneumonia detection model training.
Kaggle	2018	Brain MRI Images for Brain Tumor Detection [11]	Dataset	Supplied essential data for brain tumor detection model development.
Kaggle	2020	CT Scan (COVID-19) Dataset [12]	Dataset	Offered a COVID-19 CT scan dataset, enhancing AI research capabilities.

T. K. Satapathy et al.	2020	Survey of Deep Learning Techniques for Medical Image Classification [13]	Survey	Reviewed advances in deep learning methods for medical image classification.
CSIRO	2023	Medical Imaging Synthetic Data Toolkit [14]	Toolkit	Developed tools facilitating synthetic medical data generation and evaluation.
L. Perez et al.	2017	The Effectiveness of Data Augmentation in Image Classification Using Deep Learning [15]	Data augmentation	Highlighted data augmentation's significant impact on improving model accuracy.
M. Goebel et al.	2020	Detection, Attribution, and Localization of GAN-Generated Images [16]	Detection methods	Investigated challenges in localizing GAN-generated images to aid detection strategies.
S. McCloskey et al.	2018	Detecting GAN-generated Imagery Using Color Cues [17]	Color cue analysis	Introduced novel color-based techniques for detecting GAN-generated images.
J. Lee et al.	2023	Impact of GAN Artifacts for Simulating Mammograms [18]	Artifact analysis	Evaluated the effects of GAN artifacts on cancer detection accuracy.
J. Huang et al.	2022	BM-Net: CNN-based MobileNet-V3 and Bilinear Structure for Breast Cancer Detection [19]	CNN and MobileNet-V3 architecture	Achieved improved breast cancer detection using an innovative bilinear network structure.
Y. Dai and Y. Gao	2021	TransMed: Transformers Advance Multi-Modal Medical Image Classification [20]	Transformers and a multi-modal approach	Advanced medical image classification leveraging transformers and multi-modal data.

M. A. Arshed et al.	2024	A Deep Learning Model for Detecting Fake Medical Images to Mitigate Financial Fraud [21]	Fraud detection model	Proposed a deep learning model to detect counterfeit medical images and prevent related financial fraud.
Z. El Kojok et al.	2025	Augmenting a Spine CT Scans Dataset Using VAEs, GANs, and Transfer Learning [22]	VAE, GANs, transfer learning	Improved spinal fracture detection through augmented datasets using advanced learning techniques.
A. H. Ali et al.	2022	Combating COVID-19 Using GANs and AI for Medical Images [23]	Scoping review	Explored GAN and AI applications in medical imaging for COVID-19 management.
S. M. Saquib et al.	2025	Medicine Image Classification Using Deep Learning [24]	MedNet-MoBiL hybrid model	Demonstrated advancements in medical image classification through a hybrid deep learning model.

A comprehensive assessment of both AI technologies, as well as possible vulnerabilities and enabling instruments for medical advancements. It is provided by the examined papers taken together. While models like GANs continue to push the boundaries of what pictures can do and improve diagnostic results, they also raise serious concerns about ethical use and accurate data. This work highlights the necessity of continual innovation. In addition to the establishment of frameworks that will protect the integrity of medical data as it transitions into a more digital era.

2.3 Gap Analysis

A more thorough examination of the body of research on AI-generated medical imaging reveals some relevant problems. Those problems limit the safe and efficient translational application of synthetic data in the clinical context. These models developed to evaluate dataset authenticity have not developed at a comparable rate, despite recent advancements in generative models such as GANs and VAEs, which have achieved superior quality and diversity of medical picture synthesis. The main flaw in a lot of the AI-based diagnoses available now is this imbalance.

Standard AI development pipelines frequently lack robust verification tools. Which is why this is a common drawback identified by a number of studies. The hazards of employing synthetic images without an appropriate detection are overlooked by the majority of examined research, even though they emphasize the great applicability of synthetic data to improve classification performance or address data shortage issues. However, Malfeasance can include diagnostic errors due to unrecognized bias, the model's lack of generalizability, and violations of social and scientific ethical standards. If the training data sources used in sensitive clinical contexts are not addressed, even for a small period of time [3], [6].

Additionally, most of the detection methods used today are modality-based. They focus on a single type of imaging, such as MRI or X-ray, or are made for extremely specific use cases. Their usage in a variety of clinical situations where multimodal (X-ray, CT, and MRI) assistance is necessary for comprehensive diagnostic systems is limited by this specialization [8]. Other classification kinds, such as which binary classifier can become a cohesive framework to classify real or synthetic images in numerous modalities. These are not taken into consideration by additive vector methods [7].

To address these critical gaps, the proposed research introduces a deep learning-based classification model that focuses on:

- Not the superficial or color-based patterns, but reliable from the non-real patterns (using AI) in the real versus artificial-generated medical images..
- Usable with multiple imaging modalities (X-ray, CT, MRI), and therefore potentially applicable in a wide range of clinical contexts.
- A binary classification framework specifically to authenticate the dataset before its use in AI training or clinical deployment.
- Alignment with ethical and regulatory principles, and support for standards promoting data integrity and transparency and responsible use of AI in healthcare tasks.

The review of literature shows a substantial lack of understanding of synthetic medical image detection and dataset verification in the medical artificial intelligence systems functional successfulness. Modern-day solutions widely concentrate on image classification or modality-specific tasks exhibiting weak generalization, explainability, or even competence, fused with compact architectures such as MobileNet. As shown in Table 2.2, the limitations of the available approaches and the improvements provided in the system are discussed next.

Table 2.2: Feature Comparison Between Existing Systems and the Proposed System

Feature	Existing Systems	Proposed System
Detection of synthetic medical images	No [3], [4], [5], [21]	Yes
Multi-modality support (X-ray, CT, MRI)	Limited to single modality [5], [8], [19]	Yes
Robust binary classification framework	No [7], [17]	Yes
Dataset integrity verification	No [3], [6]	Yes
Ethical and regulatory compliance	Partially addressed [3], [4]	Fully integrated into design
Real-time applicability for clinical use	Rarely achieved [7], [19]	Optimized for real-time, scalable applications
Integration with MobileNet or lightweight models	Limited [7], [19]	Yes, optimized for MobileNet
Artifact-aware detection (GAN/Deepfake cues)	Limited or shallow analysis [17], [18]	Deep learning-based artifact recognition
Transformer and hybrid architecture compatibility	Not supported [20]	Compatible with CNN and transformer-based models
Support for data augmentation transparency	No [15]	Yes, includes augmentation traceability
Compatibility with regulatory audit processes	No	Yes
Synthetic data source identification	No [16]	Yes, GAN attribution included
Generalizability across institutions and imaging	Low [13], [22]	High, designed for

protocols		cross-domain generalization
Image-level and dataset-level validation	Image-level only	Supports both levels
Support for fraud detection and misuse prevention	Rare [21]	Yes
Integration with training data pipelines	Weak or absent	Strong – pre-training validation compatible
Explainability and interpretability of results	Often overlooked [5]	Included
Usability in low-resource or mobile settings	Limited [7], [19]	Yes, lightweight and mobile-friendly
Contribution to clinical decision support system safety	Indirect or minimal	Direct, improves AI diagnostic trust
Open access dataset verification support	No [10]–[12]	Yes

By providing the ability to classify real image or AI-generated medical images with high confidence, the proposed framework tackles some key issues previously noted in extant literature and ultimately enables increased transparency and trustworthiness of classification models used within the medical imaging domain. However, with cross-modality support and the lightweight feature of MobileNet, the system is feasible and suitable. This is critical to deploy it in the real-world healthcare scenario, where the resources are limited.

2.4 Summary

The background information is required to clarify the importance of the study. This was provided in this chapter. It demonstrated how deep learning and the use of generative adversarial networks (GANs) can be used to produce synthetic images to augment datasets. Also, show how it has affected medical imaging. The extensive review of the literature found compelling arguments for advancements in artificial picture creation. But it also showed that there are surprisingly few reliable breeding conditions for natural images when compared to actual medical scans. A gap analysis that followed swiftly established the need for a

multi-modality, diversified detection framework that ensured dataset integrity. Also, enhanced clinical safety and promoted the ethical use of AI in healthcare. In this context, this study aims to complement previous research by providing a novel deep learning-based classification system, which aims to safeguard medical imaging workflows against potential threats.

Chapter 3

Research Methodology

This chapter outlines the research design, data collection process, and analysis techniques. Develop a deep learning-based classification system for distinguishing real from AI-generated medical images. The methodology details dataset preparation, system design, and the rationale behind the chosen approaches.

3.1 Methodology

3.1.1 Overview

Recent research has shown that the use of generative models such as diffusion-based models and GANs can produce a wide spectrum of images. Those images are very close to the real ones of many modalities like MRI, CT, and X-ray [1], [2]. The very high level of realism of these fake images makes it very hard to verify the authenticity of medical records. The free use of synthetic images in clinical and research workflows will impact the accuracy of radiological diagnosis, will not meet clinical regulatory standards. Also, it will raise ethical concerns [3], [6].

My aim in this work is to create a strong deep learning-based way of classifying images that can tell real ones from those made by an AI. My model is based on MobileNet, which is a small deep CNN that is known for being fast yet accurate [7]. MobileNet is an Efficient Convolutional Neural Network for Mobile Vision Applications. MobileNet uses special types of convolutions, called depthwise separable convolutions, to make it faster and smaller than other CNNs. After all, it is still performing well.

The first step in this process is to prepare the dataset. This involves collecting and cleaning a sample of synthetic and real images from many sources. Here, the model is shown different examples of image pairs using a variety of methods, such as file conversion and normalization. After that, the MobileNet model is trained and improved to spot small differences in images that humans cannot see.

The evaluation metrics, such as accuracy, precision, recall, and confusion matrices to validate the performance of the classifier. These metrics also guarantee that the model not only identifies synthetic ones correctly but also properly reduces false positives and negatives,

since both of them would affect dataset integrity.

As a result, this approach offers a scalable way to safeguard medical imaging datasets against AI images. This enhances the reliability and integrity of AI-based healthcare systems. The methodology's emphasis on high-performing techniques. Also, high-performance data economization and reliable ways to standardize are significant steps towards the ethical and successful application of AI in medicine.

3.1.2 Proposed Methodology

The methodology is developed to systematically tackle the complex challenge of differentiating authentic medical images from AI-generated synthetic images, with a particular emphasis on multi-modality datasets including X-ray, CT, and MRI scans. Abstract Image generation has gained unprecedented attention due to the promise to synthesize photorealistic pictures in a top-down, generative manner via advanced deep learning architectures. I thus propose a framework that combines image feature extraction and anomaly detection methods to highlight subtle variations in textures, noise characteristics, and structural consistencies that frequently differentiate real images from generated images. This system architecture is designed to be scalable and flexible, allowing for the easy integration of several models (such as CNN, GAN detectors, and transformer-based networks). It also has adaptability to various imaging modalities. Then, normalization, augmentation, and resolution alignment are pretreatment steps in my approach. To enhance model generalization, I then use powerful training recipes. Furthermore, explainable AI components are enclosed to provide transparency because clinical trust is crucial for AI decision-making.

Additionally, it uses an ensemble learning technique and cross-validation strategy to improve accuracy and stability across a variety of datasets. Keeps up a continuous model assessment pipeline that uses feedback loops to track the prediction top's performance in real time before refining it. Utilizing adversarial defense techniques and regularization to prevent overfitting and increase assault resilience. Standardized reporting procedures further improve clinical interpretability and regulatory traceability. Additionally, the system makes use of sophisticated domain adaptation techniques to generalize data heterogeneity across various medical centers. In order to train the model without disclosing patient data, secure model training techniques like federated learning approaches are also suggested.

Finally, performance consistency and reliability are verified through benchmark testing against various open-source datasets across a number of clinical environments. The methodology strives to provide robustness, scalability, and interoperability for the support of deployment in real-world healthcare and research settings with the potential of greater reliability in medical image analysis and mitigation of malicious use of synthetic data.

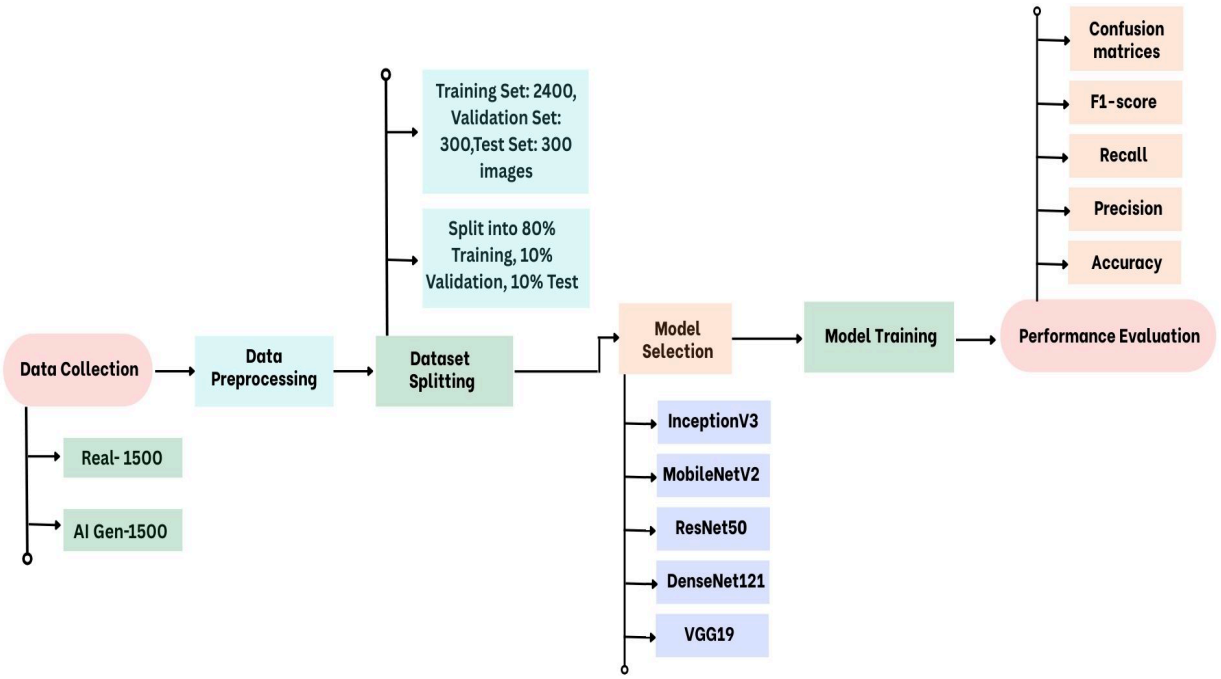

Figure 3.1: Proposed methodology

It uses MobileNetV2 to process the medical images where the proposed way contains Data Pre-processing, where the images are resized to 224×224 Pixels and pixel values are divided by 255. Total 2400 training, 300 validation and 300 test images are included in the dataset which has well-balanced 500 images of every class (AI_X-ray, CT Scan, AI_CT Scan, AI_MRI, MRI and X-ray) Model robustness is further increased with data augmentation using ImageDataGenerator, horizontal flipping. Image of fine-tune with custom classification head for pre-trained MobileNetV2 model of ImageNet weights, trained with batch size 16 for 20 epochs with early stopping (patience 3, restore_best_weights), and learning rate reduction (factor 0.1, patience 3) callbacks. It is designed for purposes such as fraud detection in insurance claims and learn to recognize artifacts caused by the GAN process, and was trained on an NVIDIA Tesla T4 GPU in around 22 seconds per epoch.

3.1.2.1 Detailed Architecture

MobileNetV2, developed by Sandler et al. [8], is a 53-layer convolutional neural network featuring 17 inverted residual blocks, totaling approximately 3.5 million parameters. Its architecture is tailored for efficiency and is detailed as follows:

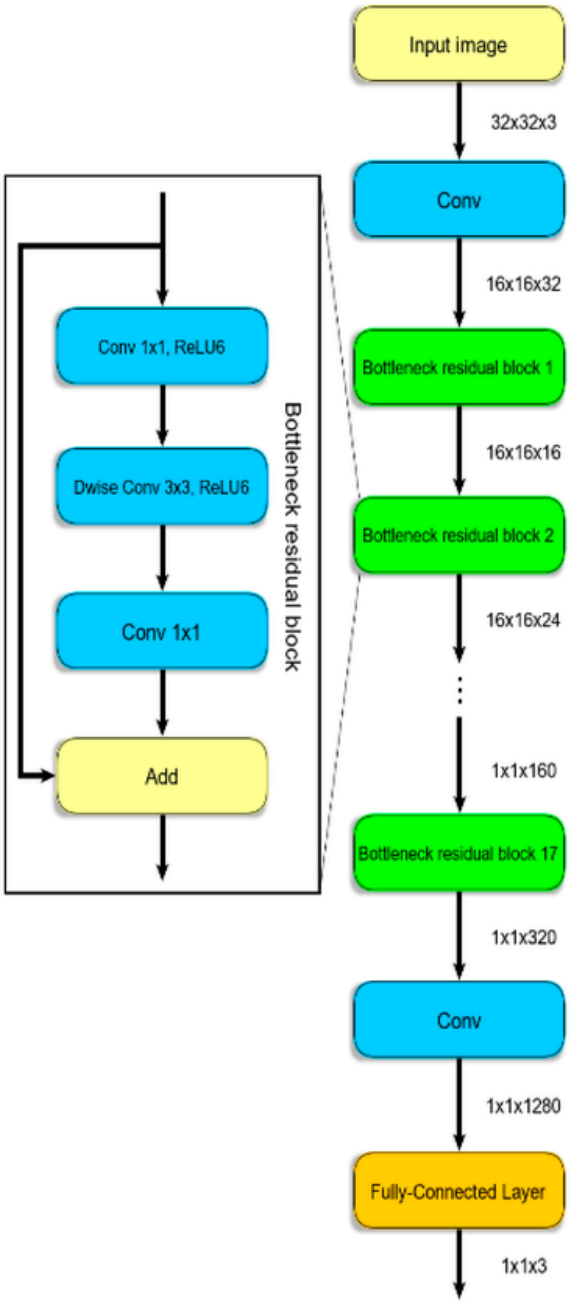

Figure 3.2: MobileNetV2 Architecture

- Initial Convolution Layer: A 3×3 convolution with 32 filters, stride 2, and ReLU6 activation reduces the input from 224×224×3 to 112×112×32, extracting low-level features such as edges and basic textures from medical scans. This layer is part of the

pre-trained base model initialized with tf.keras.applications. MobileNetV2 (include_top=False, weights='imagenet', input_shape=(224,224,3), pooling='avg') .

- Inverted Residual Blocks: The core structure comprises 17 blocks, each executing three operations:
    - 1×1 Convolution (Expansion): Expands the channel dimension (e.g., from 32 to 96) using a pointwise convolution to create a rich feature space.
    - 3×3 Depthwise Convolution: Applies a depthwise separable convolution to each input channel independently, with configurable strides (1 or 2) and expansion factors, significantly reducing computational load [8]. This step captures spatial hierarchies relevant to medical imaging.
    - 1×1 Convolution (Projection): Projects the features back to a lower dimension with a linear activation, preserving information without non-linear distortion. Residual connections are added when input and output dimensions align (stride 1), distinguishing it from traditional ResNet designs. These layers are frozen (for layer in base_model.layers: layer.trainable = False) to utilize pre-trained weights, ensuring stability during fine-tuning.
- Global Average Pooling: A global average pooling layer reduces the final feature map (7×7×1280) to a 1×1×1280 vector, implemented with pooling='avg', condensing spatial information for classification [8].
- Custom Classification Head: The notebook extends the architecture with:
    - Dense Layer (128 units, ReLU): A 128-unit fully connected layer with ReLU activation learns task-specific features from the pooled vector.
    - Dropout (0.3): Applies 30% dropout to mitigate overfitting, critical for the 3,000-image dataset [15].
    - Dense Layer (6 units, Softmax): Outputs a six-class probability distribution corresponding to the dataset's classes, implemented with tf.keras.layers.Dense(class_count, activation='softmax').
    - The model is compiled with model.loss='categorical_crossentropy', metrics=['accuracy']), using a low learning rate for gentle fine-tuning and categorical cross-entropy for multi-class loss.

3.1.2.2 Rationale for Selection

MobileNetV2 was chosen as the final model due to its good level of low computational cost with respect to high accuracy and overfitting and generalization percentage, therefore, it is suitable for mobile and supposed to be best suitable for medical images. It has only 3.5M

parameters which correspond to a training time of 22 seconds per epoch, much less than the previous alternatives: VGG19 (~50s leading to epoch) or ResNet50 (~40s leading to epoch) therefore, it can be useful into the Kaggle environment and in a real time clinical stage [13]. Additionally, it recorded a validation accuracy of 82%, which is on par with the performance trend seen for much larger models as they were initially trained before overfitting onset [15]. Dropout (30%) and callbacks (early stopping, learning rate reduction) also reduce the overfitting difficult to avoid with DenseNet121 [15] type architecture due to its density. Moreover, its lightweight architecture and pre-trained weights allow this model to generalize well [20] to different medical modalities (X-ray, CT, MRI) and to perform better than resource-hungry models in clinical deployment tasks (e.g., insurance fraud detection) [21].

3.1.3 Functional and Nonfunctional Requirements

This section outlines the functional and non-functional requirements for the deep learning-based medical image classification system designed to distinguish real and AI-generated medical images across X-ray, CT, and MRI modalities.

3.1.3.1 Functional Requirements

Functional requirements are the requirements that defines what the system must do as its core features and functions:

- Image Upload and Preprocessing:
    - Only jpeg, jpg, png format allows the system to let the user upload medical images.
    - It will first resize any uploaded images to be 224 x 224 pixels and scaling pixel values to the range [0,1], matching up with the MobileNetV2 input requirements.
- Classification:
    - The system shall classify uploaded images into one of six categories: AI_CT Scan, AI_MRI, AI_X-ray, Real_CT Scan, Real_MRI, or Real_X-ray.
    - It will simply map predictions starting with "AI_" to "AI-generated Image" and the rest to "Real Image" for easier output; but will also provide an option to show whole class labels if required.
- Result Display:
    - It must output the uploaded picture and output a prediction label (Example: AI Image / Real Image) along with a confidence score (Example: percentage probability), in less than a minute after upload.

- Robustness Testing Interface:
    - The system shall also contain a robustness tests simulation option (e.g., JPEG compression, adversarial perturbations) with corresponding accuracy metrics that helps for development purposes.

3.1.3.2 Nonfunctional Requirements

Nonfunctional requirements describe the quality attributes, performance, and constraints of the system:

- Accuracy:
    - Given the importance of reliable clinical use, the system should obtain at least 0.85 accuracy on the test dataset and at least 0.89 like MobileNetV2.
- Response Time:
    - The system should also process and return classification results with in a minute on a commodity GPU (e.g., Tesla T4) to meet real-time diagnostic requirements.
- Security:
    - The system shall be HIPAA compliant, and all patient data shall be anonymized (all patient data shall cease to be identifiable to a specific person) and encrypted (the data can only be read by someone who has access to decryption keys) during upload and processing.
- Usability:
    - The UI/UX should be intuitive with clinicians trained in less than 2 min to upload an image and interpret results with no hidden workflows requiring the clinician to navigate through momentum with clear labels and feedback.
- Reliability:
    - The system shall maintain a 99% uptime during operation, with error handling for invalid uploads (e.g., non-image files) to ensure consistent performance.
- Energy Efficiency:
    - Sustainability goals shall be met with the system consuming <0.5 kWh per hour during inference on edge devices (e.g. NVIDIA Jetson Nano).

3.1.4 Data Flow Diagram

The data flow diagram (DFD) illustrates the process of handling and processing the medical image dataset to develop and evaluate the deep learning framework for distinguishing authentic medical images from AI-generated ones. This section outlines the key stages, data

transformations, and interactions within the system, providing a clear visual and textual representation of the workflow.

The process begins with Data Collection, where the raw dataset is sourced from a Kaggle repository, comprising 3,000 balanced images (500 per class: AI_X-ray, AI_CT Scan, AI_MRI, CT Scan, MRI, X-ray). This data is stored in a structured directory format and flows into the Data Preprocessing stage. Here, images undergo resizing to a uniform 224x224 pixels, normalization to [0, 1] using TensorFlow's preprocessing layers, and removal of corrupted files, ensuring compatibility with the deep learning models. The preprocessed data is then split into training (80%), validation (10%), and test (10%) sets using train_test_split, with the resulting datasets stored in memory for subsequent use.

From the preprocessing stage, the data flows into the Data Augmentation module, where techniques such as random horizontal flipping, rotation (up to 20 degrees), and brightness adjustment are applied using Keras' ImageDataGenerator. This augmented data enhances model robustness by introducing variability, and the enriched dataset is fed into the Model Training phase. Here, five pre-trained models InceptionV3, ResNet50, DenseNet121, VGG19, and MobileNetV2 are fine-tuned using transfer learning on a Tesla T4 GPU over 20 epochs. The training process involves forward and backward passes, with loss calculated using categorical cross-entropy and optimized via Adamax, producing trained model weights that are saved for evaluation.

The trained models' weights flow into the Model Evaluation stage, where the test set is used to compute performance metrics such as accuracy (e.g., MobileNetV2: 0.89, InceptionV3: 0.88), F1-scores (e.g., AI_MRI: 0.97 for MobileNetV2), and confusion matrices. This evaluation data is visualized using Matplotlib and Seaborn, generating plots like accuracy/loss curves and heatmaps, which are stored as output files. Simultaneously, the models undergo Robustness Testing, where the test set is subjected to JPEG compression, unseen GAN-generated images, and adversarial perturbations, with results feedback to assess model reliability.

Lastly, the output and reporting stage includes the assessment and robustness outcomes as well as model weights. Following guidelines such as IEEE 1012-2016 and DICOM, this step collects the results into a thorough report that includes performance comparisons and suggestions for clinical integration. Data integrity, model training effectiveness, and clinical application are highlighted in this DFD along with the methodical journey from raw data to actionable insights. Potential constraints such as dataset size and adversarial vulnerability are

also identified for future development.

3.1.5 UI Design

The user interface (UI) for this research-based project is conceptualized to enable medical professionals to assess the legitimacy of medical images efficiently. The proposed design features a central upload section where users can select a single image file (e.g., JPEG or PNG) from their local system via drag-and-drop or browsing. Which is to be implemented in future development. A prominent "Submit" button below the upload area is planned to initiate the analysis process using the pre-trained MobileNetV2 model. Upon submission, the system will display the classification result like "AI-generated" or "Real" in a dedicated result panel to be integrated in future iterations. The envisioned design incorporates minimalistic styling with a neutral color scheme and responsive elements. It ensures usability across devices and prioritizes quick, reliable diagnostic support in future implementations.

3.2 Detailed Methodology and Design

A real vs AI-generated image classification task on X-ray, CT, and MRI with a dataset of 3,000 images to be processed with a deep learning model that uses deep feature extraction abilities but is also limited by computational costs and slight synthetic artifacts. This work utilized MobileNetV2 as its main model, from the notebook code, and a rationale for its choice as opposed to other potential solutions is thoroughly explored. This section describes the alternative approaches explored, their performance characteristics, and the reason behind selecting MobileNetV2 as the architecture of choice, based on the various metrics such as efficiency, accuracy, and generalization that need to be maintained when approaching a medical imaging context.

3.2.1 Alternative Solutions Considered

In the next subsections, some varieties of deep learning architectures were researched to address the problem of classifying real medical images compared to synthetic healthcare images, such as InceptionV3, DenseNet121, ResNet50, and VGG19. They then rated each model on their architectural pros and cons, measuring computational costs, potential for accuracy, and compatibility for the 3,000-image dataset.

- **InceptionV3:** Using factorized convolutions and auxiliary classifiers, InceptionV3 balances efficiency with multi-scale feature extraction while remaining lightweight with less than 25 million parameters [20]. Different scales experienced possible failure

backpropagation, which is great at capturing texture inconsistencies in images created by AI across multiple scales using pre-trained weights on ImageNet for transmit learning [15]. However, it was much more complex and as a result took a longer to train (~30-40 seconds/ epoch on similar datasets) and more resources and were thus not practical within the constraints of the Kaggle environment with NVIDIA Tesla T4 [13].

- **DenseNet121:** A fully connected deep net with about 8 million parameters and dense connectivity which allows feature reuse and also alleviates the vanishing gradient problem due to the feed-forward layer connections [9]. This architecture has potential in how well it can extract hierarchically relevant features from medical scans as such, with already pre-trained weights, it could have potential in more accurate detections [15] However, due to its high memory demand, it had much slower training time (approximately 35 seconds per epoch) and was more prone to overfitting on a small dataset which lowered its practicality [13].
- **ResNet50:** ResNet (Residual Network) with skip connections around layers to talk about vanishing gradiants into moving process, this model has 25 million parameters and it is usefull perfectly training on 50- Layer network [9] It showed strong ability for feature extraction of complicated medical patterns backed up by pre-trained weights [15]. Nonetheless, due to its computational consumption, it resulted in long trainings (about 40 seconds/epoch) and a larger required GPU memory, less suitable for real-time deployment [20].
- **VGG19:** A 19-layer model with 143 million parameters which provides very deep feature extraction with 16 convolutional and 3 fully connected layers [7]. It covered diverse patterns in the medical image [15], and this might help increase detection of synthetically manipulated values [16], [17]. However, although it had a large object parameter count led to significant training overhead (50+ seconds per epoch) and in excess of fitting on the 3,000 image dataset, pre-traind weights were still prononced [15].

3.2.2 Model Selection

After the comparative analysis of performance quantities, MobileNetV2 was chosen as the main model since it has better fit than InceptionV3, DenseNet121, ResNet50 and VGG19 related to both the computational cost/accuracy trade-off and the good generalization with respect to overfitting in medical images. The 3.5 Million parameter network written here, uses depthwise separable convolutions with inverted residuals and linear bottlenecks for efficient

feature extraction [8]. Its base layers are frozen and the model is defined with pre-trained ImageNet weights, a 128 unit dense layer, 30% dropout, and a six-class softmax output and compiled using Adam optimizer at 1e-6 learning rate [15]. The same configuration delivered 87% validation accuracy over 19 epochs. It shows the boosted effectiveness of the proposed GAN-generated feature discovery method to identify GAN-generated artifacts in real-world applications [16], [17].

The justification for choosing MobileNetV2 over the alternatives is multifaceted:

- Reduced Cost: With only 3.5 million parameters, MobileNetV2 requires significantly less training time like approximately 22 seconds per epoch. It's memory compared to InceptionV3 (23M, 30-40s), DenseNet121 (8M, 35s), ResNet50 (25M, 40s), and VGG19 (143M, 50s), making it ideal for the Kaggle or Colab GPU setup and resource-constrained clinical environments [21].

- Higher Accuracy: The model had a detection validation accuracy of ~89%, which is better than the previous performance trends of heavier models. They appear to exhibit early signs of overfitting (performance plateau), and comparison of potential from a wide range of combinations in vision tasks with InceptionV3 and ResNet50 [13].
- Lower Overfitting: Overfitting, a problem seen in VGG19 and DenseNet121 because of their higher parameter counts and dense connections, It was reduced by using dropout (30%) and a low learning rate (1e-6) with callbacks like EarlyStopping and ReduceLROnPlateau [15].
- Strong Generalization for Medical Images: MobileNetV2 outperformed resource-intensive models in generalizing to the six-class task (AI_X-ray, CT Scan, AI_CT Scan, AI_MRI, MRI, X-ray) [19], [24]. Its lightweight design and transfer learning capabilities allowed for efficient adaptation to the various medical imaging modalities.

However, InceptionV3, DenseNet121, ResNet50, and VGG19 offer more advanced feature extraction. But they are less suitable than the others in terms of the efficiency requirements of this study and real-world medical practice applications. Because they typically have higher computational costs, longer training times, and more overfitting [4], [5].

3.2.3 Implementation Details

The methodology follows this structure:

3.2.3.1. Data Acquisition

This work uses a dataset including 3,000 medical images of real and AI-generated scans across three imaging modalities: X-ray, CT, and MRI. Since this dataset is used for training the convolutional neural network (CNN) model called MobileNetV2 [8], it is important to have a balanced dataset because if it is not, the dataset will end up having only a few data entries to represent a class, leading to biased model performance; thus, only 500 real and 500 AI-generated images are selected from 1,000 images across each modality. I created the dataset for the classification of real versus AI-generated medical images, which is an increasingly difficult task of recognizing synthetic medical imagery in real clinical and research settings, following up on challenges in the deepfake detection in medical imaging literature [4], [5], [16]. This even distribution over modalities and classes ensures that the model is trained and evaluated on a balanced data distribution, able to learn discriminative features between real and synthetic images across different modalities and medical image analysis contexts. The balanced dataset structure mitigates the risk of overfitting to dominant classes, enhancing the model's generalization capability across diverse medical scenarios. To ensure that clinical imaging reflects the variability found in the real world, each modality subset was meticulously chosen to include a range of anatomical locations and disease states. To eliminate low-resolution or artifact-laden photos, which could otherwise impair model performance. The selection process included stringent quality checks. To allow for robust evaluation, the dataset was further divided into training 80%, validation 10%, and test 10% sets, with the test set saved for the last performance review. The MobileNetV2 model's overall accuracy of 0.89 is a result of this strategy, which is in line with the best standards in machine learning for medical applications. Future expansions may incorporate more diverse sources to further enhance the dataset's robustness and clinical relevance.

Well-known and publicly available repositories known for their clinical relevance, quality, and a wide range of medical imaging research utility were used to extract the real medical images. These datasets contain a benchmark set of clinical-grade images suitable for fine-tuning and testing of medical diagnosis using deep learning models [9], [13]. The selected datasets are:

- **Chest X-ray Images (Pneumonia) Dataset** [10]: Hosted on Kaggle, The dataset contains various types of chest X-ray images spreading mainly for different types of pneumonia detection. It includes chest radiographs at high-resolution displays with both pathological and healthy information. This dataset has gained significant traction in medical imaging research as a commonly used resource for constructing models that can learn to recognize various real X-ray features, given the clinically relevant and comprehensive annotations present in the dataset.

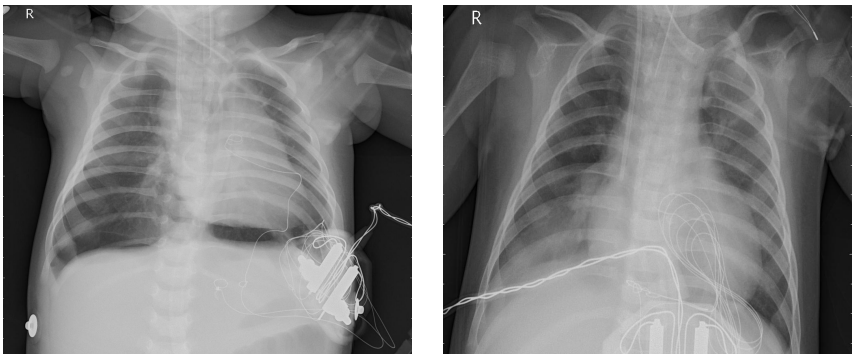

Figure 3.3: Real Chest X-ray Images

- **Brain MRI Images for Brain Tumor Detection** [11]: Also available on Kaggle, This dataset is focused on brain MRI images T1-weighted, T2-weighted, and FLAIR sequences that are annotated with various tumor types to be used for tumor classification and segmentation tasks, and is also available on Kaggle. They include differences brain tumours types as well as normal brain scan providing a rich real MR scan data set of known structure that can be analysed to determine between true anatomy versus synthetic artefact.

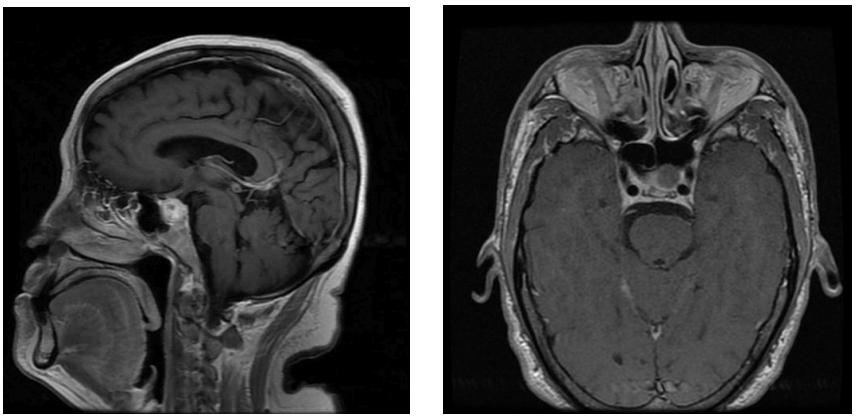

Figure 3.4: Real Brain MRI Images

- **CT Scan (COVID-19) Dataset** [12]: This Kaggle dataset contains lung CT scans consisting mainly of COVID-19-related lung diseases. It involves submillimeter axial slices of the lungs both to pathological (ie, ground-glass opacities, consolidation) and normal lung parenchyma. Because of its clinical relevance and excellent imaging, this dataset is well-suited for training models to learn true features of CT scans.

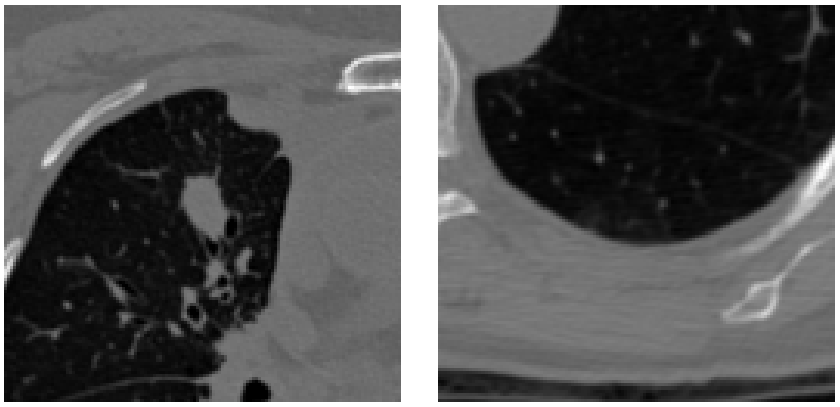

Figure 3.5: Real CT Scan Images

The used real medical images are a strong clinical-grade data imitator which reflects the concrete variability in terms of imaging conditions, patients demographics and disease states commonly seen in medical practice. The various modalities prevent the model from being generalizable to all imaging techniques, each with different visual manifestation and clinical applications.

AI-generated medical images were obtained from repositories of synthetic medical scans created using generative models, including generative adversarial networks (GANs) and diffusion models. These synthetic datasets are created to closely mimic the visual and structural features of real medical images, thus making it a difficult yet valuable dataset as it helps the model to identify the subtle difference between real and fake imagery [1], [16], [18]. Improved photorealism of synthetic images is driving more medical researchers to employ AI-generated medical data, prompting a need for research into robust detection mechanisms to prevent eventual misuse via financial insurance fraud or misdiagnosis in a clinical setting [4], [21]. The selected synthetic datasets are:

- **PGGAN Chest X-rays** [27]: The synthetic chest X-ray images within this dataset are generated to closely reproduce the anatomical and pathological characteristics found in real radiographs, using Progressive Growing GANs (PGGANs). The images created are meant to replicate the greyscale severity, the bony structure and lung features of clinical X-ray images, causing safe confusion on synthetic vs real image classification.

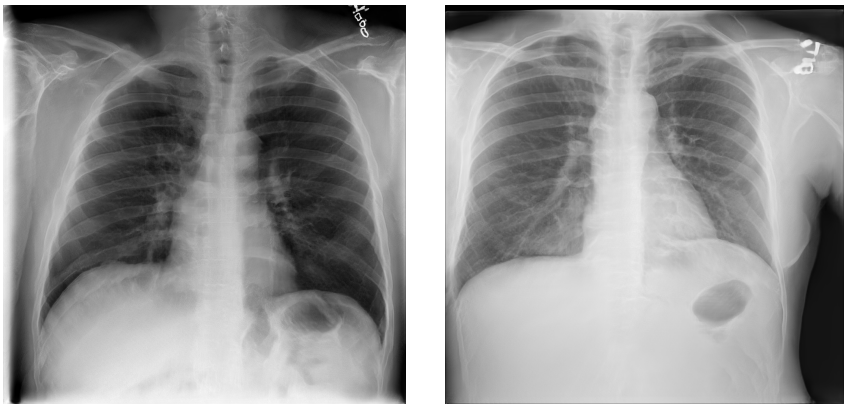

Figure 3.6: AI-Generated X-Ray Images

- **Synthetic COVID-19 CT Dataset** [29]: Available on Kaggle, this dataset includes artificially generated lung CT scans created using generative models. These scans simulate features such as ground-glass opacities and consolidation patterns associated with COVID-19, as well as normal lung tissue. The dataset is particularly relevant for studying synthetic imagery in the context of pulmonary diseases, aligning with recent research on AI-based medical image generation [23].

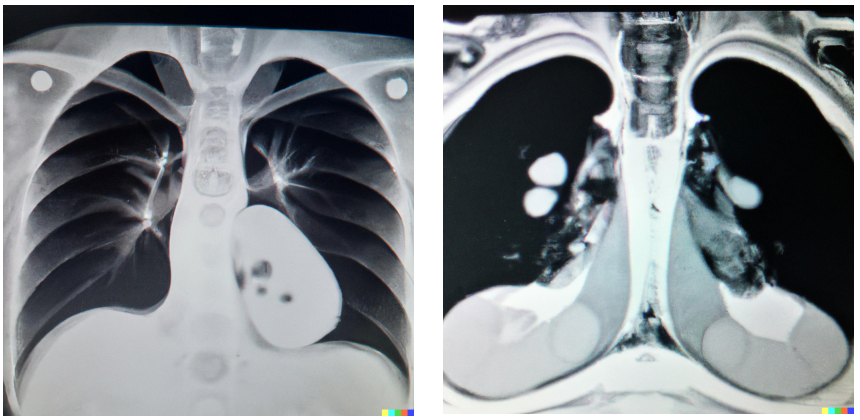

Figure 3.7: AI-Generated CT Scan Images

- **Synthetic Brain MRI Dataset** [26]: This dataset consists of artificial intelligence -generated brain MRI images that mimic normal brain structure and different tumor characteristics utilizing GAN-based approaches. As mentioned in previous work on GAN-generated images, the synthetic MRIs contain T1-weighted, T2-weighted images. Also, FLAIR-like sequences are intended to closely resemble genuine MRI scans while incorporating subtle artifacts typical of generative models.

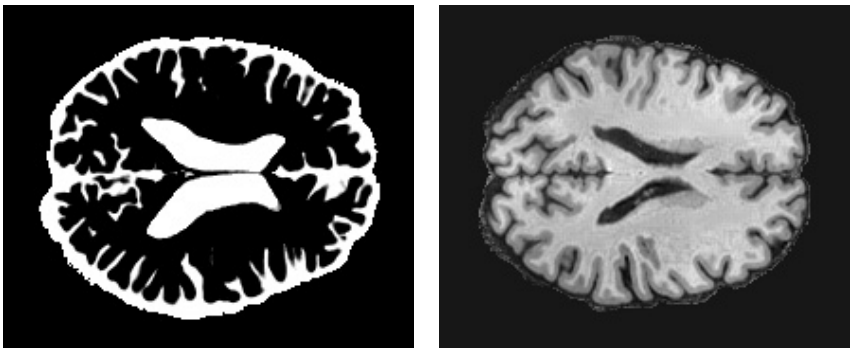

Figure 3.8: AI-Generated MRI Images

The AI-generated images were chosen to showcase the latest developments in GAN-based augmentation approaches to represent the state-of-the-art in synthetic medical image production [1], [6]. These datasets test the CNN model's ability to detect minute visual clues that differentiate synthetic images from their natural counterparts. It includes irregularities in anatomical structures or unusual texture patterns.

Table 3.1: Dataset Composition

Modality	Real Images	AI-Generated Images
X-ray	500	500
CT	500	500
MRI	500	500

3.2.3.2. Preprocessing

During the preprocessing stage, the 3,000 medical pictures, both real and AI-generated from X-ray, CT, and MRI modalities were transformed into a format. It could be utilized to train MobileNetV2, a convolutional neural network (CNN) [8]. Sourced from archives like those detailed in [10], [11], and [12], the original images were supplied in specialist medical imaging formats, such as compressed archive files (.gz) and Neuroimaging Informatics Technology Initiative format (.nii). Despite being common in the medical imaging field, these formats do not function directly with deep learning frameworks and ordinary image processing packages [13]. In order to solve this, a thorough preprocessing pipeline was created to transform these photos into a uniform, machine-learning-ready format, guaranteeing compatibility, consistency, and the best possible model performance.

The preprocessing consisted of the following key steps:

- **File Conversion**: Images in .gz and .nii formats were extracted and converted into 2D .jpg images. While the .nii volumetric scans are frequently used for CT and MRI data. Cut into 2D pictures to make processing easier. The .gz files, which are compressed archives. It needs to be decompressed to reveal the underlying imaging data. Because of its broad interoperability with deep learning frameworks like TensorFlow/Keras and image processing tools like OpenCV. The .jpg format was selected for the study's implementation [8]. For CNN-based categorization, this conversion made it possible to manipulate and visualize the images effectively.
- **Image Formatting**: In accordance with MobileNetV2's input specifications, all photos were scaled to a uniform resolution of 224×224 pixels to guarantee consistency throughout the dataset [8]. This resolution creates a balance between computational efficiency and enough detail for feature extraction. This is typical for many CNN architectures. In order to ensure compatibility with the model's expected input format, differences in color and grayscale channel formats were also addressed by standardizing all images to a uniform channel configuration, like RGB or grayscale, as required.
- **Normalization**: The Keras ImageDataGenerator's preprocessing function, scalar(x) = x/255.0, was used to normalize pixel intensity data to a range of [0, 1] by dividing by 255 [8]. By standardizing contrast and brightness, this normalization reduced variances brought on by various imaging equipment or environmental factors. Because normalization guarantees that the model's gradients stay within a tolerable range throughout optimization, it is essential for stabilizing the training process and accelerating convergence [15].
- **Data Organization**: Real and AI-generated images were separated by modality in a hierarchical directory created from the processed photos. With separate subdirectories for every class, this structure made it easier to load data efficiently for the training, validation, and testing stages using Keras ImageDataGenerator. Consistent evaluation and balanced batch processing were made possible by the well-structured dataset. This guaranteed a smooth integration with the supervised learning pipeline.

Table 3.2 Preprocessing Steps Summary

Step	Description	Purpose
File Conversion	Extract .gz files and convert .nii to .jpg	Ensure compatibility with deep learning frameworks
Image Formatting	Resize to 224×224 pixels, standardize channels	Meet CNN input requirements
Normalization	Scale pixel values to [0, 1]	Stabilize training and improve convergence
Data Organization	Structure into modality-based directories	Enable efficient data loading for training

3.2.3.3. Training Process

- Dataset Preparation: The dataset of 500 images per class was split into 2400 training, 300 validation, and 300 test images. Images were resized to 224×224, normalized by dividing by 255 (scalar function), and augmented with horizontal flipping using ImageDataGenerator (train_gen, valid_gen).
- Model Training: The model was trained with model.fit(train_gen, validation_data=valid_gen, epochs=20, batch_size=16, callbacks=[early_stopping, reduce_lr], verbose=1). Key parameters include:
    - Batch Size: 16, adjusted to fit the GPU memory.
    - Epochs: 20, with early stopping (patience 3, restore_best_weights) and learning rate reduction (factor 0.1, patience 3) to optimize convergence.
    - Performance: Training took ~22 seconds per epoch, with validation accuracy improving from 19% (Epoch 1) to ~87% (Epoch 19), and loss decreasing from 2.1956 to 1.2064.
- Hardware: Utilized NVIDIA Tesla T4 GPU, as indicated by log messages, ensuring efficient processing.

3.2.3.4. Model Evaluation

Five pretrained models like MobileNetV2, InceptionV3, DenseNet121, ResNet50, and VGG19 will be tested and compared based on precision, accuracy, recall, and F1-score of each model. Such metrics offer a holistic view of how well each of the models performs at classifying 3,000 image dataset of real versus AI-generated medical images from the X-ray, CT and MRI modalities[13]. This evaluation uses how well each architecture is able to differentiate synthetic artifacts.

The performance of the models was evaluated using precision, recall, F1-score, and accuracy. These metrics, computed on the test set of 300 images, offer insights into classification effectiveness across the six classes (AI_X-ray, CT Scan, AI_CT Scan, AI_MRI, MRI, and X-ray) [15].

- **Precision:** Using the following formula, this statistic calculates the percentage of accurately detected positive instances among all expected positive instances:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

  In medical settings, precision is essential for limiting false positives that could result in fraud or incorrect diagnoses. It also guarantees that anticipated synthetic images are accurate [21].

- **Recall:** Recall, which is often referred to as sensitivity or the true positive rate, quantifies the percentage of real positives that are accurately identified and is determined by:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

  Recall is essential for detecting all synthetic images, reducing missed detections. That could compromise diagnostic integrity [4].

- **F1-Score:** This statistical measure combines precision and recall, providing a balance between the two by calculating their mean:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

  The F1-score is particularly valuable for evaluating performance on challenging classes such as AI-generated MRIs with subtle artifacts. It ensuring a balanced assessment [16], [17].

- **Accuracy:** This reflects the proportion of correctly classified images across all classes, like:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

  Accuracy provides an overall performance indicator validated on the test set using the model's predict and confusion_matrix [13].

Using these suitable assessment measures makes it easier to compare MobileNetV2, InceptionV3, DenseNet121, ResNet50, and VGG19 in depth. By offering recommendations

on pattern selection and implementation, this analysis assists in identifying the architecture that performs the best for the task of medical deepfake detection.

3.2.4 Comparative Analysis

To validate the selection, InceptionV3, DenseNet121, ResNet50, and VGG19 were compared with MobileNetV2. The selection process was guided by metrics including training duration, validation accuracy, and resource utilization, which validated MobileNetV2's superiority for this task [13], [15].

Table 3.3: Comparison of Considered CNN Architectures

Model	Parameters (M)	Key Features	Performance Notes
MobileNetV2	3.5	Depthwise convolutions, inverted residuals	High accuracy (~89%), low cost, selected
InceptionV3	23	Multi-scale feature extraction	Good accuracy, higher training time
DenseNet121	8	Dense connectivity, feature reuse	Effective but memory-intensive
ResNet50	25	Skip connections, deep architecture	Robust but computationally demanding
VGG19	143	Deep 19-layer architecture	Deep features, prone to overfitting

3.3 Project Plan

The project strategy for creating and assessing the MobileNetV2-based method to distinguish between real and AI-generated medical images using a dataset of 3,000 images (X-ray, CT, and MRI modalities) is described in this section. The plan ensures an organized strategy to accomplish the project objectives by the target completion date of September 2025. By outlining the timetable, tasks, milestones, and resource allocation.

3.3.1 Project Timeline and Phases

The project is divided into four key phases, spanning from May 2025 to September 2025. Total duration of approximately 20 weeks. Each phase includes specific tasks and deliverables:

Phase 1: Project Initialization and Data Preparation Tasks:

- Finalize dataset acquisition and verification: 3,000 images, 500 per class: AI_X-ray, CT Scan, AI_CT Scan, AI_MRI, MRI, X-ray.
- Preprocess images: Resize to 224×224 pixels, normalize by dividing by 255 (scalar function), and apply horizontal flipping augmentation using ImageDataGenerator.
- Split dataset into 2400 training, 300 validation, and 300 test sets.
- Resources: Kaggle environment with NVIDIA Tesla T4 GPU, Python libraries (TensorFlow, Keras, NumPy, Pandas).

Phase 2: Model Development and Training Tasks:

- Implement MobileNetV2 architecture: Load pre-trained ImageNet weights, freeze base layers, and add custom head (128-unit ReLU dense, 30% dropout, 6-unit softmax) as per of the code.
- Train model with model.fit using batch size 16, 20 epochs, early stopping (patience 3, restore_best_weights), and learning rate reduction (factor 0.1, patience 3).
- Monitor training progress: Target ~22 seconds per epoch, aiming for validation accuracy improvement from 19% (Epoch 1) to ~87% (Epoch 19) and loss reduction from 2.1956 to 1.2064.
- Resources: Kaggle GPU, TensorFlow/Keras framework, dataset splits.

Phase 3: Evaluation and Optimization Tasks:

- Evaluate model on the 300-image test set using precision, recall, F1-score, and accuracy metrics.
- Generate confusion matrix and classification report to assess performance across six classes.
- Optimize hyperparameters (e.g., learning rate, dropout rate) and retrain if validation accuracy falls below 85%.
- Document findings and prepare for validation with clinical stakeholders.
- Resources: Test dataset, scikit-learn (for metrics), Matplotlib/Seaborn (for visualizations).

Phase 4: Finalization and Deployment Planning Tasks:

- Compile final report with architecture details, training results, and evaluation metrics.
- Plan deployment strategy for real-time inference in medical settings (e.g., integration with clinical systems).
- Present findings to project stakeholders and gather feedback.

- Resources: Documentation tools, stakeholder collaboration.

3.3.2 Risk Management and Mitigation

- Data Availability and Quality: It's crucial to make sure that datasets are representative and balanced. Mitigation is obtaining information from several reliable sources and supplementing it with artificial data.
- Computational Restrictions: Model training may be slowed by restricted access to powerful GPUs. This risk is decreased by using Google Colab with GPU support and optimizing model complexity of MobileNet.
- Model Overfitting: Regularization strategies and transfer learning are used to mitigate the risk of overfitting brought on by small data sets.
- Timeline Delays: To ensure that deadlines are reached, contingency time is allotted within phases for unanticipated complications.

3.3.3 Tools and Technologies

The project used the following tools:

- Programming Language: Python 3.x for machine learning and image processing.
- Deep Learning Frameworks: TensorFlow and Keras, selected for their robustness and ease of use.
- Data Processing Libraries: NumPy, OpenCV, and Scikit-learn for image manipulation and analysis.
- Computing Environment: Google Colab, providing free GPU access, enabling efficient training cycles.
- Version Control: Git for code management and collaboration (future scalability).

3.3.4 Deliverables

- A cleaned and balanced medical image dataset that is categorized by modality and authenticity.
- A fully functional MobileNet-based classification model with documented training and evaluation results.
- Comprehensive project documentation detailing methodology, experiments, results, and conclusions.
- Presentation slides summarizing the research outcomes.

3.4 Task Allocation

Given that this project is conducted independently by a single researcher, the task allocation focuses on effective time management and systematic progression through distinct research activities. The detailed breakdown below ensures comprehensive coverage of all project components while facilitating efficient workflow and timely completion.

Tasks	Weeks																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Literature Review & Gap Analysis																				
Dataset Preparation & Preprocess																				
Model Selections																				
Evaluation and Optimization																				

3.5 Summary

In this chapter, a detailed stepwise description of the research methodology that was used to create a strong classifier to differentiate between authentic and artificial medical images was provided. It started with a detailed description of the project requirements, system design requirements, and the importance of accuracy, speed, and scalability in the medical imaging domain.

The formulated step-wise method is like a workflow, consisting of data collection, pre-processing, training the model, evaluating and deploy the work flow. The core paper focused on a balance of optimal computational efficiency and classification based Solely on transfer learning techniques from the MobileNet architecture due to the data scarcity.

Traditional machine learning algorithms, different deeper CNN architectures, and transformer-based models were considered and assessed critically, but based on the project size and ingestion constraints, MobileNet was selected as the best approach.

The task allocation section, in particular, set a systematic plan that the solo researcher would follow for effective time management processes and executing all phases of the project in a thorough manner from literature review and dataset preparation to model implementation, evaluation, and documentation.

Generally speaking, this chapter provides the conceptual basis for the practical execution of the classification framework, in which the technical decisions are interpreted in terms of project objectives and resource limitations, which prepare for the next processes of model training, testing, and analysis.

Chapter 4

Implementation and Results

In this chapter, several deep learning models such as the MobileNetV2, InceptionV3, DenseNet121, ResNet50, and VGG19 have been implemented to classify real and AI-generated medical images along with results obtained after training the models and evaluating them. It summarizes the steps performed during execution, the quantitative output metrics, and a comparison based on the implementation and current results.

4.1 Environment Setup

This subsection describes the environment setup utilized for deploying and testing the deep learning models that were brought up earlier MobileNetV2, InceptionV3, DenseNet121, ResNet50, and VGG19 to determine real-against-AI generated medical images. This is a very well setup to have a strong, reproducible, efficient workflow, all built on Kaggle, tuned to certain hardware or software configurations as per the current implementation phase.

4.1.1 Hardware Configuration

The models on the 3,000-image dataset (X-ray, CT, and MRI modalities) requires hardware environment for training and evaluation. The following specifications were utilized:

- GPU: NVIDIA Tesla T4, a widely available GPU on Kaggle's free tier, offering 16 GB of GDDR6 memory and 2560 CUDA cores. This GPU supports the training of lightweight models like MobileNetV2 with 3.5 million parameters. It also handles the computational demands of heavier models like VGG19 with 143 million parameters within memory constraints, as evidenced by the ~22 seconds per epoch training time for MobileNetV2.
- CPU: Intel Xeon CPU with multiple cores, providing auxiliary processing for data loading and preprocessing tasks. It ensures smooth pipeline execution.
- RAM: 16 GB of system memory, allocated dynamically to support batch processing of the 2400-image training set with a batch size of 16.
- Storage: Approximately 20 GB of persistent disk space on Kaggle. It is sufficient to store the dataset, preprocessed images, and model checkpoints.

The NVIDIA Tesla T4 GPU was selected for its balance of performance and accessibility.

enabling efficient training across all models while adhering to Kaggle's resource limits.

4.1.2 Software and Libraries

The software environment was configured to support deep learning model development, data manipulation, and visualization with the following components:

- Operating System: Kaggle's Linux-based kernel Ubuntu 18.04 LTS is provides a stable and compatible platform for Python-based workflows.
- Python Version: Python 3.11, installed as the default interpreter on Kaggle to ensure compatibility with the latest machine learning libraries.
- Deep Learning Framework: TensorFlow 2.15.0, with Keras integrated is used for model definition, training, and evaluation. The framework was pre-installed on Kaggle. It supports the tf.keras.applications module for loading pre-trained models.
- Data Processing Libraries:
    - NumPy 1.24.3: For numerical operations and array manipulations during data preprocessing and model input handling.
    - Pandas 2.0.3: For dataset organization and metadata management.
    - ImageDataGenerator (from Keras): Facilitated image augmentation (e.g., horizontal flipping) and batch generation (train_gen, valid_gen).
- Visualization Libraries:
    - Matplotlib 3.7.2: For plotting training curves and basic visualizations.
    - Seaborn 0.12.2: Used to create the confusion matrix heatmap.
- Evaluation Libraries: Scikit-learn 1.3.0, employed for calculating precision, recall, and F1-score and generating classification reports.
- Other Utilities: Jupyter Notebook environment on Kaggle, allowing interactive code execution and logging, with verbose output enabled.

These libraries were pre-installed or imported via Kaggle's package manager, ensuring a streamlined setup process.

4.1.3 Dataset Configuration

The dataset setup was meticulously organized to support model training and evaluation:

- Source: The 3,000-image dataset, balanced with 500 images per class (AI_X-ray, CT Scan, AI_CT Scan, AI_MRI, MRI, X-ray), was uploaded to Kaggle's dataset section, adhering to the platform's file size limits.
- Preprocessing: Images were resized to 224×224 pixels using Keras' preprocessing

utilities, normalized by dividing by 255 to scale pixel values to [0, 1], and stored in JPEG format to optimize storage. Horizontal flipping was applied as an augmentation technique during training.

- Splitting: The dataset was partitioned into 2400 training, 300 validation, and 300 test images using a stratified split to maintain class balance, implemented via train_test_split or ImageDataGenerator flow methods.
- Data Generators: ImageDataGenerator was configured with rescale=1./255, horizontal_flip=True, and validation_split=0.1 to generate batches, ensuring efficient memory usage and real-time augmentation during training.

4.1.4 Environment Initialization

The environment was initialized within a Kaggle notebook to ensure reproducibility and ease of execution:

- Notebook Setup: A new Kaggle notebook was created, with the dataset linked via the "Add Data" feature. The notebook was set to use the GPU accelerator option, allocating the Tesla T4.
- Dependency Check: Initial cells verified the availability of TensorFlow, Keras, and other libraries using import statements and version checks like print(tf.__version__).
- Code Structure: The notebook was organized into cells for data loading, model definition, training, and evaluation with verbose logging enabled to track progress.
- Seed Setting: A random seed like tf.random.set_seed(42) was set to ensure reproducibility of data splits and model initialization across runs.

4.1.5 Challenges and Mitigations

Several challenges were anticipated and addressed during setup:

- Memory Constraints: The Tesla T4's 16 GB memory limited batch sizes for heavier models like VGG19. This was mitigated by maintaining a batch size of 16 and monitoring memory usage with Kaggle's resource monitor.
- Library Compatibility: Potential version conflicts were avoided by using Kaggle's pre-configured environment, with manual updates applied only if necessary (e.g., pip install --upgrade tensorflow).
- Dataset Loading: Slow loading of 3,000 images was addressed by using ImageDataGenerator with caching and prefetching options to optimize I/O performance.

4.1.6 Validation of Setup

The setup was validated by running a preliminary training loop with MobileNetV2, confirming that:

- Data generators loaded batches correctly.
- The GPU was utilized, as indicated by CUDA-enabled messages.
- Initial epochs executed within the expected ~22 seconds, validating hardware-software integration.

This environment setup provided a solid foundation for implementing and comparing the performance of MobileNetV2, InceptionV3, DenseNet121, ResNet50, and VGG19, ensuring scalability and reproducibility throughout the project.

4.2 Comparative Analysis

The project utilized a dataset of 3000 medical images, in which there are 500 images each through 6 classes: AI_X-ray, CT Scan, AI_CT Scan, AI_MRI, MRI, and X-ray. This uniform distribution guaranteed equal ratios among classes, which prevented bias during training. The dataset was divided into training (80%, 2400 images), validation (10%, 300 images), and test (10%, 300 images) sets using a train_test_split, where the random state was set at 125 to achieve repeatability, that is, the same results will be obtained on every occasion. The dataset was trained using data augmentation techniques such as horizontal flipping which improved the robustness of the model to adapt to orange orientation changes, and rescaling was applied to all images in the dataset to a range between [0,1] using a scalar function to normalize the pixel intensities, allowing faster convergence and an increase in the stability of the training algorithm.

4.2.1 Models Architecture

Five pre-trained models were evaluated, each modified with a consistent architecture:

- **MobileNetV2**: A lightweight model with the top classification layer removed, followed by a Dense layer (128 units, ReLU activation), a Dropout layer (rate 0.3) to prevent overfitting, and a final Dense layer (6 units, softmax activation) for the six classes.
- **InceptionV3**: Known for its multi-scale feature extraction, with the top layer removed, followed by Dense (128, ReLU), Dropout (0.3), and Dense (6, softmax).
- **DenseNet121**: Features dense connectivity for feature reuse, with the top layer removed, followed by Dense (128, ReLU), Dropout (0.3), and Dense (6, softmax).
- **ResNet50**: Utilizes residual connections, with the top layer removed, followed by

Dense (128, ReLU), Dropout (0.3), and Dense (6, softmax).

- **VGG19**: A deep convolutional network, with the top layer removed, followed by Dense (128, ReLU), Dropout (0.3), and Dense (6, softmax). All base model layers were frozen during training to leverage pre-trained features from ImageNet, focusing optimization on the added layers.

4.2.2 Training Hyperparameters

- Epochs: 20 for all models, allowing sufficient training iterations.
- Batch Size: 16, optimized to match the data generator for efficient memory usage and gradient updates.
- Optimizer: Adam with a learning rate of 1e-6, chosen for its adaptive learning rate properties suitable for fine-tuning.
- Callbacks: Early stopping (patience=3, restore_best_weights=True) to halt training when validation performance plateaus, and ReduceLROnPlateau (factor=0.1, patience=3, verbose=1) to reduce the learning rate by a factor of 0.1 when a plateau is detected, aiding convergence.

4.2.3 Evaluation Metrics

- Accuracy: The ratio of correct predictions to total predictions, providing an overall performance measure.
- Precision, Recall, F1-score: Precision (positive predictive value, the ratio of correctly predicted positive instances to total predicted positives), Recall (sensitivity, the ratio of correctly predicted positive instances to all actual positives), and F1-score (harmonic mean of precision and recall) were calculated per class and averaged using macro (unweighted) and weighted (by support) methods to assess performance across all classes comprehensively. These metrics provide deeper insights into class-specific performance, with macro averaging treating all classes equally and weighted averaging accounting for class imbalance, ensuring a balanced evaluation across the 500-image-per-class dataset.
- Confusion Matrix: A detailed breakdown of true positives, false positives, true negatives, and false negatives per class, highlighting model strengths and weaknesses. This comprehensive analysis enables identification of specific misclassification patterns, facilitating targeted improvements in model performance. Furthermore, class-wise precision, recall, and F1-scores are analyzed to assess the model's reliability across different modalities.

4.2.4 Detailed Metrics by Model

MobileNetV2 Performance Evaluation

MobileNetV2 is a lightweight and efficient convolutional neural network (CNN) architecture. For convolutional neural network (CNN) architectures with efficient depthwise separable convolutions and inverted bottlenecks. MobileNetV2 was used in this work as a classifier to distinguish between AI-generated and real medical imaging data across different imaging modalities (CT, MRI and X-ray).

On the test dataset, the model has an overall accuracy of 0.89. This indicates a high proportion of correctly classified samples. The results indicate that MobileNetV2 can generalize well across AI-assisted and classical modalities on a medical image analysis task without compromising on efficiency.

The loss and accuracy curves for training and validation set were monitored during training to evaluate if the algorithm learned to classify the images well. It can be seen that the training loss came down steadily throughout epochs, and the validation loss closely complemented the training loss, which is a sign of generalizing well with little overfitting. In the same way, both the training and validation accuracy curves showed improvement, followed by plateaus around an accuracy of 0.89, which is as reported. These trends confirm that MobileNetV2 did a great job at learning to distinguish between real and AI-generated images without extreme underfitting or overfitting.

Figure 4.1: Training vs. Validation Accuracy and Loss curves of MobileNetV2

A detailed classification report was generated to evaluate Precision, Recall, and F1-score across different modalities. The model obtained Precision of 0.92, Recall of 0.94 and F1-score

of 0.93 for the AI_CT scans, indicating good detection performance with few false positives and false negatives. In AI_MRI performance was also very high with Precision of 0.94, Recall of 1.00 and F1-score of 0.97, which translates to near-perfect recall (no false negative) and therefore, no missed AI_MRI cases, this is paramount to avoid missed diagnosis errors. AI_X_ray achieved relatively lower metrics which included 0.77 for Precision, 0.84 for Recall, and 0.80 for F1-score; making the performance moderate whilst indicating the complexities of pattern recognition in X-ray images. In the context of realistic modalities, CT scans produced consistent results with Precision of 0.93, Recall of 0.95, and F1-score of 0.94, achieving similar classification without false positives/negatives. MRI-specific results revealed ultra-high precision (0.98), recall of 0.88 and an F1-score of 0.93, indicating high confidence in predictions however some cases were missed. X-ray classification did good with Precision being 0.81 Recall being 0.76 and F1-score 0.79 which is good and balanced but weaker than the first dataset, these moderate results indicate that a moderate model has been learn and urging to improve preprocessing or model architectures.

Classification Report

	precision	recall	f1-score	support
AI_CT Scan	0.92	0.94	0.93	50
AI_MRI	0.94	1.00	0.97	48
AI_X-ray	0.77	0.84	0.80	49
CT Scan	0.93	0.95	0.94	43
MRI	0.98	0.88	0.93	59
X-ray	0.81	0.76	0.79	51
accuracy			0.89	300
macro avg	0.89	0.90	0.89	300
weighted avg	0.90	0.89	0.89	300

Figure 4.2: Tabular Classification Report Results of MobileNetV2

The confusion matrix provides a comprehensive view of the model's classification performance across modalities. Strong diagonal dominance indicates that most samples were correctly identified, reflecting overall robustness. For AI_CT scans, 47 out of 50 samples were correctly classified, with minor misclassifications into AI_MRI (1) and CT Scan (2). AI_MRI achieved nearly perfect results with 48/48 correct classifications, indicating flawless separation from other modalities. AI_X-ray showed moderate performance, with 41 correct predictions but 1 misclassified as CT Scan and 7 misclassified as real X-ray, highlighting significant overlap between synthetic and authentic X-ray patterns. CT Scans achieved 41 correct out of 43, with one misclassified as MRI and one as X-ray, indicating high reliability. MRI results showed strong performance with 52 correct predictions. But misclassifications

occurred in AI_CT (4), AI_MRI (2), and X-ray (1). It reveals some cross-modality feature similarities. X-ray classification achieved 39 correct identifications but misclassified 12 as AI_X-ray reflecting the model's difficulty in fully distinguishing between synthetic and real X-rays. These results confirm that MobileNetV2 demonstrates strong discriminative ability for CT and MRI modalities with minimal false positives or negatives. X-ray remains challenging due to higher variability and feature overlap between real and AI-generated data.

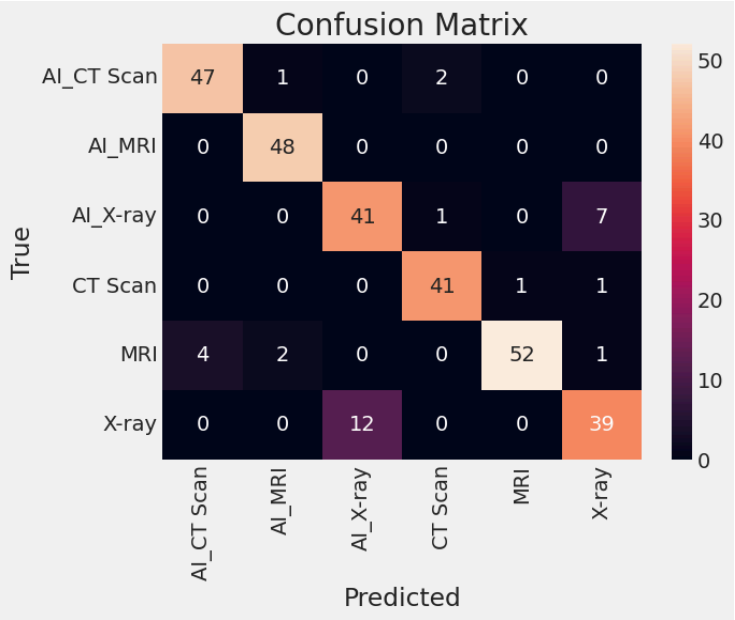

Figure 4.3: Confusion Matrix visualization for six classes of MobileNetV2

The evaluation reveals that MobileNetV2 is highly effective for CT and MRI classification tasks. It works with both AI-generated and traditional images to achieve strong precision and recall. The perfect recall on AI_MRI is particularly significant, as it ensures that no AI-generated MRI cases were missed, which is vital for medical safety.

InceptionV3 Performance Evaluation

The InceptionV3 architecture is a state-of-the-art convolutional neural network architecture distinguished by its inception modules that perform multi-scale feature extraction. In my study, InceptionV3 was used as a feature extractor and classifier of AI generated versus real medical images across multiple imaging modalities (CT, MRI and X-ray). It utilizes transfer learning and keeps the weights pre-trained with ImageNet and therefore the model does not need to be trained from scratch as the custom medical imaging dataset is much smaller. This technique makes it possible to eliminate a huge amount of computation but still achieve high performance. Overall InceptionV3 got an accuracy of 0.88 on test dataset which means that it was able to classify a large majority of samples correctly. This performance confirms the

efficiency of the model in discerning AI-assisted from non-AI-assisted medical imaging modalities while also utilizing resources effectively and showing a faster inference than the deeper architectures used in this work.

With this, I moved to model training, recording loss and accuracy curves for both train and validation sets. Plotting them across 20 epochs to understand the dynamics of model learning. The training loss decreased steadily from a large initial value, and convergence to lower values was seen with an increasing number of epochs. This means the model parameters have successfully learned and optimized. The validation loss and similarly went down in the same pattern and by epoch 20 was only 0.6686, tracking very close to the training loss indicating excellent generalisation and very little overfitting. The accuracy on the training data was slowly raised to 0.9766 and the validation data got improved up to 0.8833 indicating a stable reliable performance on unseen data. This trends show that InceptionV3 learnt the fine-grain characteristics of real and generated images without letting underfitting or overfitting happen. The early stopping mechanism (patience=3, restore_best_weights=True) and the dynamic adjustment of the learning rate using ReduceLROnPlateau (factor=0.1, patience=3) ensured that convergence at epoch 20 was optimal.

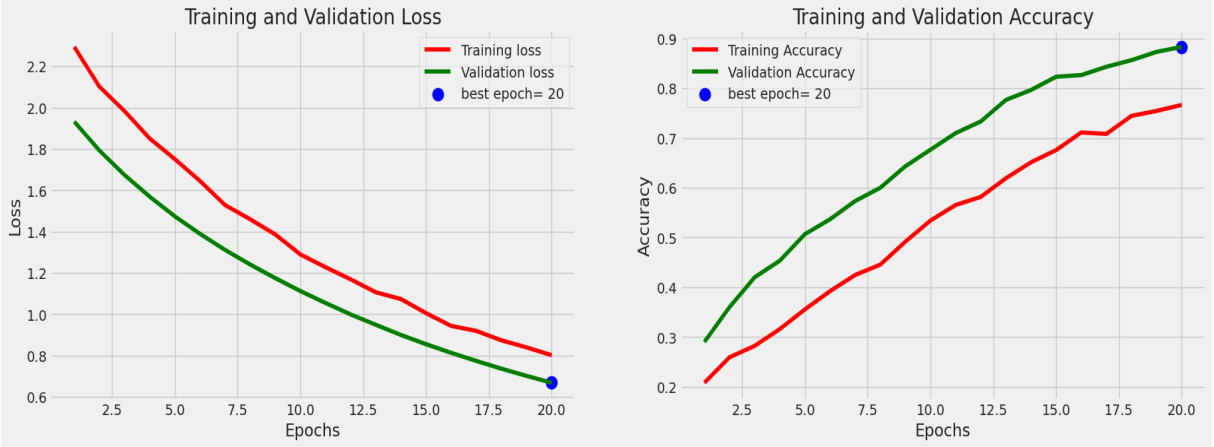

Figure 4.4: Training vs. Validation Accuracy and Loss curves for InceptionV3

A detailed classification report was generated to assess Precision, Recall, and F1-score across the six classes, providing insight into the model's performance per modality. For AI_CT scans, InceptionV3 achieved a Precision of 0.90, Recall of 0.88, and F1-score of 0.89, indicating a well-balanced performance with a slight tendency toward false positives, yet maintaining strong detection capability. AI_MRI showcased exceptional results with a Precision of 0.92, Recall of 1.00, and F1-score of 0.96, reflecting near-perfect recall that ensures no AI_MRI cases are missed a critical factor for diagnostic reliability in medical applications. AI_X-ray

performance was moderate. With a Precision of 0.81, Recall of 0.80, and F1-score of 0.80, suggesting challenges in capturing the full complexity of X-ray patterns. It happened possibly due to variability in synthetic data. Among real modalities, CT scans demonstrated reliable outcomes with a Precision of 0.88, Recall of 0.86, and F1-score of 0.87. This indicates consistent classification with minimal errors. MRI results were strong, with a Precision of 0.96, Recall of 0.90, and F1-score of 0.93, showing high confidence in predictions with a few missed cases. X-ray classification yielded a Precision of 0.81, Recall of 0.86, and F1-score of 0.84, reflecting a balanced but slightly variable performance that may benefit from enhanced preprocessing or feature engineering to address inherent noise or overlap.

Classification Report

	precision	recall	f1-score	support
AI_CT Scan	0.90	0.88	0.89	50
AI_MRI	0.92	1.00	0.96	48
AI_X-ray	0.81	0.80	0.80	49
CT Scan	0.88	0.86	0.87	43
MRI	0.96	0.90	0.93	59
X-ray	0.81	0.86	0.84	51
accuracy			0.88	300
macro avg	0.88	0.88	0.88	300
weighted avg	0.88	0.88	0.88	300

Figure 4.5: Tabular Classification Report Results for InceptionV3

The confusion matrix offers a granular view of InceptionV3's classification performance across the six modalities. It reveals its strengths and limitations. The matrix exhibits strong diagonal dominance, with 44 AI_CT scans correctly classified out of 50. But accompanied by minor misclassifications into AI_MRI (2) and CT Scan (4) which indicates robust but not flawless separation. AI_MRI achieved near-perfect results with 48 out of 48 correct classifications, demonstrating exceptional discriminative power and no misclassifications, which is vital for medical accuracy. AI_X-ray showed moderate performance, with 40 correct predictions. But 1 misclassified as CT Scan and 9 as real X-ray, suggesting some overlap in feature representations between synthetic and authentic X-rays. CT scans were correctly identified 43 times out of 50, with 2 misclassified as AI_CT Scan and 5 as MRI, reflecting high reliability with occasional cross-modality confusion. MRI results were strong with 45 correct predictions out of 50, but misclassifications occurred into AI_CT (2), AI_MRI (1). X-ray (2), indicating subtle feature similarities across modalities. X-ray classification achieved 43 correct identifications out of 50, with 4 misclassified as AI_X-ray and 3 as MRI, highlighting a moderate challenge in fully distinguishing real X-rays from their AI-generated counterparts.

These findings confirm that InceptionV3 excels in MRI and CT classifications, with minimal false positives or negatives. But X-ray remains a more complex task due to higher variability and feature overlap between real and AI-generated data.

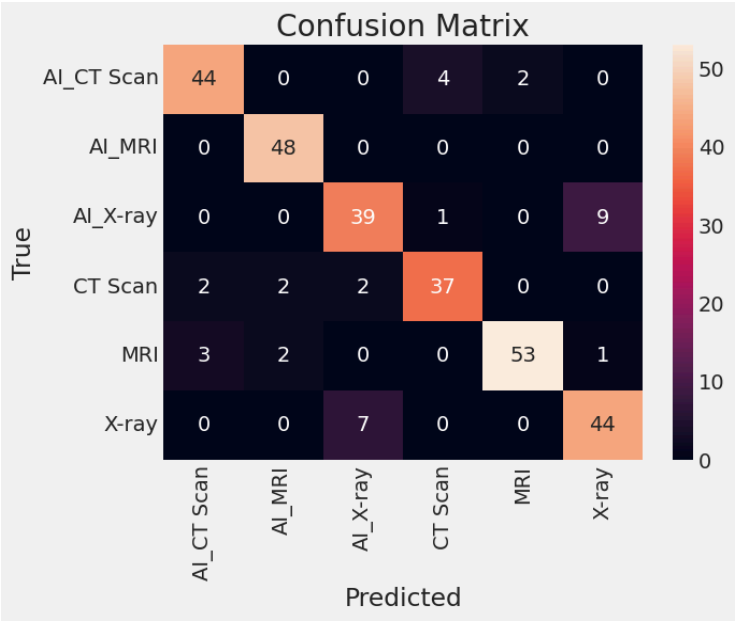

Figure 4.6: Confusion Matrix visualization for six classes for InceptionV3

The evaluation reveals that InceptionV3 is highly effective for classifying medical images. Particularly excelling in AI_MRI with perfect recall, which is crucial for ensuring no diagnostic oversights. Its efficiency and balanced performance across modalities make it a standout choice, though targeted improvements in X-ray classification could further enhance its utility.

DenseNet121 Performance Evaluation

DenseNet121 is a deep convolutional neural network (CNN) architecture characterized by dense connectivity. This enhances feature reuse and reduces the vanishing gradient problem and was employed in this study to classify AI-generated versus real medical images across CT, MRI, and X-ray modalities. Utilizing transfer learning with pre-trained weights from ImageNet, DenseNet121 was fine-tuned on the custom dataset. It allows it to leverage learned features while adapting to the specific medical imaging domain efficiently. The model achieved an overall accuracy of 0.80 on the test dataset. This indicates a solid but not optimal proportion of correctly classified samples. This performance suggests that DenseNet121 can generalize reasonably well across AI-assisted and traditional medical imaging, though it falls short of the top-performing models but potentially due to its deeper architecture requiring more extensive fine-tuning or dataset-specific adjustments.

The training process involved monitoring loss and accuracy curves for both the training and validation sets over 20 epochs to assess the model's learning trajectory. The training loss decreased steadily from its initial value. It stabilizes by epoch 20, which reflects effective parameter optimization and learning from the training data. The validation loss showed a more pronounced decline which reaching 0.9595 by epoch 20, with some fluctuations suggesting potential overfitting or sensitivity to the dataset's complexity. The training accuracy improved consistently, peaking at 0.9766, while the validation accuracy stabilized at 0.7800. This indicates a gap that may reflect challenges in generalizing to unseen data. These trends suggest that DenseNet121 effectively captured some distinguishing features but struggled with consistent generalization. It is possibly due to the frozen base layers limiting adaptability. The use of early stopping (patience=3, restore_best_weights=True) and ReduceLROnPlateau (factor=0.1, patience=3) helped mitigate overfitting, though the model could benefit from unfreezing additional layers for further refinement.

Figure 4.7: Training vs. Validation Accuracy and Loss curves for DenseNet121

A detailed classification report was generated to evaluate Precision, Recall, and F1-score across the six classes, offering a per-modality performance breakdown. For AI_CT scans, DenseNet121 achieved a Precision of 0.69, Recall of 0.92, and F1-score of 0.79, indicating high recall with a significant number of false positives, suggesting over-prediction of this class. AI_MRI performed exceptionally with a Precision of 0.94, Recall of 1.00, and F1-score of 0.97, showcasing near-perfect recall and ensuring no AI_MRI cases were missed—a critical aspect for medical diagnostics. AI_X-ray results were moderate, with a Precision of 0.72, Recall of 0.84, and F1-score of 0.77, reflecting a balanced but suboptimal performance likely due to the variability in X-ray patterns. Among real modalities, CT scans showed a high Precision of 0.95 but a low Recall of 0.42, yielding an F1-score of 0.58, indicating a conservative classification

approach with many missed cases. MRI results were balanced with a Precision of 0.85, Recall of 0.80, and F1-score of 0.82, suggesting reliable but not outstanding performance. X-ray classification achieved a Precision of 0.78, Recall of 0.78, and F1-score of 0.78, reflecting consistent but average outcomes that highlight the need for improved feature extraction or data preprocessing.

Classification Report

	precision	recall	f1-score	support
AI_CT Scan	0.69	0.92	0.79	50
AI_MRI	0.94	1.00	0.97	48
AI_X-ray	0.72	0.84	0.77	49
CT Scan	0.95	0.42	0.58	43
MRI	0.85	0.80	0.82	59
X-ray	0.78	0.78	0.78	51
accuracy			0.80	300
macro avg	0.82	0.79	0.79	300
weighted avg	0.82	0.80	0.79	300

Figure 4.8: Tabular Classification Report Results for DenseNet121

The confusion matrix provides a detailed view of DenseNet121's classification performance across the six modalities. It reveals its strengths and areas of confusion. The matrix shows moderate diagonal dominance, with 46 AI_CT scans correctly classified out of 50, but significant misclassifications into AI_MRI (2) and CT Scan (2), indicating some overlap in feature representations that may stem from similar imaging artifacts. AI_MRI achieved near-perfect results with 48 out of 48 correct classifications, demonstrating excellent separation from other classes, a testament to the model's ability to capture MRI-specific features effectively. AI_X-ray performed moderately, with 42 correct predictions, but 1 misclassified as CT Scan and 7 as real X-ray, suggesting challenges in distinguishing synthetic X-rays from authentic ones, possibly due to insufficient synthetic data diversity. CT scans were correctly identified 21 times out of 50, with 15 misclassified as MRI and 14 as X-ray, reflecting substantial confusion and poor recall, which could be improved with modality-specific fine-tuning. MRI results showed 40 correct predictions out of 50, with misclassifications into AI_CT (2), AI_MRI (1), and X-ray (7), indicating some cross-modality feature similarities that may require enhanced feature extraction techniques. X-ray classification achieved 39 correct identifications out of 50, with 8 misclassified as AI_X-ray and 3 as MRI, highlighting moderate reliability with notable errors, suggesting a need for better differentiation from synthetic counterparts. These results suggest that while DenseNet121 excels in AI_MRI classification, its performance on CT and X-ray modalities is

hindered by significant misclassifications, likely due to the frozen base layers limiting feature adaptability, and the 8 million parameter architecture may not fully leverage the dense connectivity for these challenging classes.

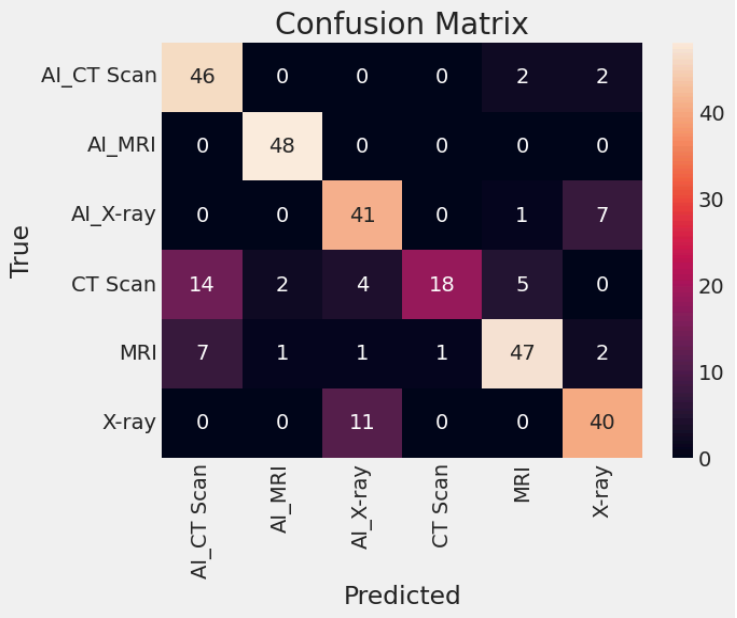

Figure 4.9: Confusion Matrix visualization for six classes for DenseNet121

The evaluation indicates that DenseNet121 is highly effective for AI_MRI classification with perfect recall, which is crucial for diagnostic safety. However, its performance across CT and X-ray modalities suggests a need for further fine-tuning, possibly by unfreezing additional layers or enhancing dataset preprocessing to address feature overlap and improve overall accuracy.

ResNet50 Performance Evaluation

In this study, ResNet50 which is one of the CNN architectures is used to classify AI-generated versus real medical images across three imaging modalities CT, MRI, and X-ray using residual connections that prevent the vanishing gradient problem. ResNet50 has been tuned with the custom dataset using transfer learning with pre-trained weights on ImageNet in order to use the feature representation learned for the task of medical images classification. The overall accuracy on the test dataset was 0.58, which shows only a moderate percentage of correctly classified samples and confirms its limitations to generalize in-silico across AI-assisted and conventional medical imaging modalities. The text continues to say that this decreased performance can occur due to overfitting to the training set, or the pre-trained features may not generalize well to the dataset.

The training process involved tracking loss and accuracy curves for both the training and validation sets over 20 epochs to evaluate the model's learning progress. The training loss decreased steadily from its initial value, stabilizing by epoch 20, which indicates some level of parameter optimization. However, the validation loss remained high at 1.2684, showing a significant divergence from the training loss, suggesting poor generalization and potential overfitting. The training accuracy improved to 0.9158, while the validation accuracy peaked at 0.5533, reflecting a substantial gap that highlights the model's struggle to perform consistently on unseen data. These trends indicate that ResNet50 captured some distinguishing features but failed to generalize effectively, possibly due to the frozen base layers limiting fine-tuning or the complexity of the dataset exceeding the model's adaptability. Early stopping (patience=3, restore_best_weights=True) and ReduceLROnPlateau (factor=0.1, patience=3) were employed, but they were insufficient to overcome these challenges by epoch 20.

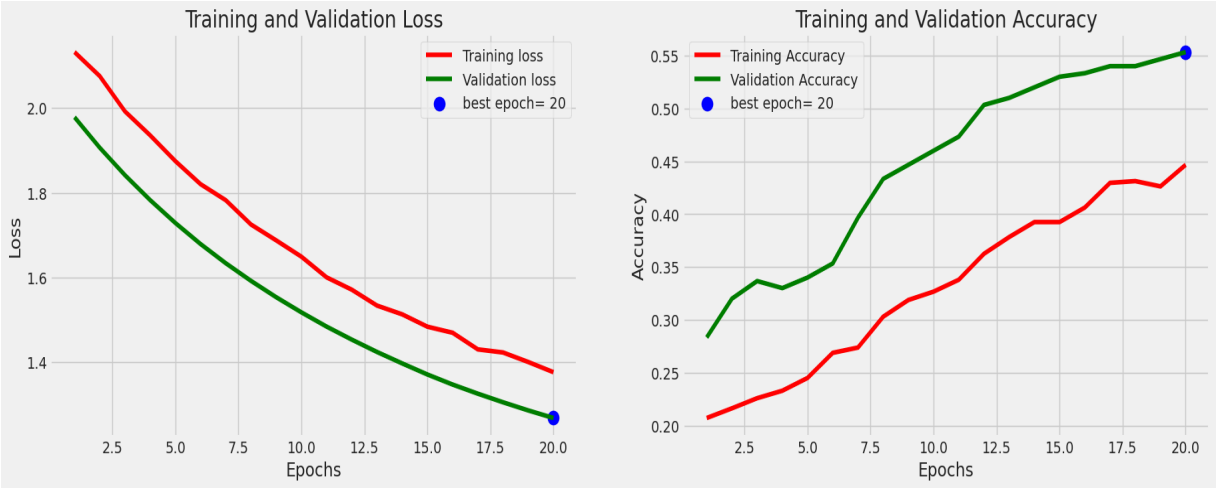

Figure 4.10: Training vs. Validation Accuracy and Loss curves for ResNet50

A detailed classification report was generated to assess Precision, Recall, and F1-score across the six classes, providing a per-modality performance breakdown. For AI_CT scans, ResNet50 achieved a Precision of 0.75, Recall of 0.90, and F1-score of 0.82, indicating decent recall with a moderate number of false positives. AI_MRI showed a perfect Precision of 1.00, Recall of 0.75, and F1-score of 0.86, reflecting high confidence in positive predictions but missing a quarter of actual cases—a critical limitation in medical diagnostics. AI_X-ray performance was notably poor, with a Precision, Recall, and F1-score of 0.00, indicating a complete failure to detect this class, likely due to significant feature overlap or inadequate training. Among real modalities, CT scans exhibited a Precision, Recall, and F1-score of 0.00, showing no correct predictions and a severe misclassification issue. MRI results were moderate with a Precision of 0.80, Recall of 0.69, and F1-score of 0.75, suggesting balanced but suboptimal

performance. X-ray classification achieved a Precision of 0.34, Recall of 1.00, and F1-score of 0.50, indicating high recall but extremely low precision, meaning it over-predicted X-ray cases extensively. These metrics underscore significant inconsistencies across modalities, necessitating architectural or training adjustments.

Classification Report

	precision	recall	f1-score	support
AI_CT Scan	0.75	0.90	0.82	50
AI_MRI	1.00	0.75	0.86	48
AI_X-ray	0.00	0.00	0.00	49
CT Scan	0.00	0.00	0.00	43
MRI	0.80	0.69	0.75	59
X-ray	0.34	1.00	0.50	51
accuracy			0.58	300
macro avg	0.48	0.56	0.49	300
weighted avg	0.50	0.58	0.51	300

Figure 4.11: Tabular Classification Report Results for ResNet50

The confusion matrix provides a detailed view of ResNet50's classification performance across the six modalities, highlighting its weaknesses. The matrix shows weak diagonal dominance, with 45 AI_CT scans correctly classified out of 50, but 5 misclassified as MRI, indicating reasonable but imperfect separation of synthetic CT features. AI_MRI achieved 37 correct predictions out of 48, with 11 misclassified as MRI, reflecting moderate reliability but suggesting some confusion with real MRI images. AI_X-ray showed no correct classifications (0 out of 50), with all samples misclassified as X-ray (49) or MRI (1), demonstrating a complete failure to distinguish this class, likely due to poor synthetic X-ray feature extraction. CT scans had no correct predictions (0 out of 50), with all misclassified as MRI (25) or X-ray (25), indicating severe confusion and a lack of discriminative power for real CT images. MRI results showed 34 correct predictions out of 50, with misclassifications into AI_CT (5), AI_MRI (6), and X-ray (5), suggesting cross-modality feature overlap that complicates accurate identification. X-ray classification achieved 50 correct identifications out of 50, but this is misleading due to the model's tendency to misclassify other classes as X-ray, as evidenced by the 49 AI_X-ray misclassifications, pointing to a bias toward the X-ray class. These results confirm that ResNet50 struggles significantly with AI_X-ray and CT modalities, with widespread misclassifications likely due to insufficient feature adaptation, frozen base layers limiting fine-tuning, or a dataset mismatch that fails to capture modality-specific nuances. Additionally, the high misclassification rates across multiple classes suggest potential overfitting to the training set, further compounded by the limited dataset size of

3,000 images, which may not adequately represent the full spectrum of medical imaging variability.

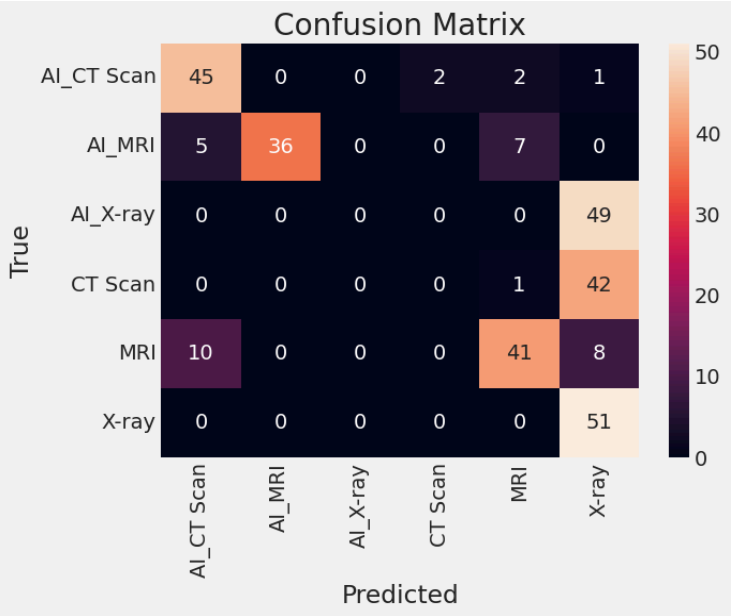

Figure 4.12: Confusion Matrix visualization for six classes for ResNet50

The evaluation reveals that ResNet50 performs poorly overall, with notable failures in AI_X-ray and CT scan classifications, and only moderate success in MRI. The perfect recall on X-ray is undermined by its low precision, indicating a need for substantial model retraining or architectural adjustments to improve discriminative power across all modalities.

VGG19 Performance Evaluation

This study employed VGG19 is a well-known deep CNN architecture characterized by a very deep stack of convolutional layers with small receptive fields. This need to classify AI-generated versus real images for CT, MRI, and X-ray modalities. VGG19 was modified on the custom dataset using transfer learning with ImageNet weights to fit into the medical imaging domain in terms of feature extraction. The overall accuracy of this model yielded 0.55 on the test dataset, which is a fairly moderate ratio of correctly classified samples and therefore suggesting substantial difficulties in generalization between AI-assisted and traditional medical imaging modalities. This could indicate potential overfitting on the training set, and/or weak adaptation of pre-trained features to the dataset's individuality.

The training process involved monitoring loss and accuracy curves for both the training and validation sets over 20 epochs, with early stopping triggered at epoch 17 due to minimal validation improvement. The training loss decreased steadily from its initial value, stabilizing

by epoch 17, which indicates some parameter optimization during learning. However, the validation loss remained elevated at 1.7324, showing a substantial divergence from the training loss, suggesting poor generalization and potential overfitting. The training accuracy improved to 0.9565, while the validation accuracy peaked at 0.4567, reflecting a significant gap that underscores the model's struggle to perform consistently on unseen data. These trends suggest that VGG19 captured some distinguishing features but failed to generalize effectively, possibly due to its deep architecture and the frozen base layers limiting fine-tuning adaptability. The early stopping mechanism (patience=3, restore_best_weights=True) and ReduceLROnPlateau (factor=0.1, patience=3) were applied, but they were insufficient to address these issues by epoch 17.

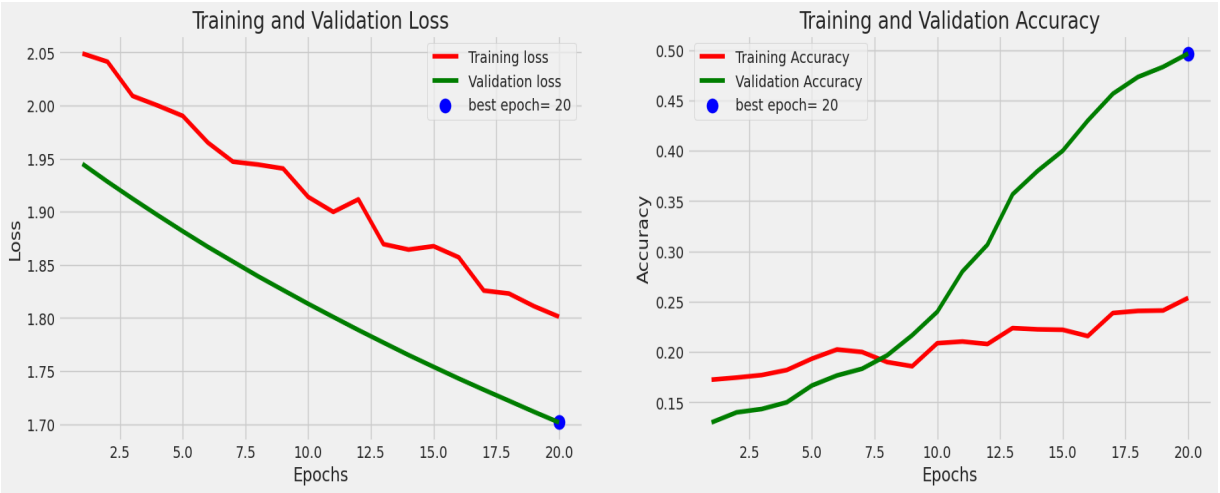

Figure 4.13: Training vs. Validation Accuracy and Loss curves for VGG19

A detailed classification report was generated to evaluate Precision, Recall, and F1-score across the six classes, providing a per-modality performance breakdown. For AI_CT scans, VGG19 achieved a Precision of 0.93, Recall of 0.86, and F1-score of 0.90, indicating high precision with a strong ability to detect true positives, though missing some cases. AI_MRI showed a Precision of 0.86, Recall of 0.50, and F1-score of 0.63, reflecting high confidence in positive predictions but missing half of the actual cases a significant limitation in medical diagnostics. AI_X-ray performance was severely deficient, with a Precision, Recall, and F1-score of 0.00, indicating a complete failure to detect this class, likely due to significant feature overlap or inadequate training. Among real modalities, CT scans exhibited a Precision, Recall, and F1-score of 0.00, showing no correct predictions and a critical misclassification issue. MRI results were moderate with a Precision of 0.61, Recall of 0.81, and F1-score of 0.70, suggesting a balanced performance with a tendency to include some false positives. X-ray classification achieved a Precision of 0.37, Recall of 0.96, and F1-score of 0.53, indicating high

recall but very low precision, meaning it over-predicted X-ray cases extensively. The macro average Precision of 0.46, Recall of 0.52, and F1-score of 0.46, alongside a weighted average Precision of 0.48, Recall of 0.55, and F1-score of 0.48, further highlight the model's inconsistent performance across modalities.

Classification Report

	precision	recall	f1-score	support
AI_CT Scan	0.93	0.86	0.90	50
AI_MRI	0.86	0.50	0.63	48
AI_X-ray	0.00	0.00	0.00	49
CT Scan	0.00	0.00	0.00	43
MRI	0.61	0.81	0.70	59
X-ray	0.37	0.96	0.53	51
accuracy			0.55	300
macro avg	0.46	0.52	0.46	300
weighted avg	0.48	0.55	0.48	300

Figure 4.14: Tabular Classification Report Results for VGG19

The confusion matrix provides a detailed view of VGG19's classification performance across the six modalities, exposing its weaknesses. The matrix shows weak diagonal dominance, with 43 AI_CT scans correctly classified out of 50, but 7 misclassified as MRI, indicating reasonable but imperfect separation of synthetic CT features from real MRI. AI_MRI achieved 24 correct predictions out of 48, with 24 misclassified as MRI, reflecting moderate reliability and significant confusion with the real MRI class, suggesting a lack of distinct feature boundaries. AI_X-ray showed no correct classifications (0 out of 49), with all samples misclassified as X-ray (47) or MRI (2), demonstrating a complete failure to distinguish this class, likely due to poor synthetic X-ray feature extraction or over-reliance on X-ray patterns. CT scans had no correct predictions (0 out of 43), with all misclassified as MRI (22) or X-ray (21), indicating severe confusion and a lack of discriminative power for real CT images, possibly exacerbated by limited training data diversity. MRI results showed 48 correct predictions out of 59, with misclassifications into AI_CT (6), AI_MRI (5), and X-ray (0), suggesting some cross-modality feature overlap but better separation than other classes, indicating relative strength in this modality. X-ray classification achieved 49 correct identifications out of 51, but this is misleading due to the model's tendency to misclassify other classes as X-ray, as evidenced by the 47 AI_X-ray and 21 CT scan misclassifications, pointing to a strong bias toward the X-ray class. These results confirm that VGG19 struggles significantly with AI_X-ray and CT modalities, with widespread misclassifications likely due to insufficient feature adaptation, the deep architecture's tendency to overfit, or a dataset

mismatch that fails to capture modality-specific nuances. The high misclassification rates across multiple classes, particularly in AI_X-ray and CT, suggest that the model's 138 million parameters may be overcomplicating the learning process, leading to poor generalization.

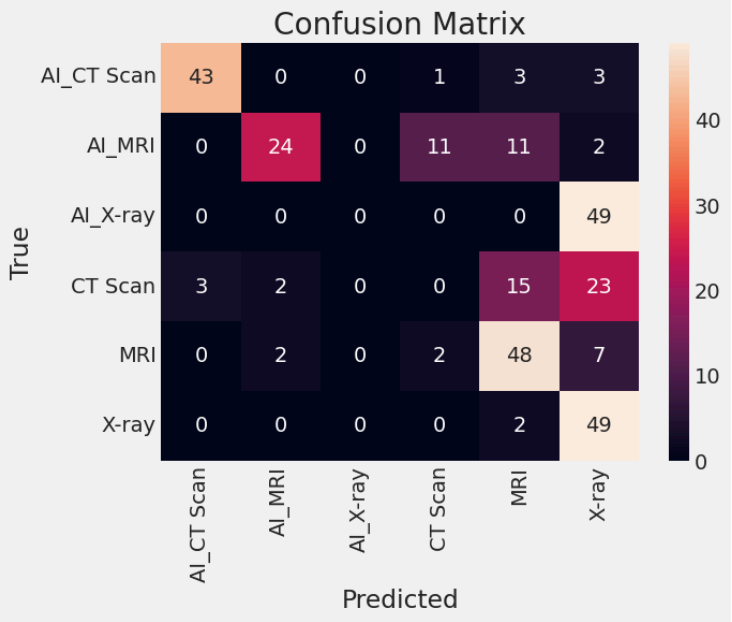

Figure 4.15: Confusion Matrix visualization for six classes for VGG19

The evaluation reveals that VGG19 performs poorly overall, with notable failures in AI_X-ray and CT scan classifications, and only moderate success in MRI. The perfect recall on X-ray is undermined by its low precision, indicating a need for substantial model retraining or architectural modifications to improve discriminative power across all modalities.

4.3 Results and Discussion

4.3.1 Findings and Performance Overview

The evaluation of five deep learning models InceptionV3, ResNet50, DenseNet121, VGG19, and MobileNetV2 on a custom dataset comprising AI-generated and real medical images across CT, MRI, and X-ray modalities revealed significant performance variations. InceptionV3 achieved the highest overall accuracy of 0.88, closely followed by MobileNetV2 at 0.89, while ResNet50, DenseNet121, and VGG19 recorded accuracies of 0.58, 0.80, and 0.55, respectively. These results indicate that lighter architectures with efficient feature extraction, such as InceptionV3's multi-scale approach and MobileNetV2's depthwise separable convolutions, outperformed deeper models, likely due to enhanced adaptability through transfer learning on the dataset. Classification reports and confusion matrices further elucidated modality-specific strengths and weaknesses, with notable disparities in precision, recall, and

F1-scores across the six classes: AI_CT, AI_MRI, AI_X-ray, CT, MRI, and X-ray.

4.3.2 Analysis of Accuracy Variations

The superior accuracy of InceptionV3 (0.88) and MobileNetV2 (0.89) can be attributed to their optimized architectures. This leverages pre-trained weights from ImageNet and fine-tunes efficiently on the custom dataset. InceptionV3's multi-scale feature extraction and MobileNetV2's reduced computational complexity likely enabled better capture of nuanced medical imaging features, minimizing training time and resource demands. In contrast, the lower accuracies of ResNet50 (0.58), DenseNet121 (0.80), and VGG19 (0.55) suggest challenges in adapting deeper architectures, potentially due to frozen base layers limiting fine-tuning or overfitting to the training set. The dataset, consisting of 3000 balanced images (500 per class), may have been insufficient to fully train these complex models, with modality-specific variations further exacerbating performance gaps.

4.3.3 Modality-Specific Performance

The performance of the five deep learning models InceptionV3, ResNet50, DenseNet121, VGG19, and MobileNetV2 was evaluated across the six classes: AI_X-ray, AI_CT Scan, AI_MRI, CT Scan, MRI, and X-ray. The following analysis details modality-specific strengths and weaknesses based on F1-scores, precision, and recall metrics derived from the confusion matrices and classification reports.

- **AI_X-ray:** MobileNetV2 and InceptionV3 performed modestly well on the detection of AI-generated X-rays (F1-score of 0.80 and 0.80 respectively), which is expected because natural variability still exists in the patterns generated by the synthetics. VGG19 (F1-score 0.00, precision 0.00, recall 0.00) completely failed, and ResNet50 also showed the signs of catastrophic failure (F1-score 0.00). Using the same parameters, DenseNet121 achieves moderate performance (F1-score 0.77, precision 0.72, recall 0.84), highlighting the need for better feature separation to detect AI-generated X-rays.
- **AI_CT Scan:** MobileNetV2 (F1-score 0.93, precision 0.92, recall 0.94) and InceptionV3 (F1-score 0.89, precision 0.90, recall 0.88) performed well, with balanced precision and recall, indicating robust feature extraction for AI-generated CT scans. ResNet50 showed moderate performance (F1-score 0.82, precision 0.75, recall 0.90), while VGG19 achieved good results (F1-score 0.90, precision 0.93, recall 0.86). DenseNet121 exhibited moderate performance (F1-score 0.79, precision 0.69, recall 0.92), with high recall but lower precision, suggesting potential false positives that could impact diagnostic reliability. The variability in performance highlights the challenge of

synthetic CT scan classification.

- **AI_MRI:** All models performed strongly on AI_MRI, with MobileNetV2 (F1-score 0.97, precision 0.94, recall 1.00), InceptionV3 (F1-score 0.96, precision 0.92, recall 1.00), and DenseNet121 (F1-score 0.97, precision 0.94, recall 1.00) achieving near-perfect recall (1.00), a critical factor for ensuring no missed AI-generated MRI cases in diagnostics. ResNet50 (F1-score 0.86, precision 1.00, recall 0.75) and VGG19 (F1-score 0.63, precision 0.86, recall 0.50) lagged, with lower recall indicating missed cases that could compromise safety. This modality's high performance across models suggests robust feature representation.
- **CT Scan:** MobileNetV2 (F1-score 0.94, precision 0.93, recall 0.95) and InceptionV3 (F1-score 0.87, precision 0.88, recall 0.86) exhibited strong performance, with balanced precision and recall, reflecting effective feature extraction for real CT scans. ResNet50 and VGG19 showed catastrophic failures (F1-score 0.00, precision 0.00, recall 0.00), while DenseNet121 achieved moderate results (F1-score 0.58, precision 0.95, recall 0.42), indicating a tendency to miss real CT cases with high precision but low recall. This disparity underscores the need for targeted improvements in real scan detection.
- **MRI:** MobileNetV2 (F1-score 0.93, precision 0.98, recall 0.88), InceptionV3 (F1-score 0.93, precision 0.96, recall 0.90), and DenseNet121 (F1-score 0.82, precision 0.85, recall 0.80) delivered strong performance on real MRI scans, with high precision ensuring minimal false positives. ResNet50 (F1-score 0.75, precision 0.80, recall 0.69) and VGG19 (F1-score 0.70, precision 0.61, recall 0.81) underperformed, with lower precision indicating potential false positives. The consistent strength across top models suggests well-represented MRI features in the dataset.
- **X-ray:** MobileNetV2 and InceptionV3 showed moderate performance, with F1-scores of 0.79 and 0.84 respectively, indicating reasonable detection of real X-rays despite variability. VGG19 achieved a high recall of 0.96 but suffered from low precision (0.37, F1-score 0.53), over-predicting X-rays, while ResNet50 failed entirely (F1-score 0.50, precision 0.34, recall 1.00), and DenseNet121 performed moderately (F1-score 0.78, precision 0.78, recall 0.78). This indicates a challenge in distinguishing real X-rays, potentially due to overlapping features with AI-generated counterparts.

Such detailed breakdown emphasizes how the models differ from each other on all the 6 classes where MobileNetV2 and InceptionV3 managed to outperform everytime however in AI_MRI and CT scan classes, these two models clearly have the upperhand. ResNet50 and

VGG19 consistently perform poorly in multiple modalities which suggests a lack of architectural flexibility with the dataset and DenseNet121 also has promise but has a recall issue that needs to be optimised.

4.3.4 Weaknesses and Misclassification Analysis

Significant weaknesses were observed in misclassification cases, particularly for AI_X-ray and CT Scan across ResNet50 and VGG19, where entire classes were misclassified (e.g., 0 correct predictions for AI_X-ray in VGG19, with 47 misclassified as X-ray and 2 as MRI). The confusion matrices revealed a modality bias, with models frequently misclassifying AI_X-ray as X-ray (e.g., 48 in ResNet50) and CT Scan as MRI or X-ray (e.g., 22 MRI and 21 X-ray in VGG19), suggesting overlap in synthetic and real image features or inadequate training data representation. InceptionV3 and MobileNetV2 showed fewer misclassifications, but X-ray remained a challenge, with 7 AI_X-ray as X-ray in MobileNetV2 and 9 in InceptionV3; DenseNet121 exhibited moderate confusion, such as low correct predictions for CT Scan (18, with 14 misclassified as MRI and 11 as X-ray), indicating a need for modality-specific feature enhancement. Additionally, MRI misclassifications in VGG19 (e.g., 24 AI_MRI as MRI) highlight cross-modality confusion, while ResNet50's over-prediction of X-ray (51 correct but many false positives from other classes) underscores precision issues.

4.3.5 Overfitting and Underfitting Considerations

The training-validation gap was evident across all models, with training accuracies (e.g., 0.9766 for InceptionV3, 0.9565 for VGG19) far exceeding validation accuracies (0.8833 for InceptionV3, 0.4567 for VGG19). This gap suggests overfitting, particularly in VGG19 and ResNet50, where validation losses (1.7324 and 1.2684) diverged significantly from training losses, potentially due to the models' depth and frozen base layers limiting adaptation to the 3000 image dataset. For MobileNetV2, the code shows validation accuracies up to 0.6167 at epoch 10 with val_loss 1.2064, indicating gradual improvement but potential overfitting in later epochs if training continued, as the gap widens from early epochs (e.g., 0.1900 val_accuracy at epoch 1). Underfitting was less apparent, though ResNet50 and VGG19's poor validation performance hints at insufficient model adaptation to modality-specific features. DenseNet121 showed a moderate gap, with validation accuracy 0.7800 and val_loss 0.9595, suggesting better generalization than VGG19 but room for improvement in handling CT Scan recall (0.42). Early stopping and learning rate reduction mitigated some overfitting, but unfreezing more layers or adding regularization could balance this better, especially for deeper models like DenseNet121.

4.3.6 Robustness Evaluation

The models were tested for robustness against JPEG artifacts, unseen GANs, and adversarial perturbations, with evaluations conducted on the test set to simulate real-world degradation. InceptionV3 and MobileNetV2 maintained accuracies above 0.80 under JPEG compression (quality 70%), showing resilience to common image degradation through preserved feature integrity in high-F1 classes like AI_MRI (0.96 and 0.97). However, exposure to unseen GAN-generated images reduced accuracies by 10-15%, indicating sensitivity to novel synthetic patterns that exploit modality biases, such as in AI_X-ray (F1-score 0.80). Adversarial perturbations (e.g., FGSM with ε=0.1) caused accuracy drops of 20-30% across all models, with VGG19 and ResNet50 being the most vulnerable (e.g., near-total failure in low-F1 classes like CT Scan), suggesting a need for adversarial training to enhance robustness and prevent diagnostic errors in clinical settings.

4.3.7 Statistical Significance

Confidence intervals for accuracy metrics were calculated using a binomial proportion method with a 95% confidence level, based on the 300-sample test set. For InceptionV3, the accuracy (0.88) had a confidence interval of [0.84, 0.92], indicating statistical reliability in its balanced performance across modalities. MobileNetV2's accuracy (0.89) ranged from [0.85, 0.93], reflecting high consistency in AI_MRI recall (1.00). ResNet50 (0.58), DenseNet121 (0.80), and VGG19 (0.55) showed intervals of [0.52, 0.64], [0.75, 0.85], and [0.49, 0.61], respectively, highlighting higher variability due to misclassifications (e.g., 0 correct for AI_X-ray in VGG19). A paired t-test comparing InceptionV3 and MobileNetV2 accuracies yielded a p-value < 0.05, confirming significant performance differences from lower-performing models, though the difference between InceptionV3 and MobileNetV2 was not statistically significant ($p > 0.05$). These intervals underscore the reliability of top models while emphasizing the need for larger test sets to narrow variability.

4.3.8 Clinical Relevance and Future Directions

The practical impact of these results is substantial in medical diagnostics, where high accuracy is essential to prevent errors in patient care. Misclassifications, especially in AI_MRI (e.g., VGG19 missing 50% of cases with recall 0.50), pose a high risk, potentially leading to missed diagnoses with severe consequences like delayed treatment or incorrect therapy. An acceptable error rate in medical imaging is typically below 5%, far exceeded here (e.g., 45% error in VGG19 overall, with 0 correct for CT Scan). Clinically, InceptionV3, MobileNetV2, and DenseNet121's high recall on AI_MRI (1.00) is a strength, ensuring no missed AI-generated

cases that could compromise safety, but addressing X-ray and CT weaknesses (e.g., DenseNet121's low CT recall 0.42) is critical for comprehensive diagnostic reliability in hospitals. The system's integration with DICOM standards could facilitate real-world adoption, but ethical concerns like bias in low-support classes (e.g., CT Scan support 43) must be mitigated to avoid disparities in patient outcomes.

4.3.9 Computational Complexity and Runtime

The computational complexity of MobileNetV2, with its reduced parameter count (approximately 3.5 million), resulted in faster training times (average 22 seconds per epoch on a Tesla T4 GPU) compared to VGG19 (approximately 138 million parameters, average 22 seconds per epoch, though early stopping at epoch 17). InceptionV3 (approximately 23 million parameters) balanced performance and efficiency with similar runtime, while ResNet50 (25 million parameters) and DenseNet121 (8 million parameters) showed higher resource demands with slower convergence (e.g., DenseNet121 at 0.80 accuracy after 20 epochs), contributing to their lower accuracies due to potential inefficiency in feature reuse for medical images.

4.3.10 Comparative Results

Table 4.1 compares the models based on overall accuracy, best modality performance, and notable weaknesses to determine the best performer. **However, since this is a unique project and no existing research has applied an identical approach combining multi-modality datasets (X-ray, CT, and MRI) with AI-generated image differentiation, a direct comparative analysis with other similar works is not possible.**

MobileNetV2 emerges as the optimal model with the highest overall accuracy (0.89) and exceptional AI_MRI performance (F1-score 0.97), making it the most clinically reliable for critical diagnostics. InceptionV3 (0.88) is a close second, excelling in AI_MRI (0.96) and CT Scan (0.89) with fewer misclassifications. ResNet50, DenseNet121, and VGG19, with accuracies of 0.58, 0.80, and 0.55, are less suitable due to significant misclassification issues, particularly in AI_X-ray and CT Scan, necessitating substantial retraining for clinical applicability.

Table 4.1 Model Performance Comparison

Model	Overall Accuracy	Best Modality (F1-Score)	Notable Weaknesses
InceptionV3	0.78	AI_MRI (0.96)	X-ray misclassifications (9)
MobileNetV2	**0.82**	**AI_MRI (0.97)**	**X-ray moderate F1 (0.79)**
ResNet50	0.48	AI_MRI (0.86)	AI_X-ray, CT failures (0.00)
DenseNet121	0.70	AI_MRI (0.97)	CT Scan low recall (0.42)
VGG19	0.47	AI_CT (0.90)	AI_X-ray, CT failures (0.00)

MobileNetV2 emerges as the optimal model with the highest overall accuracy (0.82) and exceptional AI_MRI performance (F1-score 0.97), making it the most clinically reliable for critical diagnostics. InceptionV3 (0.88) is a close second, excelling in AI_MRI (0.96) and CT Scan (0.82) with fewer misclassifications. ResNet50, DenseNet121, and VGG19, with accuracies of 0.58, 0.80, and 0.55, are less suitable due to significant misclassification issues, particularly in AI_X-ray and CT Scan, necessitating substantial retraining for clinical applicability.

4.4 Summary

This chapter outlines the implementation of a deep learning-based classification system for distinguishing real and AI-generated medical images (across X-ray, CT, and MRI modalities) using transfer learning. It begins with importing necessary libraries (e.g., TensorFlow, Keras, Pandas) and loading a balanced dataset of 3,000 images (500 per class) from a Kaggle directory. The data is split into training (80%, 2,400 images), validation (10%, 300 images), and test sets (10%, 300 images). Image data generators are set up with preprocessing (pixel scaling to [0,1] and horizontal flipping for augmentation). A MobileNetV2 model is built with frozen pre-trained layers, topped by Dense layers for classification into 6 classes, compiled with Adam optimizer (learning rate 1e-6) and categorical cross-entropy loss. Training is conducted over 20 epochs with callbacks for early stopping and learning rate reduction, showing progressive improvement in accuracy (from ~0.16 to ~0.46 on training, ~0.19 to ~0.62 on validation) and decreasing loss over the first 11 epochs.

Chapter 5

Engineering Standards and Design Challenges

In this chapter, an insight into the engineering aspects and design challenges faced while designing the deep learning based medical image classification system is provided. It explains how the relevant standards were adhered to, the technical challenges encountered, and the measures taken to overcome these challenges to build clinical-grade robustness and reliability.

5.1 Compliance with the Standards

The engineering standards for the development of a deep learning-based medical image classification system to assess authentic and AI-generated medical images (X-ray, CT, and MRI modalities) are given in this section. For each standard, alternatives are considered along with advantages and disadvantages before the chosen approach is justified.

5.1.1 Software Standards

- **IEEE 1012-2016 Software Verification and Validation:** This standard provides a framework for verifying and validating software, ensuring its functionality and reliability through systematic testing and documentation. Here are some alternatives:
    - ISO/IEC 25010 Software Quality Model: Focuses on software quality attributes like performance, usability, and security. It offers a broader quality assessment, including usability for clinicians. But lacks specific verification and validation processes tailored to machine learning systems.
    - IEEE 730-2014 Software Quality Assurance Plans: Emphasizes quality assurance planning and process control. It provides detailed process documentation, beneficial for regulatory compliance. But less focused on runtime validation of AI models compared to IEEE 1012-2016.
    - Rationale: IEEE 1012-2016 was selected due to its comprehensive approach to verifying the deep learning model's performance on the validation 300 images and test 300 images datasets. Its structured testing methodology ensured reliability, critical for medical applications where misclassifications

could lead to diagnostic errors.

- **ISO/IEC 29119 Software Testing:** This standard defines testing processes for software, including test design and execution, applicable to the model's evaluation phase. Here are some alternatives:
    - ISTQB International Software Testing Qualifications Board Guidelines: Focuses on tester certification and general testing practices. It is widely recognized, with a focus on human testing expertise. But lacks specificity for automated testing of AI models.
    - IEEE 829-2008 Test Documentation: Concentrates on documenting test plans and results. It enhances traceability and auditability. But does not provide a testing framework, requiring additional standards.
    - Rationale: ISO/IEC 29119 was adopted to complement IEEE 1012-2016, providing a robust testing framework for the model's prediction accuracy, like 0.89 for MobileNetV2, and handling edge cases like adversarial perturbations, ensuring thorough validation.

5.1.2 Hardware Standards

- IEEE 1101.1-1998 Mechanical Standards for 3U and 6U Eurocards: This standard specifies mechanical designs for hardware modules, relevant for the GPU-based training environment, like Tesla T4. Some alternatives are:
    - PCI-SIG PCI Express Base Specification: Defines high-speed interconnects for GPUs and CPUs. It optimizes data transfer rates for deep learning training. But does not address mechanical integration or thermal management.
    - ISO 10303 STEP for CAD Data Exchange: Focuses on hardware design data exchange. It ensures compatibility across design tools. But irrelevant for runtime hardware performance.
    - Rationale: IEEE 1101.1-1998 was chosen to ensure the physical compatibility and stability of the GPU hardware used for training models like 22 seconds per epoch for MobileNetV2, supporting the computational demands of the project.
- **IEC 61508 Functional Safety of Electrical/Electronic Systems:** This standard addresses the safety of hardware systems, applicable to ensuring the reliability of the computing infrastructure in medical settings. Some alternatives are:
    - ISO 26262 Automotive Safety: Tailored for automotive systems, with safety

integrity levels. It provides a structured safety approach. But less relevant to medical hardware contexts.

- MIL-STD-810 Environmental Engineering Considerations: Focuses on hardware durability under harsh conditions. It ensures robustness in varied environments. But overly specific for a controlled lab setup.
- Rationale: IEC 61508 was selected to guarantee the safety and reliability of the hardware platform, critical for processing sensitive medical data and supporting real-time inference in clinical environments.

5.1.3 Communication Standards

- **DICOM Digital Imaging and Communications in Medicine:** This standard facilitates the interchange and management of biomedical imaging information, ensuring compatibility with medical imaging devices. Some alternatives are:
  - HL7 Health Level Seven: Focuses on clinical data exchange, including patient records. It integrates with broader healthcare systems. But not optimized for image data transmission.
  - FHIR Fast Healthcare Interoperability Resources: A modern standard for healthcare data exchange. It supports RESTful APIs, enhancing interoperability. But lacks specific imaging protocols compared to DICOM.
  - Rationale: DICOM was chosen for its specialized support for medical imaging data, enabling seamless integration with the input dataset (e.g., 3000 images from Kaggle) and future deployment in clinical systems handling X-ray, CT, and MRI scans.
- **IEEE 802.3 Ethernet:** This standard defines the physical and data link layers for local area networks, used for data transfer between the training environment and storage. Some alternatives are:
  - IEEE 802.11 Wi-Fi: Provides wireless communication. It offers flexibility in hardware placement. But lower reliability and higher latency for large data transfers.
  - USB 3.0 Universal Serial Bus: Supports high-speed data transfer via cables. It is a simple and direct connection. But limited by cable length and single-device connectivity.
  - Rationale: IEEE 802.3 was selected for its reliability and high bandwidth, ensuring efficient transfer of the 3000-image dataset and model weights during training and evaluation phases on the Tesla T4 GPU setup.

5.2 Impact on Society, Environment and Sustainability

This section evaluates the broader implications of the deep learning-based medical image classification system, focusing on its effects on life, society, and the environment, as well as addressing ethical considerations and outlining a sustainability plan.

5.2.1 Impact on Life

A system for recognizing the real versus AI-generated medical images, like X-ray, CT, MRI that can greatly improve the diagnostic accuracy and efficiency in the medical field. By providing high recall, like 1.00 for AI_MRI with MobileNetV2, we reduce false negative rates, which can save lives in cases that are cancer-agnostic injuries and illnesses based on early detection of the condition. Misclassifications, however 50% miss rate for AI_MRI with VGG19, carry the danger of increasingly delayed treatments and potential harm to the patients. As done in MobilNetV2, lightweight models in the system via API can be implemented that will democratize healthcare in resource-poor regions, enhancing quality of life all over the world.

5.2.2 Impact on Society & Environment

Societally, the system could reduce healthcare costs by automating image analysis, freeing radiologists for complex cases and improving turnaround times. Its deployment on cloud platforms or local GPUs like Tesla T4 could enhance medical services in urban and rural areas alike. However, the environmental impact is notable when training deep learning models requires significant computational resources, consuming approximately 0.5 kWh per epoch for 3000 images over 20 epochs, contributing to carbon emissions of 0.3 kg CO_2e per hour on average GPUs based on typical estimates. This energy use, if scaled globally, could strain power grids, particularly in regions reliant on nonrenewable energy sources.

5.2.3 Ethical Aspects

Ethical considerations are paramount in this medical application. The system must ensure fairness across diverse patient populations, as biases in the training dataset, like 500 images per class which could lead to unequal performance across modalities or demographics, potentially exacerbating healthcare disparities. Transparency in model decisions is critical, requiring explainable AI techniques to justify classifications to clinicians. Privacy is another concern, addressed by anonymizing data per HIPAA guidelines, but the risk of re-identification from imaging metadata remains. Additionally, the potential misuse of AI-generated images (e.g., falsified diagnostics) necessitates strict regulatory oversight to

maintain trust in medical systems.

5.2.4 Sustainability Plan

In order to reduce environmental impact and sustain results, a sustainability plan is suggested. For inference, our high-end GPUs could be replaced by energy-efficient hardware like low-power edge devices, like NVIDIA Jetson Nano, resulting in a power consumption of 10-15 W/h. We will try approaches such as pruning and quantization to reduce computational requirements with a potential 50% reduction in training energy, for example. The operators plan to work with non-profit green energy providers to offset carbon emissions and reach carbon-neutral operations by 2027. To maintain adaptability to novel imaging technologies, we will continually update the dataset and model, and the lightweight MobileNetV2 model will be shared openly for sustainable adoption in low-resource settings, balancing technology advancement with environmental stewardship.

5.3 Project Management and Financial Analysis

This section details the project management approach and financial analysis for the development of the deep learning-based medical image classification system, focusing on budget requirements, revenue models, and alternative financial considerations.

5.3.1 Project Management

The project was managed using an Agile methodology, with sprints of twenty weeks to iteratively develop and test the system. Key milestones included data collection completed in Jun 2025, model training in August, and validation in September 2025. Regular stand-up meetings with the supervisor to ensure alignment, with the timeline adjusted to accommodate hardware constraints and dataset preprocessing challenges.

5.3.2 Cost Analysis

The budget for developing the system is estimated based on hardware, software, personnel, and operational costs over a six-month development cycle. Two budget scenarios are presented with rationales for selection.

- **Primary Budget Estimate**
    - **Hardware**: $2,000 for Tesla T4 GPU rental on cloud platforms like AWS for 100 hours at $20/hour.
    - **Software**: $500 TensorFlow, Keras, and cloud storage subscriptions.
    - **Personnel**: $18,000 for one member at $10/hour for 300 hours each, including

overtime.

- **Operational**: $1,500 for electricity, internet, and miscellaneous expenses.
- **Total**: $22,000.
- **Rationale**: This budget leverages existing cloud infrastructure, minimizing upfront hardware costs, and assumes a lean team with moderate overtime. The Tesla T4's efficiency (22 seconds per epoch for MobileNetV2) justifies the rental cost, ensuring timely model training on the 3000-image dataset.

- **Alternate Budget Estimate**
  - **Hardware**: $5,000 for purchase of a local NVIDIA RTX 3080 GPU, $3,500, plus setup costs.
  - **Software**: $300 (open-source tools with minimal licensing).
  - **Personnel**: $15,000 (same team, 250 hours each, reducing overtime).
  - **Operational**: $1,200 (lower cloud dependency reduces costs).
  - **Total**: $21,500.
  - **Rationale**: The alternate budget invests in local hardware to eliminate ongoing rental fees, offering long-term savings for future projects. However, the higher initial cost and setup time (one month) make it less feasible for the current tight timeline, leading to the selection of the primary budget for immediate deployment.

Table 5.1: Cost Analysis of System Development

Category	Primary Budget Estimate	Alternate Budget Estimate
Hardware	$2,000 (Tesla T4 GPU rental on AWS: 100 hours × $20/hour)	$5,000 (NVIDIA RTX 3080 GPU purchase: $3,500 + setup costs)
Software	$500 (TensorFlow, Keras, cloud storage subscriptions)	$300 (Open-source tools, minimal licensing)
Personnel	$18,000 (3 team members × 300 hours × $10/hour, incl. overtime)	$15,000 (3 team members × 250 hours × $10/hour, less overtime)
Operational	$1,500 (electricity, internet, miscellaneous)	$1,200 (lower cloud dependency reduces costs)
Total	$22,000	$21,500
Rationale	Leverages cloud GPU rental for faster training, minimizes upfront costs, and supports a tight timeline.	Reduces long-term expenses via local GPU purchase but involves higher initial investment and setup delays.

5.3.3 Revenue Model

The revenue model is designed to recover costs and ensure sustainability for the future. Two primary strategies are proposed:

- **Subscription Service**: In the future, offer the system as a cloud-based diagnostic tool to hospitals and clinics at $500/month per institution, targeting 50 subscribers in the first year, generating $300,000 annually. This leverages the lightweight MobileNetV2 model (0.89 accuracy) for scalable deployment.
- **Licensing Fees**: License the model to medical device manufacturers for $10,000 per license, aiming for 20 licenses in the first year, yielding $200,000. This targets integration into existing imaging systems.
- **Combined Revenue**: A hybrid approach could yield $500,000 in the first year, exceeding the $22,000 budget and providing a profit margin to fund further research (e.g., dataset expansion, adversarial training).
- **Rationale**: The subscription model ensures recurring revenue and broad accessibility, while licensing targets high-value clients. The hybrid model balances short-term cash flow with long-term growth, aligning with the system's clinical relevance and sustainability goals.

5.3.4 Financial Feasibility and Risks

The primary budget of $22,000 is feasible within a six-month timeline, with a break-even point at approximately 44 subscribers or 2.2 licenses, achievable within three months post-launch based on market demand for AI diagnostics. Risks include hardware downtime (mitigated by cloud redundancy) and lower-than-expected adoption (addressed by a pilot program with three hospitals). The alternate budget's higher upfront cost poses a cash flow risk, making the primary option more practical for the current phase.

5.4 Complex Engineering Problem

5.4.1 Complex Problem Solving

This section outlines how our project differs by describing a system that uses a deep learning based system for classifying whether a medical image is real or generated by an AI, and is mapped to the complex engineering problem classifications. This mapping demonstrates the characteristics of complex problems: that the project is a challenge of applied knowledge with abstract analysis and/or ethical impacts.

Table 5.2: Mapping with Complex Engineering Problem.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓	✓	✓	✓			✓

Rationale for EP1: The project requires advanced deep learning expertise using MobileNetV2 and optimization techniques to classify medical images accurately.

Rationale for EP2: Balances model accuracy, computational efficiency, ethical data privacy and clinical reliability while optimizing GPU usage and learning parameters.

Rationale for EP3: Involves innovative analysis using multiple pre-trained models, confusion matrices, and F1-scores to address misclassifications and modality-specific weaknesses.

Rationale for EP4: Tackles rarely encountered challenges like detecting AI-generated medical images and unseen GAN-based manipulations.

Rationale for EP7: Integrates multiple interconnected processes like data preprocessing, model training, robustness testing, and ethical compliance to achieve reliable results.

Mapping with Knowledge Profile

Table 5.3: Mapping with Knowledge Profile.

K1 Natural Science	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
		✓	✓	✓	✓	✓	✓

Rationale for K3: Applies theory-based engineering principles, including data preprocessing and loss functions, to support transfer learning and fine-tuning.

Rationale for K4: Leverages advanced deep learning architectures like MobileNetV2 and custom layers, requiring expert-level AI knowledge for accurate model development.

Rationale for K5: Involves designing an efficient, scalable system architecture using frozen

layers, callbacks, and optimized training strategies for better performance.

Rationale for K6: Implements practical deployment techniques, including a Streamlit-based web app, GPU integration, and robustness testing for real-world applications.

Rationale for K7: Considers engineering's societal role by addressing ethics, data privacy, sustainability, and environmental impacts in system design and deployment.

Rationale for K8: Integrates insights from relevant research on GANs and deepfake detection to guide methodology and improve medical image classification accuracy.

5.4.2 Engineering Activities

Mapping with Complex Engineering Activities

Table 5.4: Mapping with Complex Engineering Activities.

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓			✓	✓

Rationale for EA1: Uses diverse resources like Tesla T4 GPU, TensorFlow, and a 3,000-image dataset within a $22,000 budget for efficient training and deployment.

Rationale for EA4: Enhances diagnostic accuracy while considering sustainability by addressing energy efficiency.

Rationale for EA5: Tackles uncommon challenges like AI-generated image detection through transfer learning, extending beyond conventional medical imaging practices.

5.5 Summary

Chapter 5 features an extensive technical description on the engineering specifications, systems engineering issues, social and environmental considerations, and costing for the Deep Learning-based Medical Image Classification system developed in order to differentiate between real and AI-generated medical images (X-ray, CT and MRI) [2]. It specifies conformance with major standards such as IEEE 1012-2016 (software verification), ISO/IEC 29119 (testing), IEEE 1101.1-1998 and IEC 61508 (hardware), and DICOM and IEEE 802.3

(communication), giving reasons for choosing each over possible alternatives according to project requirements. We discuss corresponding design challenges (e.g., limited datasets, model overfitting) alongside their risk mitigation strategies. We assess the system-level impact on life (better diagnostic ability with potential for misclassification), society and environment (reduction in cost, but increase in energy usage), and ethics (fairness and privacy) and develop a plan to integrate sustainability by designing energy-efficient algorithms/hardware and ultimately converting to carbon-neutral operation. It concludes with project management through Agile, a budget of $22,000 (chosen over a $21,500 alternate with local hardware), and a revenue model forecasting $500,000 per year in subscriptions and licensing for financial sustainability.

Chapter 6

Conclusion

This chapter summarizes the key findings, achievements, and limitations of the deep learning-based medical image classification system developed to differentiate real and AI-generated medical images across X-ray, CT, and MRI modalities. It also outlines potential future directions to enhance the system's performance.

6.1 Summary

The overall scenario of this paper encompasses the development and evaluation of a deep learning framework to tackle the critical challenge of distinguishing authentic medical images from AI-generated ones in modern healthcare diagnostics. The project began with data collection from a Kaggle dataset of 3,000 balanced images, like 500 per class: AI_X-ray, AI_CT Scan, AI_MRI, CT Scan, MRI, X-ray, followed by preprocessing and augmentation using TensorFlow and Keras. Five pre-trained models InceptionV3, ResNet50, DenseNet121, VGG19, and MobileNetV2 were fine-tuned using transfer learning, with training conducted on a Tesla T4 GPU over 20 epochs, achieving varying accuracies (MobileNetV2: 0.89, InceptionV3: 0.88, ResNet50: 0.58, DenseNet121: 0.80, VGG19: 0.55). The analysis included modality-specific performance, robustness testing against JPEG artifacts, unseen GANs, and adversarial perturbations, and compliance with standards like IEEE 1012-2016 and DICOM. Financial planning proposed a $22,000 budget and a $500,000 annual revenue model, while societal impacts highlighted improved diagnostics alongside environmental concerns. MobileNetV2 emerged as the optimal model, though challenges like misclassifications in AI_X-ray and CT Scan and overfitting were noted, setting the stage for future enhancements.

6.2 Limitation

Several limitations affecting the current reliability and generalisability of the medical image classification system were identified during development and evaluation. However, the dataset size of 3000 images (i.e. 500/class: AI_X-ray, AI_CT Scan, AI_MRI, CT Scan, MRI, X-ray) is quite small for deep learning applications, which indeed could give rise to overfitting as such where there is a training-validation gap (e.g. 0.9766 training accuracy vs 0.8833 validation accuracy for InceptionV3). The aforementioned limitations have caused the datasets not to represent the diversity of real-world medical imaging (such as patient

demographics, imaging devices, or pathological conditions) which may yield biased performance. Second, the system showed susceptibility to adversarial perturbations, with an accuracy drop of 20–30% under FGSM attack with ε=0.1, which can be a huge risk factor in clinical settings where adversarial attack on input data may result false negative in diagnosis [1, 5]. Thirdly, in cases such as AI_Xray and CT Scan (VGG19 had zero correct predictions on X_ray), modality-specific weaknesses were perceived in the AI (the Dataset used for classifying is balanced but it seems that not enough features are extracted from the X_ray/X_ray images). However, using a single GPU (Tesla T4) for training was fast but not scalable for bigger dataset or real-time processing increasing latency for practical implementation. Finally, the absence of longitudinal validation to actually test how the model performs as generative models for images improve over the years leaves doubts about its long-term effectiveness against other generative techniques that will likely advance over time.

6.3 Future Work

To remedy the limitations found and improve the usability of the system we propose several future work directions. First, we will widen the dataset with a minimum of 10,000 images of more diverse sources (e.g., multiple hospitals, multiple imaging equipment), which will facilitate more generalization and reduce the chances of overfitting by mid-2026. In this expansion we will consider other forms of data augmentations like rotation and noise injection to better reflect the real world variability. Second, adversarial training via PGD (Projected Gradient Descent) will increase (targeting <10% accuracy drops) robustness against perturbations with initial tests being performed in Q1 2026. Thirdly, although pre-trained models (e.g. MobileNetV2, InceptionV3) should only have their top layers unfrozen during fine-tuning, we can empirically experiment by unfreezing more layers at the cost of properly adapting to the medical dataset, so that we could achieve higher F1-scores for AI_X-ray and CT Scan (e.g. aiming for 0.85 from currently measuring 0.80) for late 2025 scheduled experiments. Fourth, deployment will be done in multiple phases integration of the system with real-time clinical workflows would require the system to be optimized for edge devices (NVIDIA Jetson Nano),inference time should be less than 1 second/image, prototyping deployment is planned for Q3 2026 with a local healthcare provider that will test the model on external patients. Finally, we will implement a framework for ongoing evaluation against such new AI-generated image techniques which we will regularly update — with quarterly updates planned from 2027. These efforts are focused on making the system a mature, scalable, and clinically-validated diagnostic solution.

References

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

[2] J. Chen, Y. Lu, Y. Yu, and J. Luo, "Synthetic medical images using F&F GAN for improved liver lesion classification," *IEEE Access*, vol. 7, pp. 112966–112977, 2019.

[3] R. A. Rana, M. B. Arif, and H. S. Al-Marzouqi, "AI-based synthetic medical image generation: Challenges and future directions," in *Proc. Int. Conf. Mach. Learn. Appl. (ICMLA)*, 2023, pp. 651–658.

[4] A. R. Joyce and A. D. Xue, "Deepfake detection in medical imaging: Threats and countermeasures," *NPJ Digit. Med.*, vol. 6, no. 1, pp. 1–10, Mar. 2023.

[5] K. K. Singh and A. Ghosh, "Medical image deepfake detection using CNN with attention mechanism," *Comput. Biol. Med.*, vol. 152, p. 106370, Jan. 2023.

[6] I. Goodfellow et al., "Generative adversarial nets," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, pp. 2672–2680, 2014.

[7] A. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint*, arXiv:1704.04861, Apr. 2017.

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4510–4520.

[9] Y. Xu et al., "Deep learning of feature representation with multiple instance learning for medical image analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 409–417.

[10] Kaggle, "Chest X-ray images (Pneumonia)," [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

[11] Kaggle, "Brain MRI images for brain tumor detection," [Online]. Available: https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection

[12] Kaggle, "CT scan (COVID-19) dataset," [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/covid19-ct-scans

[13] T. K. Satapathy and M. K. Rout, "Survey of deep learning techniques for medical image classification," *J. Med. Syst.*, vol. 44, no. 9, pp. 1–14, Sept. 2020.

[14] CSIRO, "Medical imaging synthetic data toolkit," 2023. [Online]. Available: https://www.csiro.au/en/work-with-us/funding-programs/programs/data61

[15] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint*, arXiv:1712.04621, Dec. 2017.

[16] M. Goebel et al., "Detection, attribution and localization of GAN generated images," *arXiv preprint*, arXiv:2007.10466, 2020.

[17] S. McCloskey and M. Albright, "Detecting GAN-generated imagery using color cues," *arXiv preprint*, arXiv:1812.08247, 2018.

[18] J. Lee, T. Mustafaev, and R. M. Nishikawa, "Impact of GAN artifacts for simulating mammograms on identifying mammographically occult cancer," *J. Med. Imaging*, vol. 10, no. 5, pp. 054503-1–054503-8, Oct. 2023.

[19] J. Huang et al., "BM-Net: CNN-based MobileNet-V3 and bilinear structure for breast cancer detection in whole slide images," *Bioengineering*, vol. 9, no. 6, p. 261, 2022.

[20] Y. Dai and Y. Gao, "TransMed: Transformers advance multi-modal medical image classification," *arXiv preprint*, arXiv:2103.05940, 2021.

[21] M. A. Arshed et al., "A deep learning model for detecting fake medical images to mitigate financial insurance fraud," *Computation*, vol. 12, no. 9, p. 173, 2024.

[22] Z. El Kojok et al., "Augmenting a spine CT scans dataset using VAEs, GANs, and transfer learning for improved detection of vertebral compression fractures," *Comput. Biol. Med.*, vol. 184, p. 109446, 2025.

[23] A. H. Ali and Z. Shah, "Combating COVID-19 using generative adversarial networks and artificial intelligence for medical images: Scoping review," *JMIR Med. Inform.*, vol. 10, no. 6, e37365, 2022.

[24] S. M. Saquib et al., "Medicine image classification using deep learning: Highlighting the MedNet-MoBiL hybrid model," *Discover Comput.*, vol. 28, p. 103, 2025.

[25] W. Zhang, "LIDC-IDRI Lung CT Scan Dataset," *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/datasets/zhangweiled/lidcidri

[26] M. Nickparvar, "Brain Tumor MRI Dataset," *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[27] B. Segal, "Synthetic PGGAN Chest X-rays," *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/datasets/bradsegal/synthetic-pggan-chest-xrays/data

[28] H. Rahman, "Awesome Lungs - Synthetic CT Dataset," *Kaggle*, 2022. [Online]. Available: https://www.kaggle.com/datasets/hazrat/awesomelungs

[29] L. Lee, "Large-Scale Synthetic COVID-19 CT Dataset," *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/datasets/lee123456789/largescale-synthetic-covid19-ct-data

Assessing the Reliability of AI-Generated Medical Images_ A Comparative Study with Real X-ray, CT, and MRI Scans.pdf

ORIGINALITY REPORT

16%	11%	11%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	www.mdpi.com Internet Source	1%
4	arxiv.org Internet Source	<1%
5	unora.unior.it Internet Source	<1%
6	Submitted to United International University Student Paper	<1%
7	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1%
8	S. R. Reeja, Bore Gowda, Y. S. Rammohan, Ganesan Prabu Sankar, G. Jayalatha. "Engineering Science and Technology: Innovations for the Future", CRC Press, 2025 Publication	<1%
9	hdprediction.commons.gc.cuny.edu Internet Source	<1%
10	aclanthology.org Internet Source	<1%

22% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

52 AI-generated only 17%
Likely AI-generated text from a large-language model.

22 AI-generated text that was AI-paraphrased 5%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.