Master thesis

Master of Science (M.Sc.)
Major: Software Engineering

Topic:

Comparative Analysis and Explainability of Deepfake Image Detection: XceptionNet, EfficientNet-B0, and ViT

Author: Yujin Jeon

Email: yujin.jeon.developer@gmail.com
Matriculation Number: 38933788
First supervisor: Prof. Dr. Raja Hashim Ali
Second supervisor: Prof. Dr. Sami Ur Rahman
Submitted on: 17.07.2025

Statutory Declaration:

I hereby declare that I have developed and written the enclosed Master Thesis completely by myself and have not used sources or means without declaration in the text. I clearly marked and separately listed all the literature and all the other sources which I employed when producing this academic work, either literally or in content. I am aware that the violation of this regulation will lead to failure of the thesis.

Potsdam, 17.07.2025

Yujin Jeon
.....
(Signature)

Abstract

Due to the development of deepfake technology, it has become difficult to distinguish manipulated images from reality, which is causing social problems such as personal information infringement and false information dissemination. Accordingly, the development of deepfake detection technology and improvement of accuracy have emerged as a very important research topic. In particular, the artificial intelligence based automatic detection system is of great significance in that it can effectively discriminate even advanced manipulated images that are difficult to distinguish with human eyes. The development of these reliable deepfake detection technologies is essential to protect the safety and social trust of the media environment.

In previous studies, detection performance was evaluated based on CNN based models, but direct comparative analysis between Attention based Vision Transformer (ViT) and CNN models was insufficient. In addition, studies on the application of Explainable AI (XAI) techniques to explain the basis for judgment of various models and the comparison of the results were also insufficient. Therefore, this study systematically compared and analyzed three representative models, XceptionNet, EfficientNet-B0, and ViT on the same dataset. In addition, I intend to contribute to improving the interpretability and reliability of deepfake detection models by deeply analyzing the XAI visualization results for each model.

In this study, ViT, XceptionNet, and EfficientNet-B0 models were trained using a balanced dataset, and the deepfake detection performance was compared and analyzed. Grad-CAM and Attention Rollout were used as XAI techniques to visualize the area that the model focuses on when detecting. All models were trained with the same experimental environment and hyperparameters to ensure fairness in performance comparison. In addition, the interpretability of each model was evaluated through quantitative and qualitative analysis of the XAI visualization results.

Experimental results show that ViT has the highest accuracy in certain deepfake detection tasks, and EfficientNet-B0 has excellent performance in terms of computational efficiency. As a result of XAI visualization analysis, it was found that ViT focused on subtle feature changes in the face, while XceptionNet and EfficientNet-B0 extract features across a relatively wider face area. The results of this study will contribute to the methodological development and practical applicability of deepfake detection in terms of model performance and explainability. Specifically, ViT's fine grained attention mechanism clearly demonstrates its distinction from existing CNN based models. Moreover, the high computational efficiency of EfficientNet-B0 supports practical service applicability. These findings will provide an important basis for model selection and securing interpretability in the field of deepfake detection in the future.

This paper simultaneously presents performance improvement and explainability in the field of deepfake image detection through comparison between the latest deep learning model and XAI techniques. Through systematic experiments, I provide practical guidelines for selecting appropriate models and interpretation methods for reliable deepfake detection.

Contents

List of Figures ix

List of Tables xi

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Objective . . . 3
- 1.3 Problem Statement . . . 4
- 1.4 Thesis Scope . . . 5
- 1.5 Purpose and goal . . . 6
- 1.6 Outline . . . 6

2 Theoretical Background 9
- 2.1 Background . . . 9
  - 2.1.1 Definitions . . . 9
  - 2.1.2 Deep Learning Models for Image-based Deepfake Detection . . . 10
  - 2.1.3 Explainable AI (XAI) Techniques . . . 11
  - 2.1.4 Evaluation Metrics and Computational Efficiency . . . 11
  - 2.1.5 Hyperparameter Settings . . . 12
- 2.2 Existing Methods . . . 13
  - 2.2.1 Classical Approaches . . . 13
  - 2.2.2 CNN-based Deep Learning Methods . . . 14
  - 2.2.3 Temporal Analysis and Recurrent Models . . . 14
  - 2.2.4 Transformer-based Methods . . . 15
  - 2.2.5 Fine-tuning of Pre-trained Models . . . 16
  - 2.2.6 Explainable AI (XAI) Techniques . . . 17
  - 2.2.7 Benchmark Datasets and Evaluation Protocols . . . 17
- 2.3 Evaluation Metrics . . . 18
  - 2.3.1 Classification Metrics . . . 18
  - 2.3.2 Confusion Matrix . . . 19
  - 2.3.3 Computational Efficiency Metrics . . . 19

3 Related work 21
- 3.1 XceptionNet . . . 21
- 3.2 EfficientNet-B0 . . . 22
- 3.3 Vision Transformer . . . 22
- 3.4 Gradient-weighted Class Activation Mapping (Grad-CAM) . . . 23
- 3.5 Attention-based methods . . . 23
- 3.6 Ensemble Learning Approaches . . . 24

4 Research question 29
- 4.1 Problem statement . . . 29
- 4.2 Research question . . . 30
- 4.3 Novelty of this study . . . 31

4.4 Significance of Our Work . . . 31

5 Methods **33**
5.1 Methodology . . . 33
5.2 Dataset . . . 33
5.3 Detailed Methodology . . . 34
5.4 Evaluation Metrics . . . 37
5.5 Experimental settings . . . 38

6 Results **43**
6.1 Model Performance . . . 43
6.1.1 XceptionNet . . . 43
6.1.2 EfficientNet-B0 . . . 44
6.1.3 ViT . . . 45
6.1.4 Model Comparison . . . 46
6.2 Model Complexity . . . 48
6.3 eXplainable AI (XAI) . . . 49

7 Conclusion **53**
7.1 Summary . . . 53
7.1.1 Limitations . . . 54
7.2 Conclusion . . . 55
7.3 Dissemination . . . 55
7.4 Problems Encountered . . . 55
7.5 Outlook . . . 56

Bibliography **61**

List of Figures

5.1 This figure illustrates the overall workflow of the deepfake image detection experiment, from dataset preprocessing and model training (ViT, XceptionNet, EfficientNet-B0) to performance evaluation and XAI-based visual explanation (Grad-CAM and Attention Rollout). 34

5.2 This figure shows sample images from the Kaggle dataset used in this study. The first row (real images) has been blurred to protect personal information, while the second row shows fake images with synthetic distortions that models learn to detect. The dataset includes diverse faces, lighting, and backgrounds to enhance generalization. (Dataset source: `https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images`) . 35

5.3 This figure shows the entire segmentation structure of the Kaggle image dataset used in this study. The data were divided into a ratio of 80% for training, 10% for validation, and 10% for testing. Each segmentation set is configured to maintain balance between classes and provides a basis for quantitatively measuring the performance evaluation and generalization ability of the model. 35

5.4 This figure presents the structure of the XceptionNet architecture, illustrating how depthwise separable convolution is applied throughout the entire network, from the Entry Flow to the Middle Flow and Exit Flow stages. Each stage leverages depthwise separable convolution to efficiently extract hierarchical features while significantly reducing the number of parameters and computational cost, compared to standard convolutions. 39

5.5 This figure shows the EfficientNet-B0 architecture, from the Stem and MB-Conv blocks (with depthwise separable convolution, SE modules, and skip connections) to the Head, pooling, and classifier, illustrating how it balances depth, width, and resolution efficiently. 40

5.6 This figure shows the Vision Transformer (ViT) architecture, illustrating the flow from image patching and embedding with positional encoding, through transformer encoder blocks, to final classification. 41

6.1 Training and validation metrics (Accuracy, Precision, Recall, F1-score, AUC, and Loss) per epoch for the XceptionNet model. 43

6.2 Confusion matrix for XceptionNet on the test set, showing the distribution of true and predicted labels for Fake and Real classes. 44

6.3 Receiver Operating Characteristic (ROC) curve and the area under the curve (AUC) for XceptionNet, indicating the model's overall discrimination ability. 44

6.4 Training and validation metrics (Accuracy, Precision, Recall, F1-score, AUC, and Loss) per epoch for the EfficientNet-B0 model. 45

6.5 Confusion matrix for EfficientNet-B0 on the test set, showing the distribution of true and predicted labels for Fake and Real classes. 46

6.6 Receiver Operating Characteristic (ROC) curve and the area under the curve (AUC) for EfficientNet-B0, indicating the model's overall discrimination ability. 46

6.7 Training and validation metrics (Accuracy, Precision, Recall, F1-score, AUC, and Loss) per epoch for the ViT model. 47
6.8 Confusion matrix for ViT on the test set, showing the distribution of true and predicted labels for Fake and Real classes. 47
6.9 Receiver Operating Characteristic (ROC) curve and the area under the curve (AUC) for ViT, indicating the model's overall discrimination ability. 48
6.10 Bar graph comparing the deepfake image detection performance of three models (ViT, XceptionNet, and EfficientNet-B0). ViT shows the highest performance, followed by XceptionNet, with EfficientNet-B0 recording the lowest values among the three. 49
6.11 This figure is Grad-CAM visualization result of XceptionNet model. The left is a deepfake image and the right is a result of a real image. Closer to red in each image means an area that the model considers more important for classification, and closer to blue indicates an area of relatively low importance. 50
6.12 This figure shows Grad-CAM visualization results of the EfficientNet-B0 model. The left side shows the deepfake image, and the right side shows the area of attention for the real image. 51
6.13 The visualization result of the Attention Rollout of the Vision Transformer (ViT) model. The left side represents the deepfake image, and the right side represents the area of interest for the real image. The brighter the area, the greater the contribution of the model to classification, and the darker the area, the lower the importance. 51

List of Tables

2.1 This table shows the structure of a confusion matrix, which is used to evaluate the performance of classification models. The confusion matrix provides an intuitive overview of how many samples were correctly or incorrectly classified for each class by crossing the actual and predicted classes. 20

3.1 This table is a literature review table that compares and organizes the datasets, methods used, experimental results, major contributions, and limitations of major papers published in Deepfake detection and eXplainable AI. 24

5.1 This table shows the split distribution of the dataset used in this study. It was divided into training (80%), validation (10%), and test (10%) 34

5.2 This table shows the results of comparing the performance of the three models (XceptionNet, EfficientNet-B0, and ViT) for Deepfake image detection based on Accuracy, Precision, Recall, and F1-score. ViT recorded the best performance in all indicators. All metrics are reported in abbreviated form: Acc. (Accuracy), Prec. (Precision), Rec. (Recall), F1 (F1-score), with (W) indicating weighted average and (M) indicating macro average. 38

5.3 This table summarizes the network training environments and key hyperparameters used in this study, ensuring reproducibility and fair comparison across models. 39

6.1 Comparison results of XceptionNet, EfficientNet-B0, ViT's number of parameters (in millions), FLOPs (in billion), and total learning time (in minutes). The computational efficiency and resource consumption characteristics of each model are shown at a glance. 49

1 Introduction

The sophistication of deepfake images is increasing day by day due to advances in deep learning and generative artificial intelligence technologies [BCDZ25]. As a result, it becomes difficult to distinguish between actual images and manipulated images, and various social problems such as personal information infringement, identity theft, and the spread of false information are occurring. [Kha24] In particular, the dissemination of deepfake images through SNS or media can increase social confusion and undermine public trust. [Boe25] Against this background, the need for technology development capable of effectively detecting advanced deepfake images has emerged significantly. This study uses the latest deep learning based models such as ViT, XceptionNet, and EfficientNet-B0 to systematically compare and analyze the detection performance for various deepfake images. All models were trained and evaluated in the same experimental environment based on a balanced dataset, increasing the fairness and reliability of the results. In addition to the detection performance of each model, practical applicability and efficiency were also considered. In addition, by analyzing the structural differences and performance characteristics between the latest trends, Vision Transformer and CNN based models, I intend to provide practical insights to the field and research field [SSTE+24]. The core purpose of this study is to simultaneously explore technological advances and practical applications in the field of deepfake image detection. This research approach is expected to contribute to building a safe and reliable digital environment.

In addition, the interpretability of how each model discriminates deepfake images is also a major research topic, not just comparing the performance of the model. Existing studies have mainly focused on performance evaluation based on CNN based models, but they have not sufficiently dealt with the application of XAI techniques and comparison of results that explain the basis for judgment of various models [BYDA25]. Therefore, in this study, representative XAI techniques such as Grad-CAM and Attention Rollout were used to visualize the image area that each model pays attention to. Through this, I systematically analyzed what features ViT, XceptionNet, and EfficientNet-B0 focus on when detecting deepfakes. This XAI based analysis plays an important role in improving the reliability of deepfake detection models and securing transparency in actual introduction [GT24]. Furthermore, by incorporating interpretable AI technologies, both experts and general users can easily understand the model's judgment process. This paper aims to provide practical guidelines for establishing deepfake response strategies in the future by integrating the two aspects of performance and interpretability. The approach of this study, which combines deep learning and XAI techniques, has the potential to present new standards in the field of deepfake detection. Finally, the results of this study will have practical application value in related research and industrial sites in the future and contribute to enhancing the reliability of society as a whole. As such, this paper aims to simultaneously develop theoretical and practical developments in the field of deepfake image detection.

1.1 Motivation

Deepfake technology is having a great influence on each field of society along with the rapid development of artificial intelligence [HB21]. In particular, deepfakes that manipulate im-

age and video data are highly likely to be exploited in various fields such as politics, entertainment, and finance [GYV+24]. This is causing serious social problems such as fake news, false evidence, and identity theft [SD23]. Recently, images generated by deepfakes have become so sophisticated that they are hardly distinguished from real person photographs, and have reached levels that are impossible to identify with the human eye [BRS+24]. As a result, existing manual based detection methods are facing limitations, and the need for automated artificial intelligence based detection techniques is rapidly increasing. In particular, reliable deepfake detection technology can meet various industrial demands such as information security, digital forensics, and media verification [ABB+25]. In order to protect social safety and public trust, the development of advanced detection technologies has become a necessity, no longer an option [Tiw25]. Existing simple image filtering or metadata verification methods are difficult to apply to real deepfake image detection due to technical limitations [LAP+11]. Accordingly, deepfake image detection research using the latest deep learning models has become a very important research topic [HJNDU24]. This study started from a practical motive to solve these social backgrounds and technical challenges.

One of the recent research directions in the field of deepfake detection is to systematically compare and analyze the effects and limitations of various model structures beyond simply high detection accuracy [TC22]. CNN based models have shown excellent performance for a long time in the field of image classification, but there are limitations in capturing fine manipulation features such as deepfakes [NNN+22]. On the other hand, attention based models such as Vision Transformer (ViT) are emerging as new alternatives as they can effectively utilize the global features of images [KRS+23]. However, most of the preceding studies so far deal with CNN and ViT models separately or lack direct comparative studies [SW22]. In addition, the analysis of how the structural differences of the model affect the actual deepfake detection performance and efficiency has not been sufficiently conducted. Accordingly, research to compare and evaluate the latest models of various architectures under the same conditions is urgently required. Not only accuracy, but also computational efficiency and applicability in real world environments should be considered extensively. This systematic comparative analysis suggests the direction of development of deepfake detection technology and can provide practical selection criteria for field application. This study selects representative deep learning models such as ViT, XceptionNet, and EfficientNet-B0 to comprehensively analyze detection performance and efficiency in a fair experimental environment. Through this, I intend to contribute to overcoming the limitations of existing research and suggesting new research directions.

Due to the 'black box' nature of deep learning models, securing reliability and interpretability of deepfake detection results has recently emerged as an important research issue [Lu24]. In particular, the function of explaining on what basis the model judges deepfakes is essential, not just providing detection results [HSC22]. Explainable AI (XAI) techniques help to visualize or quantitatively analyze the judgment basis of deep learning models [MPK20]. However, deepfake detection studies so far have been limited in XAI applications, and systematic comparative studies of the interpretability of different models are rare [AJP22]. Research is needed to visualize and analyze differences by applying XAI techniques such as Grad-CAM and Attention Rollout for each model. This allows us to understand in detail which areas of the image the model pays attention to and what features it discriminates deepfakes based on. This interpretability can increase the reliability of the model and improve the receptivity of both professionals and general users [WHHJ22]. In particular, in high trust environments such as medical, legal, and media verification, XAI based explanatory features are important grounds for decision making [Lim21]. Therefore, this study aims to present a practical and transparent deepfake detection method by

visualizing and comparing the judgment basis of each model beyond simple performance comparison. This approach could become a standard research direction in the field of deepfake detection in the future.

Finally, this study aims at practical and comprehensive contributions considering both performance and interpretability in the field of deepfake image detection. While existing studies mainly focused on single model performance, this paper integrated the detection performance, computational efficiency, and interpretability of the latest deep learning models in one framework. To this end, I applied the same experimental conditions as the balanced dataset, resulting in practically applicable results in the field. In addition, by quantitatively and qualitatively evaluating the results of XAI visualization for each model, the basis for securing both reliability and transparency in actual use was laid. Such research design has great industrial and social utilization value as well as academic contribution [WMMD24]. Deepfake detection technology is expected to become a key infrastructure in various areas such as information security, media verification, and digital forensics in the future [KLI+25]. The results of this study can provide practical guidelines for all field workers, policy makers, and researchers. In particular, explainable AI based deepfake detection models will greatly contribute to enhancing the reliability of society as a whole and creating a safe digital environment [KPL22]. As such, this study presents a new paradigm of deepfake response strategies through the fusion of advanced deep learning technology and XAI. Through this, I intend to help effectively respond to social and technical challenges caused by deepfakes.

1.2 Objective

The main goal of this study is to systematically compare and analyze the performance and interpretability of the latest deep learning models in the field of deepfake image detection at the same time. Specifically, three representative models of Vision Transformer (ViT), XceptionNet, and EfficientNet-B0 are trained and evaluated in the same experimental environment for fair and reliable performance comparisons. By using a balanced dataset, it is designed to derive objective results that are not biased toward a specific class. Furthermore, it does not stop at simple accuracy evaluation, but uses explainable artificial intelligence (XAI) techniques such as Grad-CAM and Attention Rollout to visually analyze the basis for discrimination of each model. This approach allows us to specifically identify which image regions the model detects deepfakes by paying attention to. By analyzing both performance (accuracy, efficiency) and interpretability (explainability), the core purpose is to present the optimal model selection criteria applicable to the real world environment. Furthermore, this study aims to lay the groundwork for building a practically reliable and explainable detection system, not just to reveal which model is better. In this process, all experimental conditions were unified to ensure maximum fairness and reproducibility of the experimental results. This goal of the study is focused on providing practical guidelines that can be used in actual services or industrial sites, as well as academic contributions. Ultimately, this study intends to contribute to the development of deepfake image detection models with both reliability and transparency.

In addition, this paper aims to present specific and actionable guidelines necessary for the future development and practical application of deepfake detection techniques. By analyzing the strengths and limitations of each model in various situations from various angles, I derive practical trade-offs between accuracy, computational efficiency, and interpretability. In particular, I examine in depth how each model responds to microscopic and complex manipulations that are common in modern deepfake images. Through XAI based visualization analysis, I also provide insights on how to increase the interpretability of the

model and how it can be applied to real world expert judgment. In addition, this study aims to overcome the current research limitations and lay the groundwork for introducing reliable explainable deep learning models in sensitive fields such as security. For this goal, I present comprehensive research results by integrating not only quantitative results but also qualitative analysis. It provides practical benefits to policy makers, industry, and researchers alike by proposing realistic standards and methodologies that can be applied directly in the field. In particular, the ultimate goal is to promote the dissemination of AI based detection technologies where reliability and transparency are essential in various applications. This study aims to present future oriented research directions in the field of deepfake detection through the fusion of deep learning and XAI. Through this, it is expected to contribute to the creation of a safer and more reliable digital environment.

1.3 Problem Statement

This study aims to present the optimal model selection criteria by quantitatively comparing the deepfake detection performance of the three representative deep learning models (ViT, XceptionNet, and EfficientNet-B0) and analyzing the visual explanation power of each model. In particular, the performance difference between models is evaluated fairly using a balanced image dataset, and visual grounds and judgment structures are derived through model with specific XAI techniques. To this end, the Attention Rollout technique is applied to ViT and Grad-CAM is applied to XceptionNet and EfficientNet-B0 to visualize which features each model pays attention to. Most of the existing studies focused on only a single model, so the comparison of explainability and computational efficiency according to structural differences was insufficient. Therefore, this study aims to promote the overall understanding of the deepfake detection model through comparative analysis between multiple architectures. In addition, the reliability of the visual explanation is verified by evaluating whether the model is focused and the actual manipulated image area. In this process, the consistency between the prediction results and visualization results of each model is also analyzed. In addition, the evaluation is performed from a practical point of view by considering not only the accuracy of the model, but also the calculation time and resource efficiency. Ultimately, this study aims to provide basic data for the development of a deepfake detection framework suitable for real-world applications by integrating performance, efficiency, and explainability. This is meaningful in presenting a practical and applicable guide in the field of deep learning based image forensics.

In particular, this study aims to point out the practical limitations and necessity of deepfake detection research by comparing and analyzing deep learning models of various structures under the same conditions, away from the existing single model-centered performance evaluation. Existing studies have tended to emphasize only the accuracy of the model, but various factors such as computational efficiency and interpretability must be considered comprehensively in actual application environments. Therefore, in this paper, along with the detection performance of each model, I systematically evaluated how much XAI based visual explanation power matches the actual operation area and its reliability. In addition, not just numerical comparisons, but also detailed interpretations of what features each model focuses on and detects deepfakes were conducted in parallel. Such an approach can select the optimal model suitable for field application and provide a practical criterion for establishing deepfake response strategies in the future. Furthermore, by incorporating quantitative evaluation and qualitative analysis, I have improved our understanding of the strengths and limitations that the model can show in various real world environments. Through this, I intend to present guidelines that are practically helpful not only to researchers, but also to industry, policymakers, and practitioners. In particular, it

is expected that this study will be able to contribute to fields that require securing reliable interpretability and efficient use of resources. Ultimately, this paper aims to lay the groundwork for building a new deepfake detection framework that encompasses all three axes: performance, efficiency, and explanatory power. This aims to achieve both academic development and practical social contributions in the field of deep learning-based video forensics at the same time.

1.4 Thesis Scope

The scope of this study is limited to deepfake image detection using deep learning based models. Specifically, three representative structures, Vision Transformer (ViT), XceptionNet, and EfficientNet-B0, are selected for experimentation and analysis. All experiments use the same deepfake image dataset, which is balanced so that there is no class imbalance. The learning, validation, and testing processes were all performed under the same hyperparameters and environmental settings, and the fairness of performance comparison between models was ensured as much as possible. The performance evaluation indicators covered in this paper were selected based on accuracy, computational efficiency (FLOPs, number of parameters, execution time), and practical applicability. In addition, representative XAI techniques such as Grad-CAM and Attention Rollout were applied to evaluate the interpretability of each model. The scope of the experiment is limited to image based deepfake detection, and other modalities such as video based deepfakes and voice and text were excluded from the research scope. Rather than applying additional techniques such as model structure design or data augmentation, the focus was on the comparison and interpretation of existing widely used representative models. Experimental data were carefully selected from the published benchmark dataset, and the preprocessing and postprocessing processes were also controlled consistently. Through this setting, this study aims to clarify the differences in detection performance and explanatory power according to structural differences in deep learning models.

In addition, this paper also focused on systematically analyzing the visual explanatory power of each model. The results of applying the XAI technique for each model were evaluated both quantitatively and qualitatively (example visualization, interpretation). It includes an analysis of how much the visualization results match the actual deepfake manipulation area and whether the judgment basis is logically valid. The practical reliability of interpretable artificial intelligence was verified by evaluating the consistency between the prediction results and visualization results of each model within the experimental range. In this study, not only performance but also computational efficiency and explanatory power were considered comprehensively in consideration of the applicability in the actual field. However, due to the nature of the research scope, more complex multimodal data, vedio sequences, and real time detection were left as follow-up studies. Furthermore, special purpose datasets (medical, legal, etc.) or custom model design are not directly analyzed in this study. Consequently, this paper focuses on an integrated analysis of "structural characteristics, detection performance, and visual explanatory power of representative models" in the field of deepfake image detection. Through this clear scope of research, I would like to present practical grounds and criteria for both field application and subsequent research. Finally, I reveal that all results and implications derived within the scope of this study should be interpreted only in the field of image based deepfake detection.

1.5 Purpose and goal

The purpose of this study is to analyze the performance and interpretability of image based deepfake detection simultaneously in depth using the latest deep learning models. Recently, as deepfake technology has become increasingly sophisticated, it is becoming increasingly difficult to distinguish false images using conventional simple detection methods. Accordingly, the primary goal was to clarify the strengths and limitations of each model by fairly comparing three representative models, ViT, XceptionNet, and EfficientNet-B0, under the same conditions. In particular, not only simple performance figures of the model but also important factors in the actual service environment such as computational efficiency and learning speed were considered. The dataset is designed to be balanced for each class so that objective evaluation that is not biased toward a specific class is possible. The experimental results of each model were derived under a unified hyperparameter and experimental environment to ensure reproducibility and fairness. Through this process, I present the most suitable model for deepfake image detection and increase its applicability in various environments. In addition, it is also an important purpose to provide practical selection criteria for both field practitioners and researchers through the analysis of the judgment basis and characteristics of each model. This study aims to present practical criteria that can contribute to solving actual social and industrial problems beyond simple performance comparisons. Ultimately, the core goal of this paper is to promote comprehensive development that encompasses the reliability, efficiency, and explanatory power of artificial intelligence based deepfake detection.

In addition, this study intends to practically evaluate the interpretability of each model by incorporating explainable artificial intelligence (XAI) techniques. While existing studies mainly focused on detection accuracy, this paper applied visualization techniques such as Grad-CAM and Attention Rollout to analyze the image area that the model actually pays attention to. Through this, I want to understand in detail what features and areas each model judges deepfakes based on. The XAI results for each model are evaluated from various perspectives, including the degree of alignment with the actual manipulated image regions, the clarity of interpretation, and practical applicability. This analysis can serve as a basis for enhancing the reliability of artificial intelligence models for both field experts and general users. In addition, based on XAI based analysis, the limitations of the model and the direction of improvement are presented, and it is intended to contribute to the development of deepfake response technology in the future. By simultaneously addressing the two axes of performance and interpretability, this paper aims to present realistic and practical research results applicable in real world fields. The various insights and suggestions derived from this process can be widely used not only for follow-up research but also for policy establishment and industrial application. Therefore, this study aims to contribute to overcoming technical and practical limitations in the field of deepfake image detection and driving future-oriented development. Ultimately, the final goal of this study is to lay a practical foundation for creating a reliable and transparent artificial intelligence based media environment.

1.6 Outline

This chapter describes the development background and social needs of deepfake image detection technology. With the development of deep learning, it emphasizes that deepfakes are emerging as a problem in real society. The purpose of the study, the main research questions, and the overall scope of the study in this paper are clearly presented. It briefly introduces the differences from previous studies and the practical contribution and signif-

icance of this study. Finally, I guide the entire structure of the paper and the flow of each chapter.

2 Theoretical Background

This chapter provides the necessary background knowledge to understand the key technical concepts applied in this paper. In order to accurately grasp the subject matter, experimental design, and interpretation of the results of the paper, a sufficient understanding of basic terms and core principles must first be preceded. In particular, deep learning and generative artificial intelligence technologies are developing very rapidly, so the latest research flows and terminology definitions are essential. This chapter introduces the theoretical background of various fields, including deepfakes, explainable artificial intelligence (XAI), and image classification models, step by step. I will explain the principles and development process of each concept, and I will also cover how it is applied in real world research and service environments. In addition, the structural features and differences of the model used in this paper, as well as the meaning of the performance indicators, are described in detail. These theoretical background explanations play an important role in increasing the reliability of the experimental design and interpretation of the results of the paper. In addition, we highlight the originality and necessity of this paper by briefly summarizing existing research trends and limitations in the relevant fields. This chapter will make it easier for the reader to understand the main experimental methodologies and interpretations of the paper. In the end, this chapter serves as an important starting point on which to understand the flow of the entire paper.

2.1 Background

2.1.1 Definitions

Deepfake refers to a technology that uses artificial intelligence based generative models (mainly GAN, VAE, etc.) to create manipulated images that are difficult to distinguish from reality [AS21]. Deepfake detection is a problem of classifying whether such manipulated images are real or fake, and has recently emerged as a very important research field in the digital media environment [KKK24]. Explainable AI (XAI) refers to a technology that visualizes or numerically interprets the judgment basis of complex deep learning models so that humans can understand them [HJMK24]. Model Interpretability refers to the degree to which an artificial intelligence model can clearly explain why it produces a specific result and is essential for practical reliable and responsible AI utilization [MWL+22]. Besides this, this paper uses various deep learning and XAI terms such as CNN, Transformer, Grad-CAM, and Attention Rollout, so an accurate definition of each concept must be preceded. Recently, deepfake technology has been recognized as a real threat in various fields such as social media, politics, and entertainment [IHB+24]. In particular, it is spreading to social problems such as copyright infringement, dissemination of false information, and invasion of privacy [AM23]. Accordingly, both deepfake generation and detection technologies are rapidly developing, and technology competition is also intensifying [Zha22]. The importance of explainable artificial intelligence is emphasized not only in legal and ethical aspects but also in securing real user trust [AKK23]. Based on this social and technical context, this study actively introduced XAI to overcome the limitations of existing deepfake detection research.

Terms related to deep learning and XAI are essential to understand the core concepts of this study. Convolutional Neural Network (CNN) is an artificial neural network structure that automatically extracts and analyzes spatial features within an image, and is widely used in computer vision [ZWZ+24]. Transformer was originally developed in natural language processing, but recently, such as Vision Transformer (ViT), it is used to effectively model the relationship between entire patches by receiving images as input [HGL+23]. Generative Adversarial Network (GAN) is a structure in which two neural networks compete with each other to learn, and can generate very realistic fake images [Dur21]. Variational Autoencoder (VAE) is used to model the latent space of data and generate new images using probabilistic encoding methods [LSM21]. Gradient-weighted class activation mapping (GRA-CAM) is an XAI method that visually highlights the areas noted by CNN-affiliated networks when classifying images [ZO23]. Attention Rollout is an interpretation technique that visualizes the importance of each input patch in a Transformer based model and intuitively shows the model's prediction basis [FN24]. These technical terms are repeatedly mentioned in the experimental design and result interpretation process of this paper. Therefore, a clear understanding of the definition and principles of each term is very important for correctly interpreting and evaluating the content of this study. In this paper, based on the key terms defined above, Deepfake detection and XAI analysis experiments were systematically performed.

2.1.2 Deep Learning Models for Image-based Deepfake Detection

Convolutional neural networks (CNNs) have been widely used mainly in the field of deepfake image detection [AMG21]. CNN has the advantage of being able to effectively extract local features within an image, but there are limitations to understanding global relationships [AZH+21]. To overcome these limitations, models using the self attention mechanism, such as Vision Transformer (ViT), have recently emerged [JWZ24]. ViT is advantageous for complex manipulation feature detection as it can learn the relationship between each patch globally after segmenting the entire image in patch units [YZJ+22]. In addition, lightweight and high performance CNN structures such as XceptionNet and EfficientNet-B0 also show excellent performance on various benchmarks [HL22]. Understanding the structural characteristics and advantages and disadvantages of each model is a starting point for experimental design and interpretation of results. The accuracy of deepfake detection models is highly dependent not only on model architecture, but also on the quality and diversity of training data. Recent studies have emphasized the importance of data augmentation techniques to enhance the robustness of models for various manipulations [KBMB24]. In addition, pretraining on large datasets helps the model to acquire more generalized features, which is particularly advantageous in deepfake detection tasks [KDN23]. As the scope of image manipulation techniques continues to expand [LWD+23], continuous updates to both datasets and model architectures are needed to maintain detection performance.

Recently, hybrid approaches to maximize performance by combining models of different structures have also been actively studied [JZDL24]. For example, a structure is proposed that simultaneously leverages the robust local pattern recognition capabilities of the CNN family and the global context understanding capabilities of Transformer based models [YZF23]. These hybrid models complement the limitations of a single structure and can significantly improve the real Deepfake image detection performance. In addition, as data diversity and complexity increase, various state-of-the-art techniques such as adaptive learning, transfer learning, and ensemble are also being introduced [MAKM22]. Lightweight networks that optimize model parameters and computational volumes, such as EfficientNet-B0, are in the spotlight in terms of real-time processing and mobile environment application [LZG+25]. XceptionNet is a representative example of simultaneously

securing computational efficiency and classification performance using deepwise separate convolution [GSA+24]. Transformer structures such as ViTs require large datasets and high-performance hardware, but have strength in capturing complex operational characteristics [HJC+23]. The choice of model structure should comprehensively consider various factors such as data characteristics, experimental purpose, and usage environment. The emergence and development of such diverse model architectures is overcoming the limitations of Deepfake detection technology and increasing the applicability of actual services. In this paper, the advantages and disadvantages of major deep learning-based structures were compared, and the characteristics and performance of each model were systematically analyzed through actual Deepfake detection experiments.

2.1.3 Explainable AI (XAI) Techniques

XAI techniques are being actively studied to solve the "black box" problem [VE21] of complex deep learning models [ŞAO25]. Gradient weighted class activation mapping (Grad-CAM) is a representative method for visually highlighting critical areas in input images of CNN based models [HS23]. In the case of Transformer based models, methods such as Attention Rollout can be used to visualize which image parts the self attention mechanism focuses on [SCMS23]. These XAI techniques make the model's judgment base interpretable and increase reliability and transparency in practical application [Chi25]. In this study, the manipulated image detection process and concentration area of each model are analyzed in depth through the XAI technique.

Recently, various interpretation techniques have been developed in the field of XAI, and the reliability and availability of the model's decision process are increasing significantly [SYK+22]. In particular, Grad-CAM is not just a simple visualization, but is widely used to check whether the judgment basis of the expert and the judgment basis of the model match in actual field applications such as medical image analysis [EM25]. In the case of Transformer models, not only Attention Rollout but also several analysis techniques such as Layer-wise Relevance Propagation (LRP) and Integrated Gradients are being studied together [GDS+24]. These methods enable quantitative identification of not only which area in the image the model was focused on, but also the interactions between input features. The visual rationale derived through the XAI technique helps to detect weaknesses or data errors in the model early on, which are unknown by simple accuracy figures. In addition, it increases utilization in real world services by providing reliable explanations to both users and developers. As the complexity of deep learning models increases, the role of XAI becomes more important, especially in areas where social reliability is required, such as Deepfake detection. Based on the experimental results, I can see how XAI visualization actually reveals the misjudgment area or misclassification pattern of the model. As such, XAI technology is not just an auxiliary tool, but a key factor in ensuring the safety, reliability, and legal/ethical responsibility of the model. Therefore, in this study, the interpretability and practical utility of the Deepfake detection model were closely verified through the application and comparison of various XAI techniques.

2.1.4 Evaluation Metrics and Computational Efficiency

The performance evaluation of the deepfake detection model is performed using not only simple accuracy, but also various classification indicators such as Precision, Recall, F1-score, and AUC. In addition, operational efficiency indicators such as the number of parameters of the model, the amount of computation (FLOPs), and training time are also considered for practical environmental application. It is necessary to understand that the balance of model performance, efficiency, and interpretability is a key factor in de-

termining its availability in the actual service field. Accurate performance evaluation of deepfake detection models should be done from various perspectives. In addition to simple classification accuracy, considering the actual service environment, it is necessary to balance the efficiency and interpretability of the model. In particular, the complexity of the model, the number of parameters, and the amount of computation (FLOPs) act as very important evaluation criteria in hardware-constrained environments [QCNS22]. In systems or mobile environments that require real-time processing, the speed of inference and the weight reduction of the model are key factors [SICM22]. In addition, the quality or diversity of datasets and the application of prior and transfer learning also have a significant impact on actual performance. In addition, the ease of update, scalability, and maintenance perspective of the model must be considered together to increase practical applicability. Therefore, in this paper, the criteria for comprehensively evaluating practical environmental applicability along with various quantitative performance indicators were applied.

In real world applications, it is difficult to judge the superiority of the model simply with high accuracy [Mya24]. The complexity, computational efficiency, and interpretability of the model are all considered for field application. In particular, in environments with limited computational resources or services requiring real time processing, the weight reduction and inference speed of the model are very important evaluation criteria. FLOPs or models with fewer parameters can save hardware resources and have advantages in terms of maintenance and deployment. On the other hand, oversimplified models may suffer from poor generalization performance or complex manipulation detection capabilities, which requires an appropriate level of performance and efficiency balance. In addition to the learning speed and inference time in actual operation, the ease of model update and management, scalability, etc. are also important factors. And it should be possible to transparently explain the basis for judgment of the model through interpretable artificial intelligence techniques such as XAI. In order for users or experts to be able to trust the results, it is necessary to provide comprehensive interpretation indicators along with performance figures. Therefore, in this paper, various quantitative performance indicators and comprehensive evaluation criteria including computational efficiency and interpretability were applied. Through this, I tried to specifically present the conditions of the Deepfake detection model that can be used in the actual service environment.

2.1.5 Hyperparameter Settings

The performance and generalization capabilities of deep learning models depend heavily on hyperparameter settings. Hyperparameters are values that the researcher directly determines before the model's training process begins. Representatively, these include the learning rate, batch size, number of epochs, optimization algorithms, and loss functions. The learning rate is an important factor in determining how quickly the model converges to the minimum value of the loss function. The batch size refers to the number of data samples input to the network at one time, and affects computational efficiency and learning stability. The number of epochs determines how many times to learn the entire dataset repeatedly, which poses a risk of underfitting if insufficient and overfitting if excessive. The optimization algorithm presents a method for determining how to update the parameters to reduce the value of the loss function. The loss function guides the direction in which the model learns by quantifying the difference between predicted and real values. These hyperparameters are closely connected to each other and can have a complex impact on the learning outcomes of the model, either individually or in combination. Effective hyperparameter setting and tuning is therefore an essential process for the development of deep learning models.

Hyperparameters are more than just set values. Appropriate hyperparameter adjustments depend not only on the learning efficiency of the model, but also on its final performance and its applicability in real world environments. Incorrect hyperparameter settings can lead to instability in training, degradation in performance, and even complete failure of the model. In deep learning research, hyperparameter exploration is often done through repeated experiments and evaluation processes. Hyperparameter tuning becomes more important, especially when applying complex datasets or new model structures. In some recent studies, automated hyperparameter optimization techniques are also utilized. Suitable hyperparameter selection increases the generalization ability of the model and effectively controls overfitting or underfitting problems. In a real service environment, hardware constraints, learning time, and maintenance aspects must also be considered, so hyperparameter settings should be more strategic. After all, hyperparameters are a key factor that exists at the intersection of maximizing model performance and securing practical applicability. Theoretically, a systematic understanding of hyperparameters can also be said to be the starting point for reliable artificial intelligence research.

2.2 Existing Methods

2.2.1 Classical Approaches

First, in the field of deepfake detection, traditional image processing and feature based approaches were mainly used in the early days. These methods have mainly focused on extracting hand designed features such as texture, borderline, color distribution, etc. from face regions. For example, local binary patterns or histogram based features have been utilized a lot [WKKG21]. The extracted features were fed into traditional machine learning algorithms such as SVM, KNN, and decision tree for classification [NM16]. This approach had the advantage of not requiring large-scale training data and being relatively simple to implement. However, relying only on human designed features has made it less adaptable to new manipulation techniques or complex variations. In particular, there have been limitations in capturing the fine grained modulation of deep learning based manipulation techniques that have recently emerged. These approaches have experienced a sharp decline in performance when data is lacking in diversity or when traces of manipulation are microscopic. In addition, the false detection rate could be increased by various factors such as facial outer area and background. This has led to the need for new methods with more automated feature extraction and high adaptability.

Moreover, classical approaches are still meaningful in that they were easier to interpret and implement compared to complex deep learning models. In practice, it is also used as a basic filtering tool in combination with a simple preprocessing process. However, the fixed feature extraction method had fundamental limitations in responding to the developing manipulation techniques. In particular, it was difficult to reflect variables in real world environments such as facial expressions, pose changes, and various lighting conditions. This has led researchers to gradually pivot to deep learning approaches that enable data driven automatic learning. Early research has also introduced hybrid models that combine traditional methods with deep learning. These hybrid methods sought to complement their performance by adding representations obtained from deep learning to existing features. However, as the performance of deep learning based models gradually increases, the proportion of existing techniques is decreasing. Nevertheless, in some applications, traditional methods are still used in pre-filtering or pre-processing phases. Eventually, advances in deepfake detection techniques have led to the emergence of new frameworks that can overcome the limitations of existing methodologies and respond to more complex variations.

2.2.2 CNN-based Deep Learning Methods

With the development of deep learning, CNN based image analysis techniques have become central to the field of deepfake detection. Convolutional Neural Networks (CNNs) have the advantage of being able to automatically extract local features within images and learn hierarchically complex patterns. Representatively, structures such as XceptionNet, EfficientNet, and ResNet have been widely applied in deepfake image detection. These models can effectively capture even microscopic traces of manipulation through deep and complex network structures. Specifically, XceptionNet is evaluated for simultaneous computational efficiency and performance with Depthwise Separate Convolution. EfficientNet is designed to optimize model size and computation at the same time to maintain high accuracy in limited resource environments. CNN-based models have shown superior performance improvements over traditional techniques when large scale training data is available. It also has the advantage of mitigating the overfitting problem through various tricks such as data augmentation and normalization. However, CNN models are limited by their inherent dependence on local features, making it difficult to fully reflect global contextual information. As a result, there has been a discussion about the need for new structures that can learn a wider range of relationships and patterns.

CNN based methods have established themselves as standard approaches, performing well on real-world contests and benchmark tests. In particular, it showed robust characteristics even under various noise conditions or poor image quality. However, with increasingly advanced manipulation techniques, simple CNNs alone have become difficult to fully capture even fine variations of the latest deepfake techniques. Due to this, researchers have attempted to build deeper CNN structures or make various structural improvements such as Residual/Attention blocks. Furthermore, hybrid approaches that combine CNN and RNN or Transformer families of networks have also been actively studied. Recently, more and more attempts have been made to apply lightweight CNN models to mobile environments and real time systems. In practice, EfficientNet-B0 and others show sufficiently high detection accuracy despite low parameters and low computational power. However, the diversity of datasets and generalization to real-world data remain challenges. In this flow, CNN is still used as a key foundation technique for deepfake detection, but complementary and extended research continues. Consequently, the success of CNN-based models has made important contributions to the popularization and commercialization of deepfake detection techniques.

2.2.3 Temporal Analysis and Recurrent Models

In video based deepfake detection, temporal coherence and variations in motion between frames serve as important clues. For this purpose, networks capable of processing time series data such as Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) have been applied. In particular, since LSTM can effectively learn long-term dependence, it can also detect minute variations between successive frames in the image. 3D CNN is also attracting attention for its structure that can learn spatial and temporal information simultaneously. These models have strength in detecting unusual sequence patterns such as blinking of a person's eyes, mouth movement, and natural movement of facial muscles. This is because it can capture not only frame-by-frame features but also high-dimensional patterns that reflect the flow of the entire video. Time series based models have mainly contributed to the improvement of detection performance for various video modulation techniques such as Deepfake, FaceSwap, and Face2Face. However, time series structures have long training and inference times, and there are practical constraints on the need to secure large video datasets. In addition, the

possibility of false detection can be increased in real image processing environments such as frame drop and encoding noise. Nevertheless, time series-based analysis has established itself as an essential key technique in video deepfake detection.

These Temporary Analysis based models can effectively capture traces of manipulation that are not identified by the static per-frame features alone. In particular, it is useful for identifying abnormal patterns in scenes containing various motions or emotional expressions. The natural flow and temporal consistency of videos are among the difficult parts for deepfake creators to emulate. Thus, the information obtained through Temporary Analysis plays an important role in increasing the reliability of deepfake detection. Recently, structures combining CNN and LSTM and hybrid models based on 3D CNN have also been actively studied. Moreover, self-supervised learning has enabled effective learning on large video datasets that lack labels. Video based detection shows higher practicality in real-world environments compared to still image-based models. However, there is also a disadvantage that real-time processing may be difficult due to complex network structures. In the future, the development of efficient computational structures and lightweight time series analysis models is expected to become more important research topics. Overall, the combination of Temporary Analysis and Recurrent Models has made significant progress in video-based deepfake detection.

2.2.4 Transformer-based Methods

Recently, Transformer architecture has been actively applied in image and video analysis as well. Specifically, Vision Transformer (ViT) and Swin Transformer are representative models proposed to overcome the limitations of CNN structures. These models can effectively learn global patterns and correlations within images by leveraging the self attention mechanism. ViT divides the input image into several patches and then processes the embedding of each patch by inputting it into the Transformer structure. This method has the advantage of being able to capture the contextual information of the entire image without relying solely on local features such as CNN. Transformer based models show the strength of deepfake detection to even subtle relationships between manipulated and non-manipulated areas. Indeed, several benchmark tests have also reported cases in which ViT family models outperform the existing CNN family. In addition, as the data size increases, the performance of the Transformer family model becomes more prominent. However, it is pointed out as a limitation that large computational resources are required and overfitting can easily occur with less data. Nevertheless, Transformer based approaches present a new paradigm in the field of deepfake detection.

Another advantage of the Transformer family is its scalability to various applications. As the attention mechanism verified in natural language processing has also been successfully applied in the field of computer vision, hybrid structure research is also actively underway. For example, models that combine CNN based feature extractors and Transformers to simultaneously exploit the advantages of both structures are emerging. Furthermore, Swin-Transformer introduces window based local attention to improve computational efficiency and is effectively applied to a variety of vision tasks. Deepfake detection is particularly important in that it can capture even abnormal global patterns or fine-grained texture variations in the face. Along with this, more and more research has recently extended Transformer structures to video analysis. Spatio-temporary Transformer, which can process the temporal pattern and spatial information of videos at the same time, is a prime example. These models further increase their applicability in real time deepfake detection and large data environments in the future. Overall, Transformer based methodology is positioning itself as a key research topic driving the future of deepfake detection technology. Various structural innovations are expected to continue in the future, focusing on

improving computational efficiency and data efficiency.

2.2.5 Fine-tuning of Pre-trained Models

In deep learning based deepfake detection research, fine-tuning using pre-trained models is one of the most widely used techniques. Pre-training refers to the process of obtaining the weights and characteristics of a pre-trained models from a large dataset such as ImageNet. This pre-trained model already fully learns common low and intermediate level features (edge, pattern, etc.) within the image. Finetuning further trains these pre-trained networks with a dataset specialized for deepfake detection, enabling them to reflect even domain-specific features. With this method, high performance can be expected with much less data and computation than training the model from start to finish. In addition, in situations where the data is limited or the labeling cost is high, fine-tuning increases both generalization performance and training efficiency. CNN as well as recent Transformer structures (ViT) have been applied to deepfake detection using fine-tuning methods after large-scale dataset-based pre-training such as ImageNet. As such, fine tuning contributes greatly to the fast convergence of deep learning models, reduction of overfitting, and improvement of domain adaptability. The effectiveness of fine tuning is demonstrated by superior performance differences over non pre-trained models in realworld benchmark experiments. Therefore, in recent years, both deepfake detection research and practical services, fine tuning has become a de facto essential approach.

Deepfake image and image data have very diverse characteristics depending on the generation method or operation technique. Due to this, it is difficult for one model to fully capture all variant types, and there is also a problem that the distribution varies from dataset to dataset. By applying fine tuning, in addition to the characteristics generally learned by the pre-learning model, the special patterns and artifacts of the dataset to be detected may be learned in detail. For example, GAN-based deepfakes, VAE-based variations, face synthesis, and other different types of manipulation can lead to different detection points, which enhance this detailed discrimination. In addition, fine tuning has the advantage of being able to quickly adapt the model with only a small amount of additional data, even if a new deepfake generation technique emerges that did not exist before. This flexibility greatly improves the maintenance, management, and scalability of deepfake detection models in real-world service environments. Several studies have reported that the fine-tuned model also has superior domain generalization ability among different datasets compared to single structured or non pre-trained models. In particular, when transfer learning strategies and data augmentation techniques are combined, practical detection performance can be maximized while preventing overfitting even in limited samples. As such, fine tuning is establishing itself as a powerful tool for reliable practical application in the field of deepfake detection. In the future, strategic use of fine tuning is expected to become increasingly important to cope with more diverse generation techniques and data environments.

The implementation method of fine tuning in real deepfake detection pipelines is also gradually being advanced. Basically, methods are used to fine-tune all the entire network weights, or selectively learn only specific layers (mainly upper layers). Depending on the situation, a multitasking fine tuning strategy that combines intermediate feature maps may be applied to learn by replacing only the classifier head. Recently, automated fine tuning methodologies such as hyperparameter optimization, adaptive learning rate adjustment, and fixed/unrelease layer selection have also been actively studied. One thing to be aware of in the fine tuning process is that if the difference between the original pre-training data and the domain data is too large, the initial weight may rather interfere with convergence. In this case, it is common to fine tune only some layers or freeze specific layers. The

effectiveness of fine tuning can also be confirmed through XAI (explainable artificial intelligence) analysis, which clearly shows a tendency to focus on real manipulation areas compared to pre-learning models. In practice, not only the performance of the fine tuning model, but also the ease of update, management cost, and suitability to the hardware environment are considered important evaluation points. In conclusion, fine tuning is the latest technological trend in deepfake detection based on deep learning and a practical essential element for practical service application. In this paper, a fine tuning-based model was used to analyze the practical improvement and applicability of deepfake detection performance.

2.2.6 Explainable AI (XAI) Techniques

The interpretability of deep learning based deepfake detection models has emerged as a very important issue for practical application and reliability. Accordingly, XAI (Explainable AI) techniques are being actively studied to visually explain on what basis the prediction results made by the model were derived. Gradient weighted class activation mapping (Grade-CAM) is typically used in CNN based models. Grad-CAM visualizes the image regions that contributed the most to a particular classification result, helping to determine the model's decision making basis. This method is useful for verifying that the manipulated regions in deepfake images actually match the regions the model focused on. In Transformer models, various interpretation techniques such as Attention Rollout and Attention Map visualization are applied. These methods visually represent the location within the image where self-attention is concentrated, helping to identify global patterns. The XAI technique is utilized as an important tool to provide trust not only to researchers but also to real users or domain experts. In particular, it plays a major role in ensuring transparency of deepfake detection results in areas where reliability is important, such as medical imaging and legal evidence. Recently, studies have also been conducted to quantitatively evaluate the performance of various XAI techniques.

The application of XAI goes beyond simple model interpretation and is also being used as a tool to diagnose model limitations and errors. For example, if the model made an incorrect prediction, XAI visualization can confirm that the area of focus is independent of the actual manipulation. This gives us clues to improving dataset quality or model structure. Moreover, the XAI technique is also utilized in the design of training data augmentation strategies to enhance the model's generalization capabilities. Recently, various XAI metrics and benchmarks have also been proposed to quantitatively analyze the decision making process of the model. XAI goes beyond just showing results and is recognized as an integral part of the actual system deployment and commercialization phase. The importance of transparent and interpretable deepfake detection technology is further emphasized to secure user trust and meet legal and ethical needs. It is expected that research to improve the performance, precision, and user-friendliness of XAI techniques will continue in the future. Ultimately, XAI is positioned as a key foundation technique for securing the reliability and practicality of deepfake detection models. This enhancement of interpretability is creating another innovative flow of deep learning-based deepfake detection research.

2.2.7 Benchmark Datasets and Evaluation Protocols

Utilization of standardized benchmark datasets is critical for research advances in deepfake detection and comparison of model performance. Representatively, various public datasets such as FaceForensics++, Celeb-DF, and DeepFake Detection Challenge are widely used. These datasets contain a variety of manipulation techniques, resolutions, and compression conditions similar to real-world environments, which play a major role in objectively evaluating the generalization performance of the model. FaceForensics++ offers a set of

tests segmented by different modes of operation, difficulty, and compression levels, helping researchers validate their models in different situations. Celeb-DF contains more delicate and sophisticated modulations along with high-quality realistic manipulation images, which are utilized to evaluate the limitations of state-of-the-art deepfake detection models. The DeepFakeDetection Challenge dataset reflects the real service experience by providing the vast amount of video data and various variations used in real competitions. These benchmark datasets allow researchers to compare model performance under the same conditions and facilitate technological advances. In addition, some datasets provide additional annotations such as manipulated parts, frame-by-frame labels, and thus are widely used in explainable AI studies such as XAI. Continuous expansion and update of benchmark datasets is an important foundation for securing reliability and improving practical applicability of deepfake detection techniques. As such, standardization of datasets is a key factor in supporting performance evaluation and technological competitiveness in deepfake detection.

Performance evaluation of deepfake detection models is performed using various classification metrics. Accuracy, Precision, Recall, and F1-score are typically used, and the importance of comprehensive indicators such as macro-average and weighted-average is emphasized in class imbalance situations. In real world environments, different analytical metrics such as ROC-AUC and Confusion Matrix are also used to evaluate the detailed strengths and weaknesses of the model. Moreover, efficiency indicators such as the number of parameters, amount of computation (FLOPs), and inference time of the model also play an important role in determining practical system applicability. In the field of deepfake detection, performance evaluation is required considering various realistic conditions such as real-time detection, mobile environment application, and hardware constraints. Along with quantitative indicators, qualitative evaluations are also being conducted to analyze the areas where the model is actually focused and the prediction basis through the XAI technique. Researchers are increasing the reproducibility and transparency of their findings by clearly disclosing datasets, evaluation protocols, and performance metrics. Recently, new evaluation methodologies reflecting real-world attack scenarios such as advertising attack and open-set testing have also been introduced. Diversification of evaluation criteria and strengthening reality play a crucial role in increasing the reliability, safety, and practical utilization of deepfake detection techniques. Advances in these different evaluation methods are expected to continue to promote technological advancement and practical social application in the field of deepfake detection.

2.3 Evaluation Metrics

2.3.1 Classification Metrics

The evaluation of the classification performance of the model was focused on Accuracy 2.1, Precision 2.2, Recall 2.3, and F1-score 2.4. Accuracy refers to the percentage of all predictions that are actually matched, and Precision is the percentage that is actually positive among samples predicted as positive. Recall represents the percentage of actual positive samples that the model correctly predicted with positive, and F1-score reflects the trade-off between the two indicators as a harmonized average of precision and reproducibility. These performance indicators allow the predictive power of each model to be evaluated from various perspectives. In particular, the overall performance of the model was comprehensively analyzed using methods such as macro/weighted average. Since macro average considers all classes equally regardless of the number of samples in each class, performance for minority classes can be evaluated in a balanced manner. On the other hand, since the weighted average is calculated by reflecting the number of samples per class, it reflects the overall

model's performance more realistically when a class imbalance exists. By considering the macro and weighted averages together, it is possible to objectively evaluate whether the model is biased toward a specific class. In addition to these standard indicators, detailed analysis by class can specifically identify vulnerabilities in specific classes. As such, quantitative indicator based evaluation is particularly important in applications where even small misclassification has a large impact, such as healthcare and security.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.1}$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.2}$$

where TP and FP denote the numbers of true positives and false positives, respectively.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.3}$$

where TP and FN denote the numbers of true positives and false negatives, respectively.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

where Precision and Recall are defined as above.

2.3.2 Confusion Matrix

The detailed distribution of classification results for each class was visually confirmed using the Confusion Matrix. The confusion matrix is a table 2.1 that shows at a glance what predictions the model actually made for each class and is an essential tool for analyzing the performance of the classifier in detail. Through this, it is possible to understand not only the degree of agreement between the correct answer class and the prediction class, but also the case of misclassification for each class. In particular, the model is very useful for error pattern analysis, such as in which particular class the misclassification occurs frequently and which class confusion is noticeable. This analysis allows us to systematically diagnose which classes the model has strengths or weaknesses. If there is a data imbalance, it is possible to visually evaluate the frequency of misclassification in minority classes and the impact on overall performance. In addition, even when the similarity between classes is high, misclassification patterns can be grasped in detail, helping to improve dataset quality or establish follow-up learning strategies. The confusion matrix makes it easy to calculate TP, FP, FN, and TN values for each class even in multi class classification situations. This information specifically reveals the detailed operating characteristics and problems of the model, which are difficult to know with simple performance indicators. Therefore, the confusion matrix is used as a key analysis tool in evaluating the performance of deepfake image detection models and deriving improvement directions.

2.3.3 Computational Efficiency Metrics

To evaluate the practical applicability of the deepfake detection model, comparisons were also conducted in terms of computational efficiency. In this paper, parameters, FLOPs (Floating Point Operations), and inference time were selected as major efficiency indicators. Params(M) was calculated as the total number of trainable parameters divided by one million, and FLOPs(G) was calculated as the sum of floating-point operations across all

Table 2.1: This table shows the structure of a confusion matrix, which is used to evaluate the performance of classification models. The confusion matrix provides an intuitive overview of how many samples were correctly or incorrectly classified for each class by crossing the actual and predicted classes.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

layers divided by one billion, as shown in Equations 2.5 and 2.6. As such, each model used several indicators to evaluate not only overall performance but also reliability and balance for a specific class. The performance evaluation method in this study is focused on reflecting various situations that are important when applied in practice in clinical settings, not just the percentage of correct answers. Reasoning time refers to the time it takes for input images to be processed in a real environment and to produce results, and it is used as a key measure in applications where securing real time performance is important. These efficiency indicators are essential for determining practical applicability in various limited resource environments such as mobile, embedded, and cloud as well as server environments. In particular, the need for models with high accuracy as well as computational efficiency is further emphasized in environments with limited computational resources. Since the efficiency value for each model may vary depending on the model design and hardware environment, comparison under the same conditions is important. In this paper, all models were measured in the same experimental environment, and the trade-off between performance and efficiency was clearly analyzed. Through this, it was possible to comprehensively evaluate the applicability in actual service sites or industrial applications beyond simple performance comparison.

$$\text{Params(M)} = \frac{\sum_{l=1}^{L} P_l}{10^6} \tag{2.5}$$

where P_l denotes the number of trainable parameters in layer l, and L is the total number of layers in the model.

$$\text{FLOPs(G)} = \frac{\sum_{l=1}^{L} F_l}{10^9} \tag{2.6}$$

where F_l denotes the number of floating point operations required for layer l, and L is the total number of layers in the model.

3 Related work

In the field of deepfake detection, various deep neural network structures and interpretable AI techniques are being applied. XceptionNet is a deep CNN architecture that effectively separates fine patterns and manipulation traces in images, showing very good performance in deepfake detection. EfficientNet is a model that achieves both parameter efficiency and computational optimization at the same time, recording high accuracy despite its relatively light structure and gaining popularity in practical applications. Vision Transformer (ViT) processes the entire image on a patch-by-patch basis based on the self-attention mechanism, enabling sensitive capture of global manipulation features. The eXplainable AI (XAI) technique helps to visually interpret what areas and features deep learning models have made predictions based on, such as Grad-CAM or LIME. XAI provides transparency to enable users to trust deepfake detection results, and plays an important role in preventing dangerous misjudgment in real-world applications. In addition, various auxiliary techniques such as facial landmark recognition, frame extraction, data augmentation, and ensemble are utilized to improve performance and general thermal power. Each model and technique has different advantages and disadvantages depending on dataset characteristics, real-world application environments, and state-of-the-art deepfake generation techniques, requiring comprehensive evaluation and further research. In conclusion, XceptionNet, EfficientNet, Vision Transformer, and eXplainable AI are positioning themselves as key technologies for modern deepfake detection. Please refer to the literature review table 3.1 for further details and assessment of previous studies.

3.1 XceptionNet

XceptionNet has been utilized as a very powerful CNN-based model in deepfake image detection. Combining XceptionNet and facial landmark recognition techniques, Saxena *et al.* [SYG+23] achieved over 96% accuracy and AUC 0.97 in image-based deepfake detection, and proposed a robust deepfake classification framework with a multi-input structure. A notable limitation of their work is that dataset diversity has been limited and generalizability for new manipulation techniques is limited. Ganguly *et al.* [GGM+22] presented a ViXNet model that combines Vision Transformer and XceptionNet, recording more than 98% of intra-dataset AUC and excellent generalization performance on various deepfake datasets (FF++, Celeb-DF, DFDC, etc.). However, the cross-dataset evaluation showed limitations focusing on inter-dataset performance degradation and facial region specificity. Ashok *et al.* [AJ23] reported using a single XceptionNet network to show high accuracy and strong generalization ability for unseen data on a wide range of image and video deepfake datasets. However, we mention the lack of validation on specific datasets and state-of-the-art deepfake techniques. Yasser *et al.* [YHEG+23] compared and analyzed EfficientNet and XceptionNet to verify the performance of XceptionNet through high accuracy and AUC and log loss-based evaluation in the authenticity classification of images and images. However, the limitation is that the experiment is limited to some datasets and two models, and that the quantitative result figures are not sufficiently presented. As such, XceptionNet has become a representative benchmark for deepfake detection, but additional research is still required in the diversity of datasets, adaptability to the latest manipulation technologies,

and cross-domain generalization.

3.2 EfficientNet-B0

EfficientNet is a lightweight, high-performance CNN structure widely utilized in recent deepfake detection studies, with various variations and combination methods being attempted. Yasser *et al.* [YHEG+23] compared and analyzed EfficientNet-B4 and XceptionNet to confirm the strengths of the two models through high accuracy and AUC and log loss-based evaluation in the authenticity classification of images and images. However, there was a limit to scalability due to experiments limited to two models and some datasets. To *et al.* [TLN+22] achieved 62.5% accuracy on the FaceForensics++ dataset through a pipeline combining MTCNN-based frame selection and EfficientNet, and verified the effectiveness of frame extraction and face separation techniques. However, one limitation of this study is that generalization performance is limited as it is applied only to a single dataset. Das *et al.* [DKT+25] proposed an ensemble model based on ResNet and EfficientNet, recording superior accuracy, precision, and reproducibility compared to existing baselines in precise face manipulation detection. This study is limited to still images and left the need to expand to multimodal detection. Koritala *et al.* [KCP+24] proposed a hybrid structure combining EfficientNet and LSTM, reporting 99.98% accuracy on the Celeb-DF dataset, showing a significant improvement in detection performance when time series information is combined. However, this study is also limited to a single dataset, so generalizability to other datasets has not been proven. As such, EfficientNet family models show excellent efficiency and accuracy in deepfake detection, but further studies are still required on dataset diversity, multimodal/video scalability, and practical generalization capabilities.

3.3 Vision Transformer

Deepfake detection research based on Vision Transformer (ViT) has recently shown strong performance on various datasets and transform structures. Ganguly *et al.* [GGM+22] demonstrated high accuracy and strong generalization performance over 98% of intra-dataset AUC on various deepfake datasets through ViXNet that combines ViT and XceptionNet, but revealed that limitations still exist in cross-dataset environments. Lin *et al.* [LHLL23] proposed a hybrid model that combines multi-scale convolution (CNN) and ViT, showing superior performance and wide versatility over existing methods on most public datasets. Their work emphasized abundant source domain-based generalization capabilities, but noted the need for further experiments in accurate numerical and external verification. Essa *et al.* [Ess24] applied feature fusion methods of the latest ViT family networks such as DaViT, iFormer, GPViT, and MLP-Mixer, achieving strong performance and high versatility, reaching up to 99.84% of the AUC on datasets such as FaceForronics++ and Celeb-DF. However, due to the high complexity of this model, the increase in computational demand in practical applications is pointed out as a disadvantage. Ramadhani *et al.* [RMU24] combined ViViT and various facial landmark features, DSC, and CBAM to record 87.18% accuracy and 92.52% F1-score on the Celeb-DF v2 dataset, and quantitatively analyzed the effects of each module. However, this study showed performance degradation in complex datasets, and the structural complexity of ViViT also remained a limitation. Overall, ViT-based deepfake detection research shows high performance and generalization ability, but it can be seen that overcoming the limitations of model complexity, computational demand, and domain generalization is still a challenge.

3.4 Gradient-weighted Class Activation Mapping (Grad-CAM)

XAI plays a key role in resolving the opacity of deep learning-based models and interpreting the model's decision making process. Abir *et al.* [AKA+23] compared the deepfake image detection performance using various CNN architectures and LIME, and demonstrated that the XAI technique is useful in improving the reliability and accuracy of prediction results. This study pointed out that it was verified only on still images limited to a specific source, and further verification of generalizability was needed. Jahmunah *et al.* [JNT+22] reported that DenseNet showed more than 95% accuracy and high interpretability through a deep learning-based myocardial infarction detection model applying Grad-CAM. Although it was possible to visually confirm that the main lead and clinically significant regions of the ECG signal were activated through Grad-CAM, it was applied to a single dataset only to mention the need for actual clinical verification. Zhang *et al.* [ZZS+22] applied MaizePestNet and Grad-CAM-based dual classification frameworks to achieve 93.85% corn pest identification accuracy and lightweight model implementation. Grad-CAM effectively eliminated non-core elements such as weeds and backgrounds, increasing their availability as real-world field systems, but revealed limited generalization performance due to background differences between training data and real-world field data. Li *et al.* [LLS+23] introduced multi-layer Grad-CAM (MLG-CAM) to achieve precise extraction of feature frequencies and positional interpretation power at the same time in vibration data-based defect diagnosis. It showed high reliability and efficiency on various datasets, but pointed out that a slight increase in computation time and further research is needed for future real-time system application. As such, XAI contributes to increasing the interpretability and practicality of models in various domains, but it requires limitations that most studies have been validated only in a limited environment, and in-depth verification of practical application and generalization performance.

3.5 Attention-based methods

Sarker *et al.* [SSAT24] combined attention mechanism based deep learning models with federated learning, and XAI to simultaneously improve the accuracy and interpretability of smart grid load prediction. This study applied 1D-CNN-GRU structure and SHAP-based analysis techniques to verify performance by achieving low MAE on various time series and geographic datasets. However, the complexity of federated learning and PSO tuning limits our generalization to other environments. Hasanpour Zaryabi *et al.* [HZMK+22] analyzed the effectiveness of the attention mechanism on the remote exploration building segmentation dataset through XAI techniques. Layer Grad-CAM, DeepLIFT, and various attention modules (SE, CBAM, etc.) are combined to demonstrate improvement in performance indicators such as accuracy, F1-score, and IoU. However, this study is limited to the building segmentation problem, and the possibility of expansion to other tasks has not been evaluated. Wang *et al.* [WYA+24] conducted a case study to predict emotional states using attention mechanism based XAI based on wearable multivariate time series data. This study integrates self-attention and graph-attention layers to provide insights of high accuracy and interpretable time series and modality levels. However, it is limited to a small sample and has a limitation that it is focused on a specific problem called emotion prediction.

3.6 Ensemble Learning Approaches

Ali *et al.* [ARuH+24] proposed an Ensemble Convolutional Neural Network-MF (ECN-MF) model for audio deepfake detection. The model combines various neural network structures such as RNN, 1D CNN, LSTM, and ConvLSTM with audio features such as Mel Frequency Copstrum Factor (MFCC), Chroma, and Zero Crossing Rate to effectively detect various audio manipulations. Experimental results demonstrate that ECN-MF achieves superior accuracy over existing individual models on multiple subsets of the Fake-or-Real dataset and improves the performance of audio deepfake detection. Recently, Srishti *et al.* [VGD+24] proposed an optical flow based ensemble deep learning model (OptiFake) to distinguish between deepfake and real images. By extracting the motion of pixels in the image, the method is demonstrated through the FaceForensics++ dataset that deepfake detection is possible with higher accuracy than the existing method. Specifically, the OptiFake model achieves 86.02% and 85.7% accuracy on deepfake and FaceSwap subsets, respectively, demonstrating the differentiation of the study as the first attempt to apply the ensemble model to the entire optical flow-based frame. Recently, weighted and evolutionary ensemble models using 3D CNN and CNN-RNN-based networks, as well as particle cluster optimization (PSO) based optimal structure and hyperparameter search techniques have been proposed by Zhang *et al.* [ZZL+24]. This study effectively extracts spatial temporal features through a new PSO algorithm, and organizes combinations of optimized networks into ensembles, showing high deepfake video detection performance compared to existing studies and other exploration techniques. The proposed PSO variant model also demonstrates statistically superior results over existing optimization methods on various artificial datasets and problems.

Table 3.1: This table is a literature review table that compares and organizes the datasets, methods used, experimental results, major contributions, and limitations of major papers published in Deepfake detection and eXplainable AI.

Year Published	**Paper Author and Citation**	**Paper Title**	**Dataset Used**	**Methods Used**	**Results**	**Contribution**	**Drawback/ Limitation**
2023	Saxena *et al.* [SYG+23]	Detecting Deepfakes: A Novel Framework Employing XceptionNet-Based Convolutional Neural Networks	Dessa, Deepfake Detection Challenge	XceptionNet, facial landmark	96% acc, AUC 0.97	Novel framework with facial landmark info	Limited dataset, only video

Year Published	**Paper Author and Citation**	**Paper Title**	**Dataset Used**	**Methods Used**	**Results**	**Contribution**	**Drawback/ Limitation**
2022	Ganguly *et al.* [GGM+22]	ViXNet: Vision Transformer with Xception Network for Deepfakes based video and image forgery detection	FF++, Celeb-DF, Deepfakes, DFDC	ViXNet (ViT +Xception), attention	AUC up to 99.26% (intra), 74.78% (inter)	Two-branch robust model, strong generalization	Low inter-dataset performance
2023	Ashok *et al.* [AJ23]	Deepfake Detection Using XceptionNet	Real/fake datasets (unspecified)	XceptionNet	High accuracy, generalizes to new data	Robust on both images/videos	Dataset details missing
2023	Yasser *et al.* [YHEG+23]	Deepfake Detection Using EfficientNet and XceptionNet	FF++, Celeb-DF	Efficient Net-B4, XceptionNet	High acc, evaluated by log loss/AUC	Comparative study of two CNNs	Details/ results limited, 2 datasets
2022	To *et al.* [TLN+22]	Deepfake Detection using EfficientNet: Working Towards Dense Sampling and Frames Selection	FF++	MTCNN, EfficientNet, frame selection	62.5% accuracy	Pipeline with face/frame selection	Single dataset, moderate acc
2025	Das *et al.* [DKT+25]	Enhanced DeepFake Detection Using CNN and EfficientNet-Based Ensemble Models for Robust Facial Manipulation Analysis	FF++	ResNet, EfficientNet Ensemble	High accuracy, improved metrics	Fine-grained detection, ensemble	Only still images, no multimodal

Year Published	**Paper Author and Citation**	**Paper Title**	**Dataset Used**	**Methods Used**	**Results**	**Contribution**	**Drawback/ Limitation**
2024	Koritala *et al.* [KCP+24]	A Deepfake detection technique using Recurrent Neural Network and EfficientNet	Celeb-DF	EfficientNet + LSTM	99.98% accuracy	Hybrid EfficientNet-LSTM approach	One dataset, generalization unclear
2023	Lin *et al.* [LHLL23]	DeepFake detection with multi-scale convolution and vision transformer	Celeb-DF, DFDC, FF++, Wild-Deepfake	Multi-scale CNN, ViT	Outperforms most methods, good generalization	Hybrid multi-scale & global learning	Datasets/ results not fully detailed
2024	Essa *et al.* [Ess24]	Feature fusion Vision Transformers using MLP-Mixer for enhanced deepfake detection	FF++, Celeb-DF	DaViT, iFormer, GPViT, MLP-Mixer	AUC up to 99.84%	Advanced ViT fusion, high generalizability	High model complexity
2024	Ramadhani *et al.* [RMU24]	Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attention	Celeb-DF v2	ViViT, facial landmarks, DSC, CBAM	87.18% acc, 92.52% F1	ViViT with landmark features, ablation	Only one dataset, complex model

Year Published	**Paper Author and Citation**	**Paper Title**	**Dataset Used**	**Methods Used**	**Results**	**Contribution**	**Drawback/ Limitation**
2023	Abir *et al.* [AKA+23]	Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods	Kaggle (Flickr/StyleGAN)	InceptionResNetV2, DenseNet-201, LIME	Acc up to 99.87%	CNN/XAI comparison, LIME explainability	Only still images, specific datasets
2022	Jahmunah *et al.* [JNT+22]	Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals	PTB ECG	DenseNet, CNN, Grad-CAM	>95% acc, DenseNet best	First explainable deep MI/ECG classification	Single database, no real-world validation
2022	Zhang *et al.* [ZZS+22]	A deep learning and Grad-Cam-based approach for accurate identification of the fall armyworm (Spodoptera frugiperda) in maize fields	Images of 36 maize pests	MaizePestNet, Grad-CAM, distillation	93.85% acc, small model size	Efficient pest ID, background removal	Real-world vs training mismatch risk
2023	Li *et al.* [LLS+23]	Multilayer Grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis	Simulated, public, self-collected	Multilayer Grad-CAM	Better impulse/-fault feature extraction	Clearer, trustworthy explanations	More compute, vibration data only

Year Published	**Paper Author and Citation**	**Paper Title**	**Dataset Used**	**Methods Used**	**Results**	**Contribution**	**Drawback/ Limitation**
2024	Sarker *et al.* [SSAT24]	Enhancing smart grid load forecasting: An attention-based deep learning model integrated with federated learning and XAI for security and interpretability	4 smart grid datasets	Attention 1D-CNN-GRU, PSO, SHAP, federated learning	Low MAE, high accuracy	Integrated privacy-preserving, interpretable smart grid forecasting	Hard to generalize; complex setup
2022	Hasanpour Zaryabi *et al.* [HZMK+22]	Unboxing the Black Box of Attention Mechanisms in Remote Sensing Big Data Using XAI	Remote sensing (buildings)	CNN + various AMs; XAI (Grad-CAM, DeepLIFT)	Improved all metrics; XAI confirms	Shows AMs' benefits; interprets via XAI	Limited to buildings
2024	Wang *et al.* [WYA+24]	Attention-Based Explainable AI for Wearable Multivariate Data: A Case Study on Affect Status Prediction	Wearable data (21 students)	Attention xAI (self-attention, GAT); PA/NA models	78% accuracy; interpretable	Reliable, interpretable health MTS analysis	Small data; only affect prediction
2025	Yujin Jeon	Proposed Work	deepfake and real images from Kaggle	XceptionNet, EfficientNet-B0, ViT	XceptionNet: 91%, EfficientNet-B0: 89%, ViT: 94%	ViT and CNN based models performance and explainability comparison	Limited to a single dataset

4 Research question

Research questions (RQs) set in this study focus on major issues and limitations that must be resolved for the development of the field. First, as a result of a close analysis of previous studies, it was confirmed that there are several problems that have not yet been completely solved. Accordingly, this paper intends to define these outstanding problems in the form of specific research questions. Research questions were set beyond just technical questions and aimed at solving practical problems and improving performance in real world application environments. In addition, the research questions fully reflected the data characteristics used in this study, experimental design, and structural differences between the applied deep learning model. Each question clearly presents the direction of the paper's experiment and serves as a reference point to lead the entire research process consistently. In particular, in this paper, detailed performance indicators and interpretability were widely considered so that new contributions could be derived through comparison with existing studies. Beyond simple performance comparisons, this approach provides a comprehensive view of the model's practical applicability and future research directions. By preparing clear evaluation criteria for each research question, the answer to each question was experimentally verified. As a result, the research question of this study was selected in consideration of both theoretical meaning and practical value, and it aims to contribute to the development of the field.

4.1 Problem statement

This study aims to present the optimal model selection criteria by quantitatively comparing the deepfake detection performance of the three representative deep learning models (ViT, XceptionNet, and EfficientNet-B0) and analyzing the visual explanation power of each model. In particular, the performance difference between models is evaluated fairly using a balanced image dataset, and visual grounds and judgment structures are derived through model with specific XAI techniques. To this end, the Attention Rollout technique is applied to ViT and Grad-CAM is applied to XceptionNet and EfficientNet-B0 to visualize which features each model pays attention to. Most of the existing studies focused on only a single model, so the comparison of explainability and computational efficiency according to structural differences was insufficient. Therefore, this study aims to promote the overall understanding of the deepfake detection model through comparative analysis between multiple architectures. In addition, the reliability of the visual explanation is verified by evaluating whether the model is focused and the actual manipulated image area. In this process, the consistency between the prediction results and visualization results of each model is also analyzed. In addition, the evaluation is performed from a practical point of view by considering not only the accuracy of the model, but also the calculation time and resource efficiency. Ultimately, this study aims to provide basic data for the development of a deepfake detection framework suitable for real-world applications by integrating performance, efficiency, and explainability. This is meaningful in presenting a practical and applicable guide in the field of deep learning-based image forensics.

4.2 Research question

1. RQ1: How do the architectural differences between ViT, XceptionNet, and EfficientNet-B0 impact their accuracy in detecting deepfake images?

   ViT, XceptionNet, and EfficientNet-B0 are architectures representing the Transformer and CNN based models, respectively, and have structural different characteristics. This research question aims to clarify how these structural differences affect accuracy in deepfake image detection. ViT utilizes a self attention mechanism to capture global information of images, while XceptionNet and EfficientNet-B0 extract local features via convolutional operations. I will compare and analyze what these differences actually lead to through experiments. Specifically, I quantitatively evaluate how well each model detects manipulated regions or fine modulations within an image. Prior studies have reported the strengths and weaknesses of each architecture, but direct comparisons under the same dataset and conditions are rare. Therefore, in this study, I systematically analyze the performance of the three models in the same experimental environment. Through this, the goal is to reveal the practical impact of each model's structural characteristics on deepfake detection accuracy. Ultimately, I want to demonstrate empirically how important architectural selection is for improving deepfake detection performance. These analysis results will provide an important criterion for model selection in the future development of deepfake detection systems.

2. RQ2: Which model among ViT, XceptionNet, and EfficientNet-B0 achieves the optimal balance between accuracy and computational efficiency in deepfake image detection?

   Deepfake image detection models must consider not only high accuracy but also practical computational efficiency. This research question focuses on how efficient each of the three models works and how the balance between accuracy and efficiency varies. I measure important efficiency indicators in a practical environment, such as the number of parameters, the amount of computation (FLOPs), and inference time. Some models show high accuracy, but require too much computational resources and time, making their practical application difficult. On the other hand, the lightweight model consumes less resources but can be concerned about performance degradation. Therefore, I compare and analyze how quickly and accurately each model behaves in real world situations. To this end, experiments are conducted on the same hardware and dataset to derive objective results. When considering both accuracy and efficiency at the same time, finding the best model is key. This analysis can provide practical help in implementing deepfake detection systems that can operate reliably even in resource-limited environments. Consequently, this work highlights the importance of model selection that achieves optimal balance of performance and efficiency.

3. RQ3: What insights can Explainable AI (XAI) methods, specifically Grad-CAM (for CNN based models) and Attention Rollout (for ViT), provide into the feature extraction strategies of each model for deepfake detection?

   Deep learning based deepfake detection models show high accuracy, but their internal operating principles remain "black boxes." Therefore, I want to apply the Explainable AI (XAI) technique to visually analyze how each model interprets and judges images. Grad-CAM visualizes important image areas in CNN based models, and Attention Rollout allows us to analyze areas of interest in ViT. This research question focuses on what insights these XAI techniques actually provide in the deepfake detection process. I compare which parts of the manipulated image each model pays attention to, and whether the information matches the judgment of human experts. Through

this, the basis for the prediction can be clarified, not just looking at the prediction results. In particular, this interpretability is essential for reliable model development and practical application. Experiments evaluate how closely the interpretation results of Grad-CAM and Attention Rollout relate to real deepfake detection. Such an approach can also help identify weaknesses or limitations of the model. In the end, this study aims to further increase the reliability and practicality of the deepfake detection system by transparently presenting the model's judgment basis through XAI.

4.3 Novelty of this study

The main novelty of this study lies in its comprehensive approach to balancing performance, efficiency, and explainability in deepfake detection. Existing studies have mainly focused on improving accuracy or analyzing individual elements separately. This study experimented with three architectures, ViT, XceptionNet, and EfficientNet-B0, under the same conditions to quantitatively compare performance and computational efficiency. This integrated comparison can be said to be a new attempt in that it allows you to grasp the structural characteristics and resource consumption of each model at a glance. In addition, the visual analysis of explainability by applying the appropriate XAI technique for each model is an important difference in this study. Through visualization analysis, the basis for judgment of the model was clarified, and as a result, the foundation for increasing user trust was laid. By verifying the consistency between the prediction result and the visualization result, the reliability of judgment was systematically secured beyond simple numerical evaluation. This verification procedure has not yet been sufficiently addressed in the field of deepfake detection, further highlighting the novelty of the study. Furthermore, this study proposed a new evaluation criterion that includes not only accuracy, but also efficiency and interpretability. This original approach is of great academic and industrial value in that it provides practical and reliable guidelines for model selection in real world environments.

4.4 Significance of Our Work

In this study, ViT, XceptionNet, and EfficientNet-B0 models were trained and evaluated under the same conditions. Visual explanatory power was analyzed by applying the XAI technique to each model, and the basis for judgment was verified through this. As a result of the experiment, ViT showed a sophisticated visual concentration area with high accuracy, and EfficientNet-B0 showed excellent efficiency with relatively low computational cost. XceptionNet showed a balanced performance between the two models and produced stable results. The visualization results obtained through Grad-CAM and Attention Rollout clearly showed the basis for determining the model, and reliability could be confirmed through the degree of agreement with the manipulated image area. In addition, the possibility of model interpretation was increased through the analysis of consistency between visualization and prediction results. Overall, this study comprehensively analyzed three key elements: performance, efficiency, and explanatory power. This is a practical basis for designing a reliable deepfake detection system in practical applications. The comparison of models of various architectures within one framework is significant in terms of practicality and academic contribution. Ultimately, this study presents the direction of future development of deep learning-based image verification technology.

5 Methods

In this chapter, I would like to systematically explain the entire research method conducted in this paper step by step. First, I introduce the types and configurations of the datasets used in the experiment. Next, main steps performed in the data preprocessing process will be described in detail. The reproducibility of the study was enhanced by describing the experimental environment and the software settings used. In addition, I briefly summarize the main structure and characteristics of each deep learning network architecture. Additional details for reproduction are also presented. It also describes the various techniques used in data analysis and visualization. Key evaluation indicators applied to objectively evaluate the performance of the model will also be described. Each section was constructed to have logical consistency with the flow of the entire research process. Through this, I aim to ensure that the methodology of this study is presented sufficiently transparently and can be easily reproduced in subsequent studies.

5.1 Methodology

In this study, three models (ViT, XceptionNet, and EfficientNet-B0) were trained under the same conditions on a balanced image dataset. In the data preprocessing process, images were divided into training, validation, and test, and normalization was applied to ensure consistency and stability of model training. Model performance was compared quantitatively, and visual explanatory power was analyzed using Grad-CAM and Attention Rollout. The interpretability of predictions was evaluated by visualizing how each model paid attention to the manipulated area. In addition, the analysis was performed by comprehensively considering three criteria which are accuracy, computational efficiency, and explainability. (Please refer to Fig 5.1 for the workflow.)

5.2 Dataset

In this study, a Kaggle public image dataset for machine learning based fake image detection was used. Since the size of the original dataset is very large, only some of them were selected and used. The selected data were divided into direct training sets (80%), validation sets (10%), and test sets (10%) and used for the experiment (Fig 5.1). The dataset consists of two classes, each clearly divided into a real image and a fake image (Please refer to Fig 5.2). The entire dataset is designed to minimize learning bias as it is organized in a balanced way. All images went through size adjustment (224×224) during the preprocessing process and were then converted into a form suitable for model input through normalization. The dataset composition and distribution are visualized in Fig 5.3. The training set was focused on improving the performance of the model, and the validation set was used to determine whether it was overfit or not. The test set was used for the final performance evaluation to measure the generalization ability. This dataset contains various facial features and background information, helping the model identify fake images even under various conditions. This configured dataset became an important basis for securing the experimental reliability of this study.

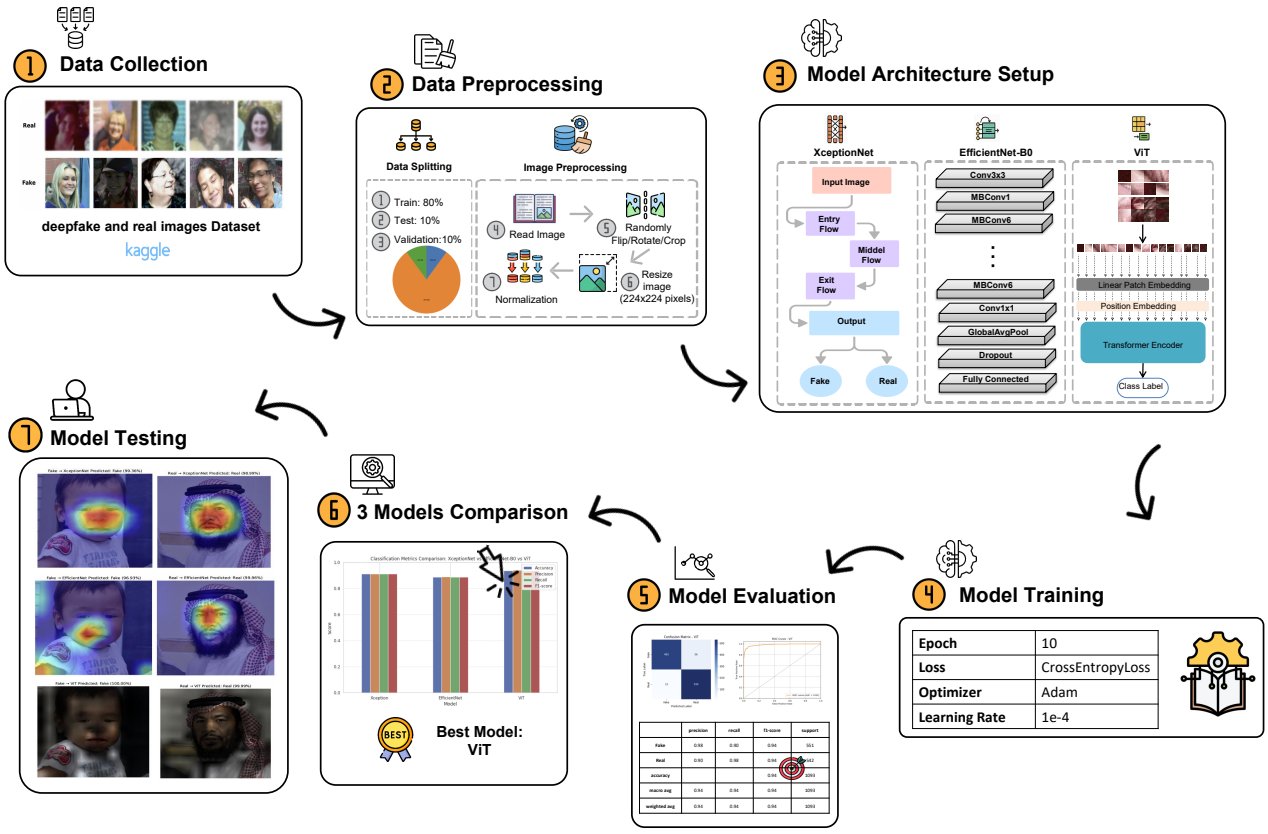

Figure 5.1: This figure illustrates the overall workflow of the deepfake image detection experiment, from dataset preprocessing and model training (ViT, XceptionNet, EfficientNet-B0) to performance evaluation and XAI-based visual explanation (Grad-CAM and Attention Rollout).

Table 5.1: This table shows the split distribution of the dataset used in this study. It was divided into training (80%), validation (10%), and test (10%)

	Real Class	**Fake Class**	**Total**
Train Set	4330	4401	8731
Validation Set	541	550	1091
Test Set	542	551	1093
Total	5413	5502	10915

5.3 Detailed Methodology

The first step of this study started with extracting some of the data for fake image detection from the public image dataset provided by Kaggle. Since the entire dataset is very massive, in this study, an appropriate quantity of data for learning and evaluation was selected in a balanced manner for each class. The selected data was configured to fit the binary classification problem consisting of a fake image and a real image. All images were resized to a size of 224×224 and went through a normalization process to be suitable for model input. After that, the data were divided into a ratio of 80% for learning, 10% for validation, and 10% for testing. To prevent the class imbalanc problem, the class ratio remained the

Figure 5.2: This figure shows sample images from the Kaggle dataset used in this study. The first row (real images) has been blurred to protect personal information, while the second row shows fake images with synthetic distortions that models learn to detect. The dataset includes diverse faces, lighting, and backgrounds to enhance generalization. (Dataset source: `https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images`)

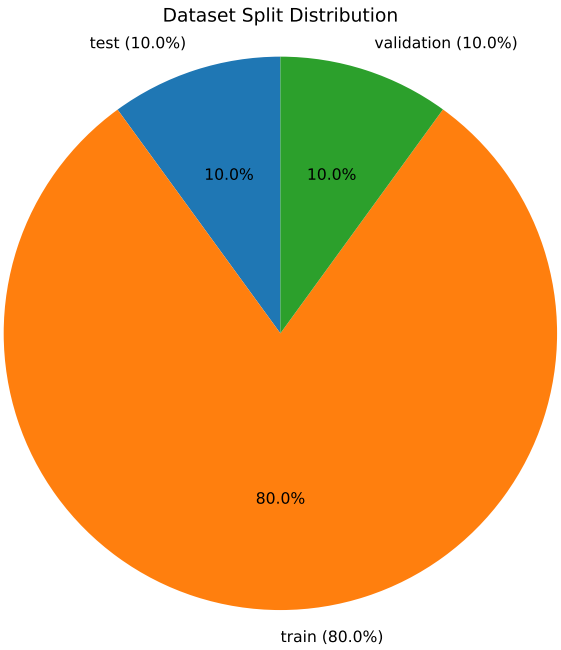

Figure 5.3: This figure shows the entire segmentation structure of the Kaggle image dataset used in this study. The data were divided into a ratio of 80% for training, 10% for validation, and 10% for testing. Each segmentation set is configured to maintain balance between classes and provides a basis for quantitatively measuring the performance evaluation and generalization ability of the model.

same even within each set. In this process, data selection and pre-processing criteria were strictly applied to increase the reliability of the experimental results. Data samples were checked for duplication, and the same image was managed not to be included in multiple sets. In addition, data segmentation was randomized, and samples from each class were evenly distributed. These data composition and segmentation methods play an important role in enhancing objectivity and reproducibility in evaluating the learning performance of the model.

For image data preprocessing and augmentation, several stages of transformation were applied using PyTorch's transforms.Compose. First, the RandomHorizontalFlip() function was used to randomly invert the left and right sides of the input image to increase data diversity. Random Rotation 20 generates images of various angles by arbitrarily rotating the image within a range of up to 20 degrees. ColorJitter changed the color properties of the image by randomly adjusting the brightness, contrast, saturation, and hue within the range of 0.2, respectively. This process is conducive to learning models that are robust to lighting or environmental changes. Furthermore, the spatial transformation of the input data was diversified by arbitrarily cropping the image through RandomResizedCrop (224) and then resizing it to a size of 224×224. The ToTensor() function transforms an image into a tensor, turning it into a form that deep learning models can process. Finally, Normalize (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]) was applied and normalized based on the mean and standard deviation for each channel. This normalization improves the stability and convergence speed of the training by consistently matching the distribution of the input data. Overall, this series of preprocessing and augmentation processes maximize the generalization performance of the model and enable feature learning robust to various situations.

In this study, three machine learning models, Vision Transformer (ViT), XceptionNet, and EfficientNet-B0, were used to compare and analyze the fake image detection performance. These models have structurally different characteristics, showing differences in interpretability as well as performance. All models applied the same dataset, learning conditions, and hyperparameters for comparison in the same training environment. CrossEntropyLoss was used as the loss function, Adam was set for the optimizer and the learning rate was set to 1e-4. Learning was conducted for a total of 10 epochs, and the overfitting of the model was determined through a verification set for each epoch. ViT processes the global information of the entire image in a Transformer structure, while XceptionNet and EfficientNet-B0 extract regional features with a convolutional based hierarchical structure. In the learning process, the loss is calculated in a batch unit and a backpropagation is performed to update the weights. Identifying how model structure differences affect detection accuracy, computational efficiency, and visual interpretation is one of the key objectives of this experiment.

After the model training was completed, the performance of the three models was quantitatively evaluated using the test set. Precision, recall, f1-score, and accumulation were used as major evaluation indicators, and the results and overall average of each class were analyzed. As a result of the experiment, ViT recorded the highest performance with an accuracy of 94%, and was selected as the best model. XceptionNet and EfficientNet-B0 also showed similar levels of f1-score, but EfficientNet showed particular strength in terms of computational efficiency. The evaluated performance results were analyzed through confusion matrices and classification reports, through which a practical criterion for model selection could be provided. In addition, the basis for visual judgment was extracted by applying Grad-CAM and Attention Rollout techniques to each model. The visualization results helped to understand which part of the image the model judged whether or not it was fake. ViT tended to focus more on the relatively manipulated face area, and the

CNN-based model considered a wider area. Explainability and reliability were evaluated together by analyzing the degree of agreement between the model's prediction results and visualization.

Lastly, the Grad-CAM technique was used to visually check which areas each model pays attention to in the test image. Grad-CAM visualization results intuitively show which parts of the image the model made the discrimination based on. This allowed us to verify whether the model actually captures meaningful features well or pays attention to the wrong parts. In addition, when misclassification occurred, it was possible to additionally analyze whether the model's visual precautions were appropriate through Grad-CAM. This analytical analysis goes beyond simple quantitative performance evaluation and plays an important role in transparently presenting the model's prediction process and judgment basis. As a result, this study secured the reliability and practical applicability of the model by performing a test set-based performance evaluation and an interpretability verification using Grad-CAM. In addition, the Attention Rollout technique was applied to the Vision Transformer (ViT) model to visually analyze the global information flow between each patch and the model's attention area. Through the Attention Rollout results, it was possible to understand in detail how the model extracts features from which part of the input image and integrates information in the judgment process. This analysis does not just interpret the prediction results, but also allows a transparent understanding of the internal operating principles of Transformer-based models. In the end, by using Grad-CAM and Attention Rollout together, it was possible to verify the predictive basis and interpretability of various models from various angles.

5.4 Evaluation Metrics

In this study, Accuracy, Precision, Recall, and F1-score were used to evaluate the performance suitable for the fake image detection problem. This problem is a binary classification problem composed of two classes, real and fake, and although the ratios between classes are balanced, each indicator can be interpreted differently. Accuracy is the correctly predicted rate of all test samples and is an indicator that evaluates the overall classification ability of the model. ViT showed the best performance with Accuracy 0.9378, followed by XceptionNet with 0.9131 and EfficientNet-B0 with 0.8893. Precision refers to the proportion of samples that a model classifies into a specific class and represents the ability to reduce false positives. Recall is the proportion of samples that the model classifies correctly among samples belonging to a specific class and represents the ability to reduce false negatives. F1-score is calculated as the harmonized average of Precision and Recall and is an indicator that reflects the performance in a balanced way when there is an imbalance between classes. These three indicators were calculated as weighted average (weighted average considering the number of samples per class) and macro average (simple average between classes), respectively, and both methods are included in the table 5.2.

Weighted average and Macro average are very important indicators in model evaluation, providing different perspectives. Since the weighted average is calculated by considering the number of samples in each class, it represents the overall performance reflecting the actual data distribution. This indicator is characterized by being greatly influenced by the dominant class when there is a class imbalance in the dataset. On the other hand, since macro average, which is a simple average for each class, evaluates each class with the same weight, it is possible to grasp the model's performance for a few classes in a balanced way. Therefore, macro average is useful for determining whether the model works fairly well for all classes. By checking both means in the experiment, it is possible to detect and improve whether there is bias against a particular class early. If you judge only by looking

at weighted average, there is a risk that only the performance of multiple classes will be highly evaluated. If macro average is further analyzed, it can be verified whether it works stably even in a small class, thereby increasing overall reliability. Especially in security and verification fields such as deepfake detection, few class false detections are even more important because they can pose large risks. Consequently, simultaneous analysis of both indicators is essential to reduce the bias of the model and to ensure its reliability and effectiveness in an actual use environment.

Table 5.2: This table shows the results of comparing the performance of the three models (XceptionNet, EfficientNet-B0, and ViT) for Deepfake image detection based on Accuracy, Precision, Recall, and F1-score. ViT recorded the best performance in all indicators. All metrics are reported in abbreviated form: Acc. (Accuracy), Prec. (Precision), Rec. (Recall), F1 (F1-score), with (W) indicating weighted average and (M) indicating macro average.

Model	**Acc.**	**Prec. (W)**	**Rec. (W)**	**F1 (W)**	**Prec. (M)**	**Rec. (M)**	**F1 (M)**
XceptionNet	0.9131	0.9131	0.9131	0.9131	0.9131	0.9131	0.9131
EfficientNet-B0	0.8893	0.8901	0.8893	0.8892	0.8902	0.8891	0.8892
ViT	0.9378	0.9407	0.9378	0.9377	0.9404	0.9381	0.9377

5.5 Experimental settings

The experiment in this study was conducted by comparing the three models (ViT, XceptionNet, and EfficientNet-B0) in the same hardware and learning environment. The experiment was conducted in the Google Colab environment based on the PyTorch framework, and the GPU used NVIDIA Tesla T4. All models were subjected to the same image size (224×224) and the same data preprocessing method (normalization and augmentation), and CrossEntropyLoss was used as the loss function and Adam was used as the optimizer. The learning rate was fixed at 1e-4, the training batch size was set to 16, and the validation and test batch size was set to 32. Each model was trained for a total of 10 epoch. The dataset was divided at a ratio of training (80%), validation (10%), and test (10%), and the ratio between classes was configured to balance. All experiments were repeated for each model under the same conditions to ensure fairness in performance comparison. Evaluation indicators were calculated based on Accuracy, Precision, Recall, and F1-score, and both weighted average and macro average methods were used. For explainability analysis, Attention Rollout was applied to ViT and Grad-CAM techniques were applied to CNN-based models (XceptionNet, EfficientNet-B0). Please refer to the Table 5.3 for network configuration.

XceptionNet is a deep convolutional neural network structure that shows excellent performance in computer vision problems such as image classification. Instead of traditional convolutional operations, the model uses Depthwise Separate Convolution to significantly reduce the number of parameters and the amount of computation, while maintaining high accuracy. The network is largely divided into three stages: Entry Flow, Middle Flow, and Exit Flow. In Entry Flow, the input image is gradually extended to more channels and features are extracted. MiddleFlow effectively learns complex features by repeating blocks of the same structure eight times. Each block consists of multiple SeparateConv2d operations, batch normalization, and ReLU activation functions. In Exit Flow, the number of channels is further extended, and a final feature map is obtained. Then, the spatial information is reduced through the Global Average Pooling layer. After that, The final Fully

Table 5.3: This table summarizes the network training environments and key hyperparameters used in this study, ensuring reproducibility and fair comparison across models.

Network Configuration	
Epochs	10
Learning rate	0.0001
batch size (train/val/test)	16/32/32
Optimizer	Adam
Loss	CrossEntropyLoss
num_classes	2
Pretrained	True

Connected (Linear) layer outputs the probability for each class. Thanks to these structures, XceptionNet is widely used as a representative deep learning model that captures both efficiency and performance. Please refer to (Fig 5.4) for the network architecture.

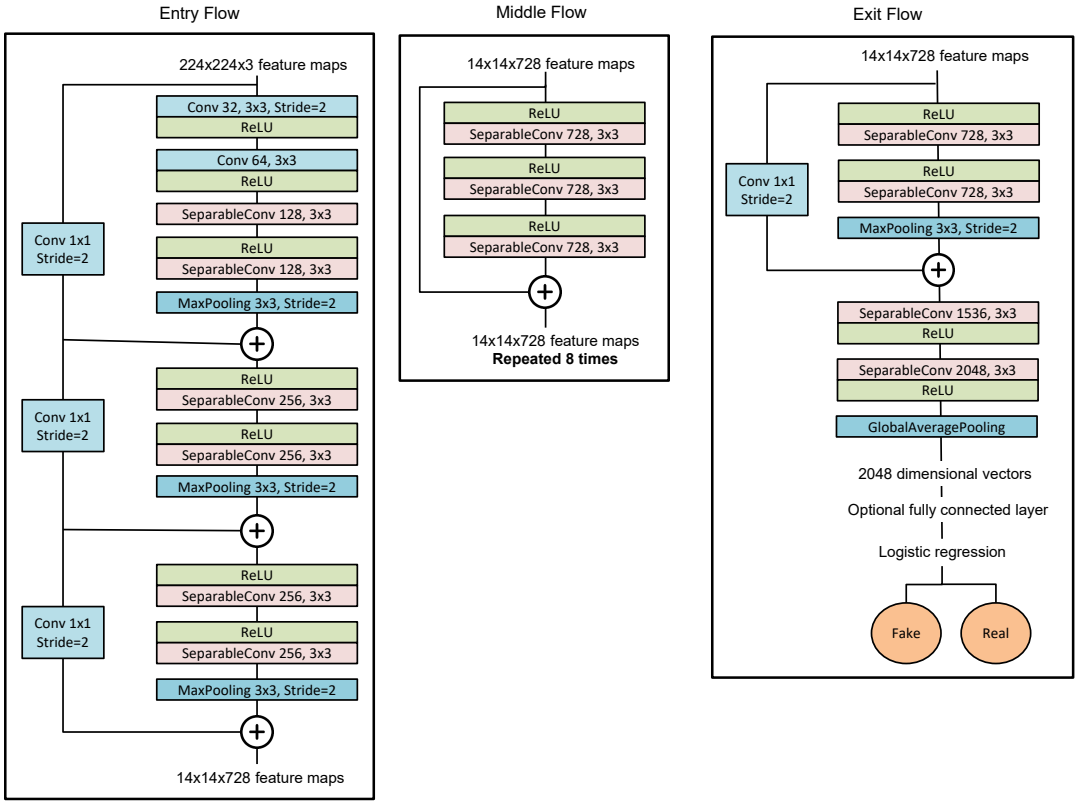

Figure 5.4: This figure presents the structure of the XceptionNet architecture, illustrating how depthwise separable convolution is applied throughout the entire network, from the Entry Flow to the Middle Flow and Exit Flow stages. Each stage leverages depthwise separable convolution to efficiently extract hierarchical features while significantly reducing the number of parameters and computational cost, compared to standard convolutions.

EfficientNet-B0 is a representative deep learning network that has achieved both model

weight reduction and performance improvement. The model effectively preprocesses the input image in a Stem step starting with Conv2d, BatchNorm, and SiLU activation functions. After that, the structure of MBConv, which combines Depthwise Separate Convolution and Squeeze-and-Excitation (SE) blocks, will be applied in earnest. In the middle of the network, the Inverted Residual block and SE module, which are used repeatedly, are included to maximize the efficiency of feature extraction. Each block has a different number of channels, kernel size, and stride value, so the spatial resolution and number of channels of the characteristic map are adjusted step by step. Specifically, the MBConv6 block is repeatedly applied several times, increasing the depth and expressiveness of the model. In the second half Head region, Conv2d, BatchNorm, and SiLU are applied once more to refine the feature map. The spatial information is then compressed into one dimension through Global Average Pooling. Finally, the class-specific probability value is finally calculated through the Fully Connected (Line) layer. Because of these structures, EfficientNet-B0 shows excellent results in both parameter efficiency and performance. Please refer to (Fig 5.5) for the network architecture.

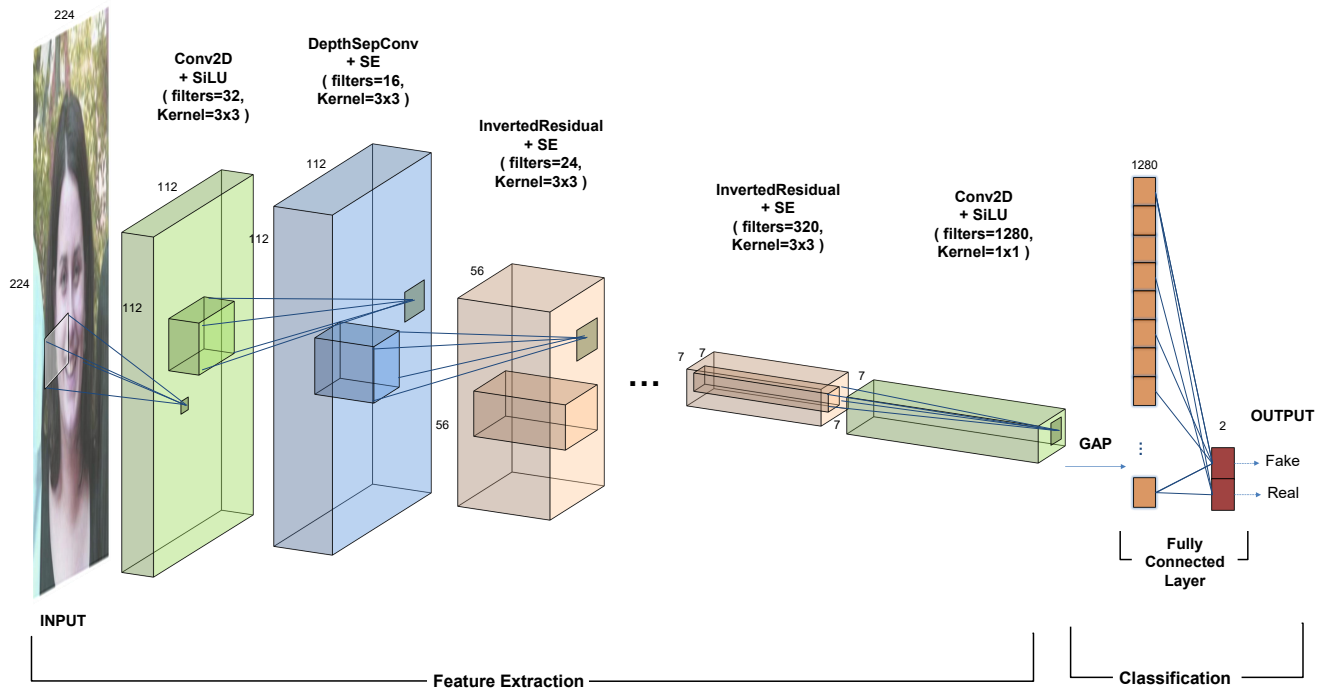

Figure 5.5: This figure shows the EfficientNet-B0 architecture, from the Stem and MBConv blocks (with depthwise separable convolution, SE modules, and skip connections) to the Head, pooling, and classifier, illustrating how it balances depth, width, and resolution efficiently.

Vision Transformer (ViT) is an innovative model that applies transformer structures to images as they are for image classification problems. The input image of the ViT is first resized to a fixed size (224x224). The image is then segmented into patches of constant size (16x16), each of which is spread into a one-dimensional vector and embedded. These embedding vectors are transformed into input sequences, and each patch is input to the transformer encoder like a word token. The model adds a positive encoding to add location information to each patch embedding. Thereafter, several transformer encoder blocks are sequentially applied. Each block consists of multi head self attention and feedforward networks, which can learn global relationships between all patches. Finally, the embedding

corresponding to the Classification (CLS) token among each patch vector becomes the feature vector representing the entire image. This CLS token is converted into a class-specific probability value through a linear layer for final classification. Unlike traditional CNNs, ViT can effectively learn global features for the entire input image without local computational constraints. Thanks to these structures, ViT demonstrates performance beyond existing CNN-based models when large amounts of data and computational resources are sufficient. Please refer to (Fig 5.6) for the network architecture.

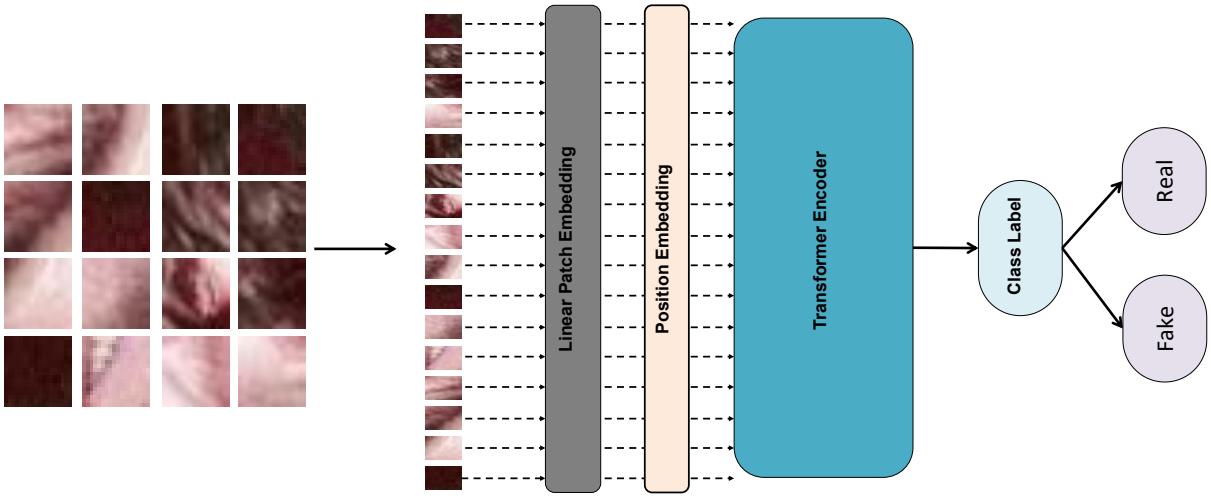

Figure 5.6: This figure shows the Vision Transformer (ViT) architecture, illustrating the flow from image patching and embedding with positional encoding, through transformer encoder blocks, to final classification.

6 Results

6.1 Model Performance

6.1.1 XceptionNet

During the training process of the XceptionNet model, Loss decreased rapidly at the beginning of training and showed stable convergence without overfitting. There was little difference between validation performance and training performance, so the generalization ability was excellent. Accuracy, Precision, Recall, F1-score, and AUC all remained above 0.90 in Training and Validation, and the performance gradually improved as Epoch increased (as shown in Fig 6.1). Confusion Matrix correctly classifies 998 out of a total of 1093 samples, showing high accuracy in both real and real samples (as shown in Fig 6.2). The model accurately predicted 503 in the Fake class and 495 in the Real class, and 48 and 47 cases of misclassification occurred, respectively. The overall accuracy was recorded at about 91.3%. The model shows overall stable classification performance without any class-to-class imbalances. On ROC curves, the AUC value is 0.98, demonstrating that XceptionNet has near perfect discriminant power in Deepfake detection (as shown in Fig 6.3). These results suggest that XceptionNet is a highly reliable model in Deepfake image detection. In particular, the high AUC and balanced confusion matrix increase its applicability in real-world environments. In conclusion, XceptionNet performed well in terms of precision, reproducibility, and overall classification performance.

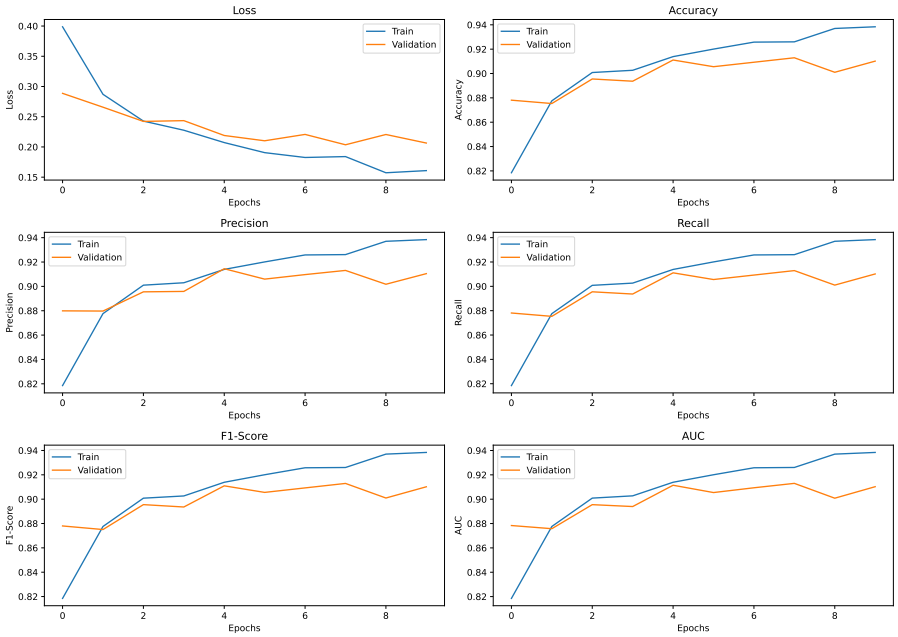

Figure 6.1: Training and validation metrics (Accuracy, Precision, Recall, F1-score, AUC, and Loss) per epoch for the XceptionNet model.

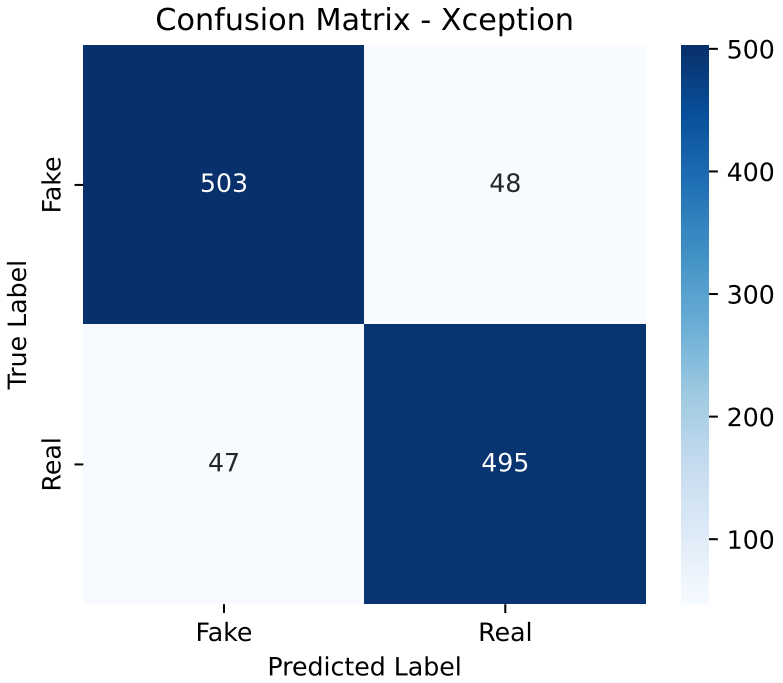

Figure 6.2: Confusion matrix for XceptionNet on the test set, showing the distribution of true and predicted labels for Fake and Real classes.

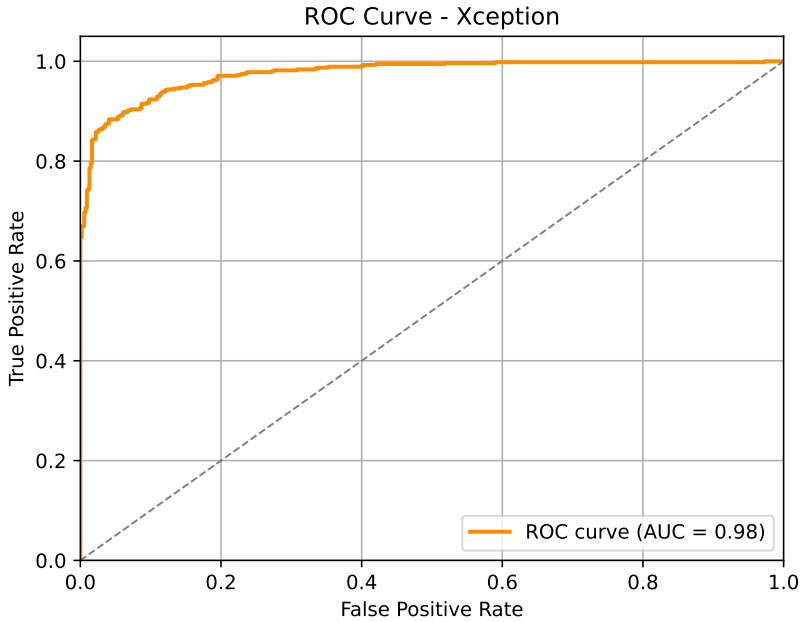

Figure 6.3: Receiver Operating Characteristic (ROC) curve and the area under the curve (AUC) for XceptionNet, indicating the model's overall discrimination ability.

6.1.2 EfficientNet-B0

In the learning process, Accuracy, Precision, Recall, F1-score, and AUC tended to increase overall as Epoch increased. Accuracy of Train and Validation gradually increased to about

0.82 levels at Epoch 1 and 0.94 levels at Epoch 10. Loss continued to decrease as the echo increased, confirming the stability of learning. Major performance indicators such as Precision, Recall, F1-score, and AUC also recorded a range of 0.88 to 0.90 based on Epoch 10 (as shown in Fig 6.4). The Confusion Matrix results of the EfficientNet-B0 model accurately predicted 503 in the Fake class and 469 in the Real class, with 48 and 73 misclassifications, respectively (as shown in Fig 6.5). Out of a total of 1093 samples, the overall accuracy was recorded at about 88.9%. Fake class Precision was 0.87, Recall was 0.91, F1-score was 0.89, Real class Precision was 0.91, Recall was 0.87, and F1-score was 0.89. In the ROC curve, the AUC value was 0.96, showing high classification performance (as shown in Fig 6.6). The EfficientNet-B0 model has shown high overall performance and stable learning results.

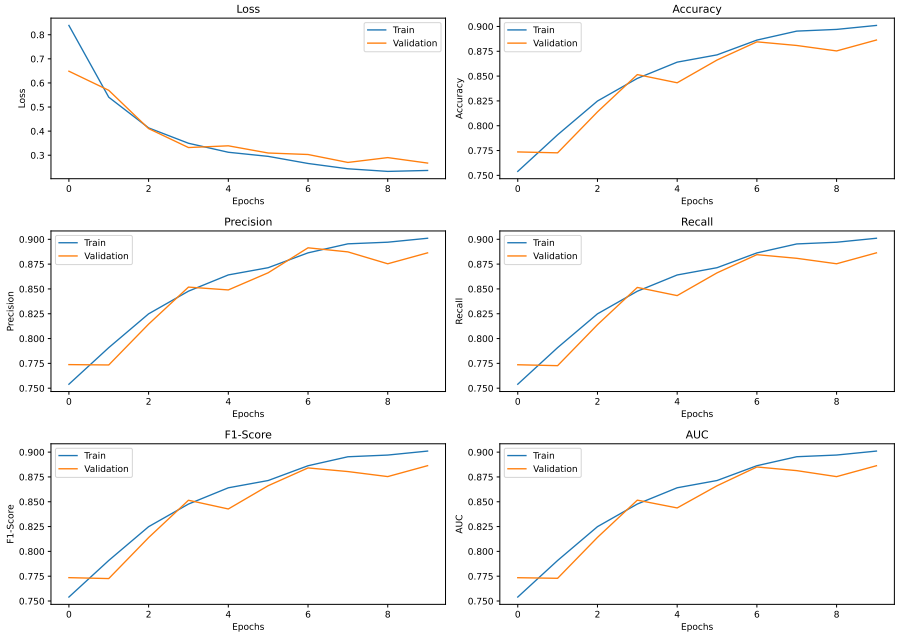

Figure 6.4: Training and validation metrics (Accuracy, Precision, Recall, F1-score, AUC, and Loss) per epoch for the EfficientNet-B0 model.

6.1.3 ViT

Accuracy of Train and Validation increased to approximately 0.89 levels at Epoch 1 and 0.99 levels at Epoch 10. Loss shows that learning was made stably by steadily decreasing as the echo progressed (as shown in Fig 6.7). According to the Confusion Matrix results of the Vision Transformer (ViT) model, 495 were accurately predicted in the Fake class and 530 in the Real class, and 56 and 12 misclassifications occurred, respectively (as shown in Fig 6.8). Out of a total of 1093 samples, the overall accuracy was recorded at about 93.8%. Fake class Precision was 0.98, Recall was 0.90, and F1-score was 0.94, Real class Precision was 0.90, Recall was 0.98, and F1-score was 0.94. In the ROC curve, the AUC value was 0.99, showing very good discrimination performance. In the learning process, Accuracy, Precision, Recall, F1-score, and AUC continued to improve as Epoch increased (as shown in Fig 6.9). The ViT model has achieved excellent performance in Deepfake image detection based on its high accuracy and discriminant power. Overall, the ViT model showed results that secured both efficiency and reliability.

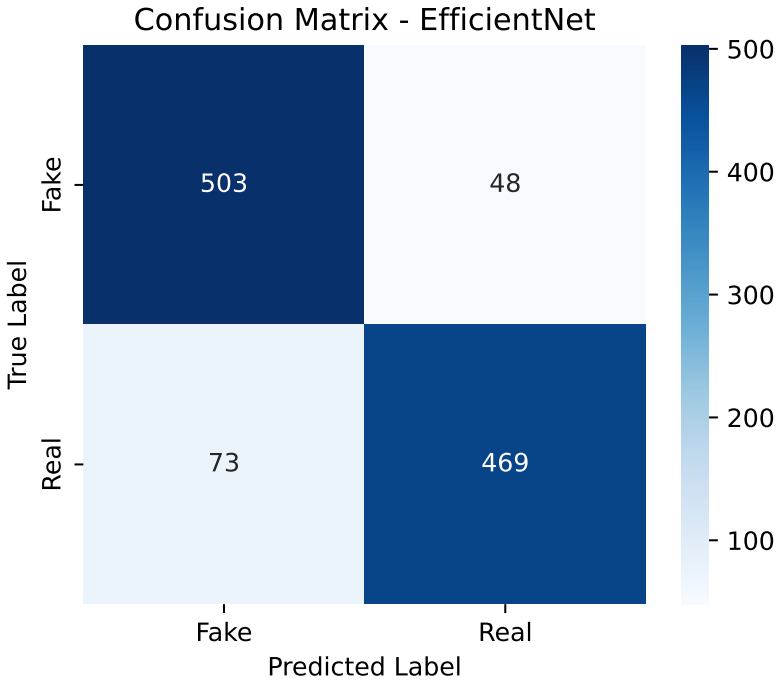

Figure 6.5: Confusion matrix for EfficientNet-B0 on the test set, showing the distribution of true and predicted labels for Fake and Real classes.

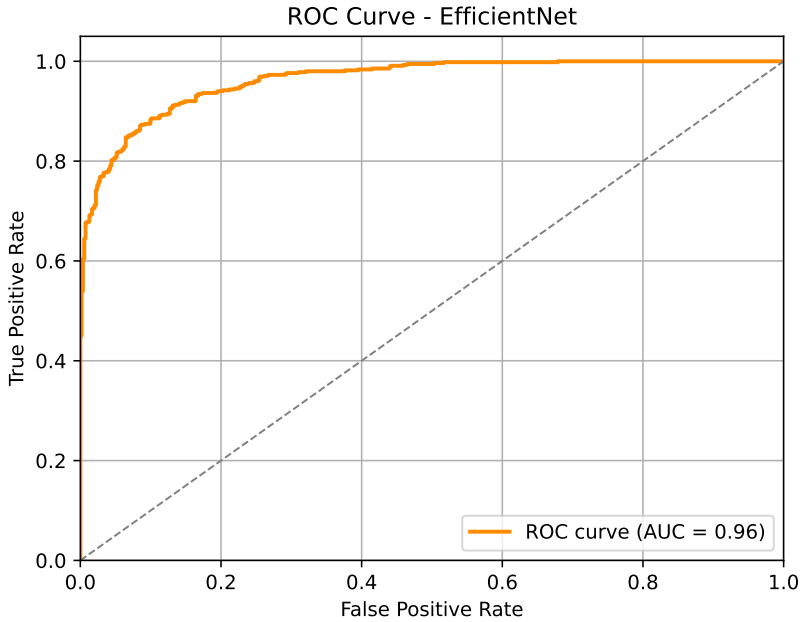

Figure 6.6: Receiver Operating Characteristic (ROC) curve and the area under the curve (AUC) for EfficientNet-B0, indicating the model's overall discrimination ability.

6.1.4 Model Comparison

Structural differences between the three models (ViT, XceptionNet, and EfficientNet-B0) have shown distinct differences in deepfake image detection performance. Experimental

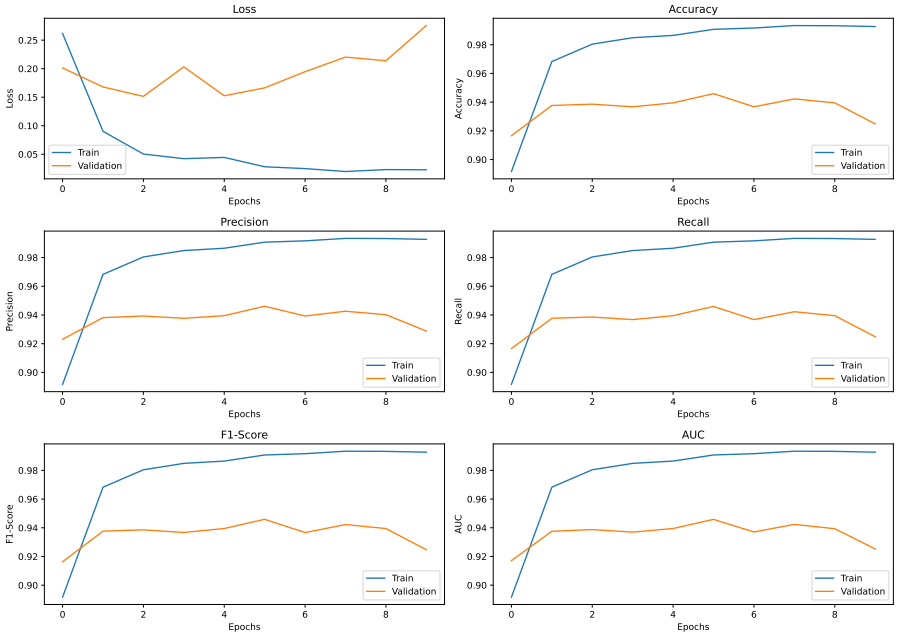

Figure 6.7: Training and validation metrics (Accuracy, Precision, Recall, F1-score, AUC, and Loss) per epoch for the ViT model.

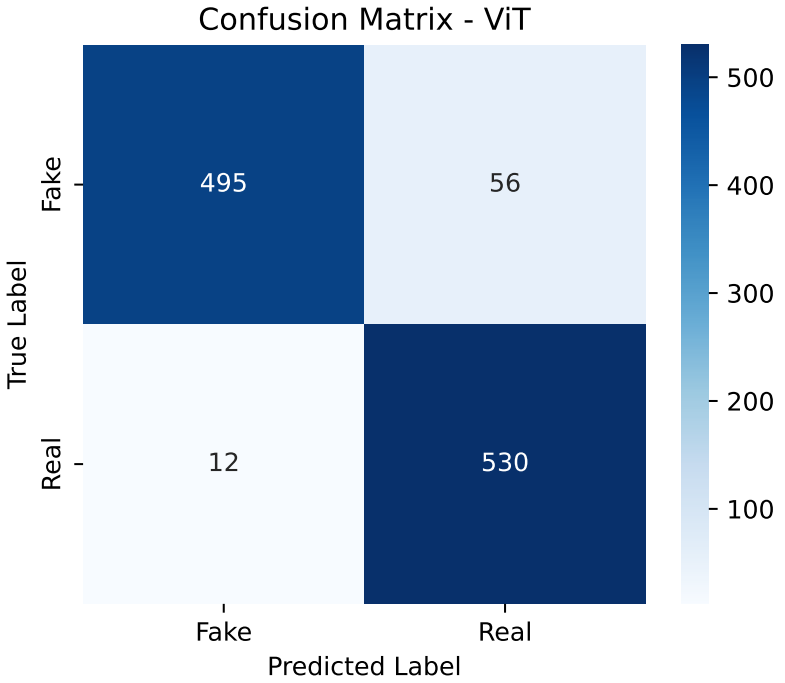

Figure 6.8: Confusion matrix for ViT on the test set, showing the distribution of true and predicted labels for Fake and Real classes.

results show that ViT achieves the highest accuracy on the entire test set. XceptionNet also recorded high accuracy, but it was slightly lower than ViT. EfficientNet-B0 performed well despite its relatively lightweight structure. In addition to Accuracy, similar trends were observed in the macro F1-score and weighted F1-score indicators. On the other

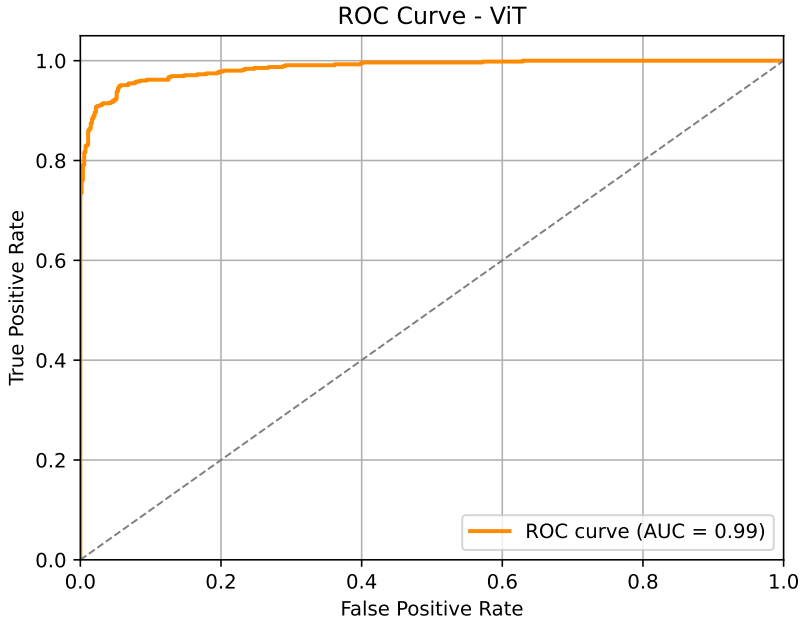

Figure 6.9: Receiver Operating Characteristic (ROC) curve and the area under the curve (AUC) for ViT, indicating the model's overall discrimination ability.

hand, XceptionNet showed high precision in some classes, but did not reach ViT in recall. EfficientNet-B0 showed the lowest accuracy among the three models, but achieved above baseline performance as a lightweight model. Fig 6.10 shows the results of comparing the main performance indicators of the three models with a bar graph. As a result, it was confirmed that the structural difference of the model directly affects the accuracy of deepfake image detection.

6.2 Model Complexity

The accuracy and computational efficiency of each model were comprehensively evaluated based on the number of parameters, the amount of computation (FLOPs), and the total learning time (min). ViT achieved the highest accuracy among the three models, recording the most parameters of 85.8M and the amount of computation reaching 16.87G. However, ViT also took 51.69 minutes to learn, which was the most burdensome in terms of resource use. XceptionNet was less than ViT in both the number of parameters (20.81M) and the amount of computation (4.6G), but it still showed a significant scale, and the learning time was relatively long at 46.86 minutes. On the other hand, EfficientNet-B0 had the smallest number of parameters (4.01M) and FLOPs (0.39G), and the learning time was the fastest among the three models at 21.59 minutes. These results mean that EfficientNet-B0 has both reasonable accuracy and very high computational efficiency even though it has a lightweight structure. XceptionNet and EfficientNet-B0 showed a practical balance in both accuracy and efficiency, but it can be seen that ViT is an option that can only be considered when you want the best performance. EfficientNet-B0 is judged to be the most competitive model in situations where computational resources are limited, such as actual service environments or real-time detection. As a result, the three models showed a clear trade-off in terms of accuracy and computational efficiency. Please refer to the table 6.1

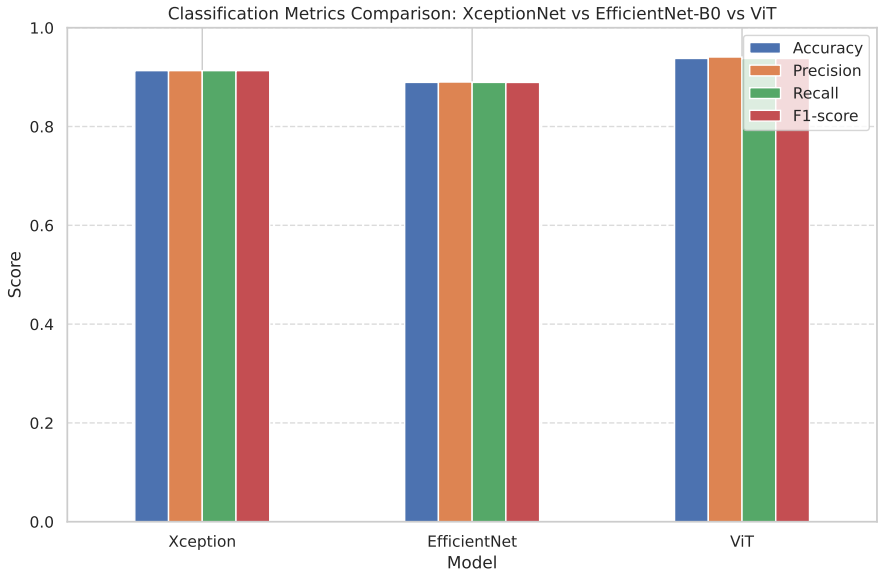

Figure 6.10: Bar graph comparing the deepfake image detection performance of three models (ViT, XceptionNet, and EfficientNet-B0). ViT shows the highest performance, followed by XceptionNet, with EfficientNet-B0 recording the lowest values among the three.

Table 6.1: Comparison results of XceptionNet, EfficientNet-B0, ViT's number of parameters (in millions), FLOPs (in billion), and total learning time (in minutes). The computational efficiency and resource consumption characteristics of each model are shown at a glance.

Model	Params (M)	FLOPs (G)	Training Time (min)
XceptionNet	20.81	4.6	46.86
EfficientNet-B0	4.01	0.39	21.59
ViT	85.8	16.87	51.69

6.3 eXplainable AI (XAI)

The feature extraction method of each model was visually analyzed through Grad-CAM (Fig 6.11, Fig 6.12) and Attention Rollout (Fig 6.13) techniques. As a result of visualization using Grad-CAM, XceptionNet and EfficientNet-B0 models showed strong activation responses in the eyes, nose, and mouth, which are the main facial regions of the image. These models showed a tendency to focus mainly on the area where deepfakes apply modulation, and showed a high degree of agreement with the actual deepfake modulation site. The results of Grad-CAM are mainly expressed as red and yellow color heatmaps, allowing users to intuitively identify modulated areas. The darker the red color in the activation map, the higher the interest the model has in the region. Both models produced more pronounced and clearer activation patterns in the strongly modulated region, suggesting that they effectively capture real modulated signals. On the other hand, it was confirmed that less activation appeared in areas where modulation was weak or did not exist. In particular, EfficientNet-B0 showed relatively more detailed region segmentation and distinct boundary representation, while XceptionNet tended to have activations distributed over a large area. These differences are attributed to the structural properties and filter composition of the two models. In conclusion, visualization with Grad-CAM clearly demonstrates

that CNN family models focus their attention on the main facial regions where real deepfake modulation has been made. In this process, it was also confirmed that each model's internal feature extraction method was reflected in the visualization results, resulting in slightly different heatmap patterns for each model.

For the ViT model, visual analysis was performed by applying the Attention Rollout technique. As a result of Attention Rollout, ViT showed a wide distribution of regions of interest across the entire face region. Unlike CNN family models, this reflects the structural properties of simultaneously distracting attention to multiple sites, rather than focusing only on a limited specific site. In particular, the activation map of ViT is represented in the form of a grayscale heat map rather than color, showing a large region of interest that is visually smoothly connected. It shows a pattern that encompasses not only the area where deepfake modulation was performed, but also the entire face, and activation was observed simultaneously in several areas. This suggests that the self attention mechanism of transformer based models effectively incorporates broad contextual information. In addition, ViTs have shown a tendency to maintain a certain level of interest not only in the vicinity of the modulation region but also in the non modulation region. This highlighted the nature of spreading modulation detection signals over a wider space compared to CNN family models. Visualization with Attention Rollout clearly shows that ViT's internal information processing method is fundamentally different from CNN-based models. In the end, ViT integrates information distributed and comprehensively across the face, while CNN based models are able to identify differences that selectively focus on highly modulated areas.

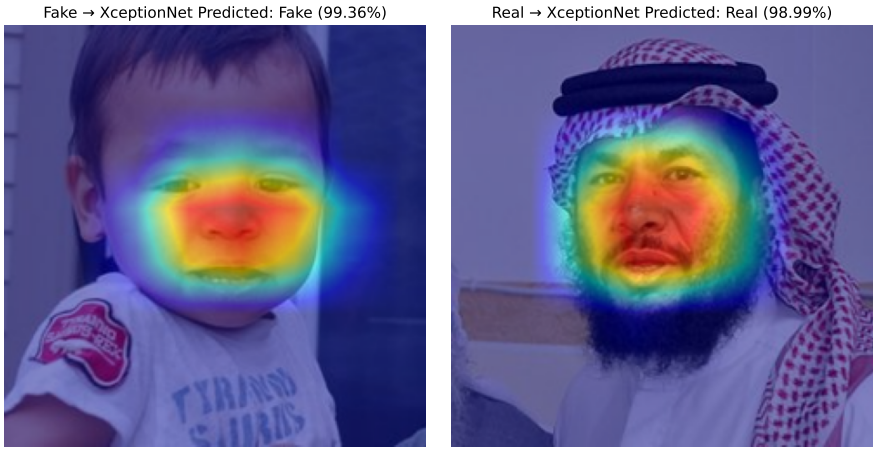

Figure 6.11: This figure is Grad-CAM visualization result of XceptionNet model. The left is a deepfake image and the right is a result of a real image. Closer to red in each image means an area that the model considers more important for classification, and closer to blue indicates an area of relatively low importance.

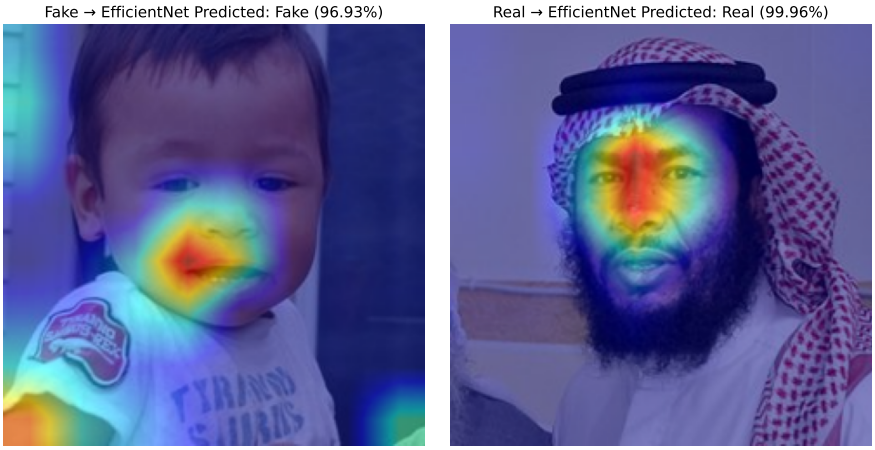

Figure 6.12: This figure shows Grad-CAM visualization results of the EfficientNet-B0 model. The left side shows the deepfake image, and the right side shows the area of attention for the real image.

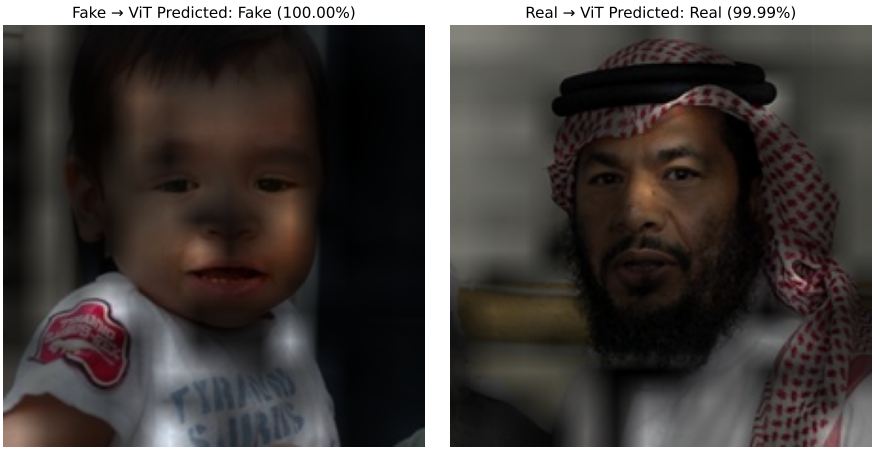

Figure 6.13: The visualization result of the Attention Rollout of the Vision Transformer (ViT) model. The left side represents the deepfake image, and the right side represents the area of interest for the real image. The brighter the area, the greater the contribution of the model to classification, and the darker the area, the lower the importance.

7 Conclusion

7.1 Summary

In this study, three machine learning models, ViT, XceptionNet, and EfficientNet-B0, were applied to deepfake image detection problems to compare and analyze performance and interpretability in multiple ways. The answers to the research questions were derived focusing on the model structure, prediction accuracy, computational efficiency, and visualization analysis results according to the explainable artificial intelligence (XAI) technique.

First, looking at the results of The effect of differences in model structure on accuracy (RQ1), the Vision Transformer (ViT) model recorded the highest classification accuracy. ViT showed better performance than existing CNN-based models (XceptionNet, EfficientNet-B0) by effectively integrating information from various regions in an image through a self-attention-based structure. This means that if sufficient training data and adequate pre-training are provided, the Transformer family model can overcome the existing CNN in complex visual pattern recognition problems such as deepfake image classification. XceptionNet also recorded high accuracy, but did not perform as well as ViT. EfficientNet-B0 showed the relatively lowest accuracy, revealing the limitations of its lightweight structure.

Second, looking at the accuracy and computational efficiency of models (RQ2), ViT achieves the highest accuracy, but instead, it consumes a lot of computation and memory and has a slow learning and inference speed. XceptionNet showed a balanced result for both computational efficiency and accuracy, and EfficientNet-B0 showed the best computational efficiency, but the performance was regrettable. Therefore, in practical applications, rather than selecting a model based solely on accuracy, it can be concluded that computational efficiency must also be considered according to the purpose of use (real-time detection, large-scale distribution, etc.).

Third, analyzing XAI (Explainable Artificial Intelligence) Application Results (RQ3), both Grad-CAM (for Convolution-based models) and Attention Rollout (for ViT) provided the basis for the model's decision making visually. In the Grad-CAM results of XceptionNet and EfficientNet-B0, patterns focusing mainly on key areas such as the eyes, mouth, and nose of the face were observed. ViT's attention rollout was in the form of a black-and-white heat map, showing a pattern of attention distributed widely over the entire image area, and tended to be more sensitive to subtle artificial traces of deepfakes. As a result, it is interpreted that ViT was able to achieve higher accuracy. However, ViT's attention visualization also has a limitation in that its intuition is somewhat inferior to CNN-based Grad-CAM.

The distinction and contribution of this study is that three models with completely different structural principles were directly compared under the same conditions, and two XAI techniques, Grad-CAM and Attention Rollout, were applied in parallel to deal with the visual interpretation comprehensively. In particular, the biggest difference from previous studies is that the differential characteristics from CNN-affiliated models were specifically compared by applying Attention Rollout to ViT.

In terms of comparison with existing methods, XceptionNet and EfficientNet-B0 are CNN-based classification models that have already been verified in previous studies, showing similar performance trends in this study. However, ViT has also demonstrated the

strength of the latest Transformer family in the field of deepfake detection, and in particular, it has been able to interpret the cause of high accuracy in conjunction with XAI visualization.

On the other hand, there are several assumptions and limitations in the analysis of this study. It was also confirmed that the results can vary depending on the size and distribution of the used dataset and the up-to-date deepfake generation technique, and that the performance of ViT can fluctuate very sensitively depending on the quality and quantity of data. For actual environmental application, reflecting various deepfake generation methods, securing additional data, and supplementing interpretation power through a combination of various XAI techniques remain future tasks.

7.1.1 Limitations

There are several limitations to this study. Due to the insufficient size and diversity of the image dataset used, it is difficult to completely verify the generalization ability of the model. In particular, since it did not reflect both the latest deepfake generation techniques and various variations, there may be differences in detection performance in the real environment. Transformer models such as ViT showed high accuracy, but large-scale data and computational resources are essential, so when applied to real systems, there are limitations in terms of cost and efficiency. The data experimented in this study were limited only to static images, so in fact, the deepfake image data used more frequently have not been verified. Since the video data requires additional considerations such as temporal information and consistency between frames, a follow-up study is required to reflect this. Although the prediction basis of the model was visualized through the XAI technique, some of the results were ambiguous in interpretation or could not sufficiently explain the decision-making process of the actual model. And it is regrettable that additional exploration such as various data augmentation, ensembles, and post-processing techniques was insufficient. Since only three representative deep learning models were compared in this study, direct comparative analysis with other state-of-the-art models is needed.

Despite these limitations, this study sought to quantitatively compare model performance with XAI analysis results from various perspectives. I tried to ensure maximum reliability of the results by responding flexibly to data quality issues and technical constraints that occurred during the experiment. In particular, I clearly recognize the limitations of the data and the constraints of the model, and I have also made further experimental plans to supplement them. I plan to re-experiment with data reflecting the latest Deepfake generation method. In addition, the introduction of lightweight models and hardware optimization techniques applicable to real world environments is also considered as a future research project. Additional XAI techniques or Attention Mechanism analysis will also be attempted to strengthen model interpretation. Additionally, I want to improve the stability and generalization performance of the model by experimenting with various methods such as data augmentation, ensemble techniques, and post-processing applications. In the future, I plan to expand the scope of research by expanding it to video data as well as non-static images. Direct performance comparisons with various state-of-the-art deep learning models will also be conducted step by step. Through this follow-up research direction, I intend to get closer to the development of reliable Deepfake detection technology in a real environment.

7.2 Conclusion

In this study, three models, ViT, XceptionNet, and EfficientNet-B0, were used to systematically analyze the deepfake image detection problem. As a result of the experiment, ViT showed the highest classification accuracy and the strength of the transformer based model could be confirmed. XceptionNet showed high accuracy and ease of interpretation, and EfficientNet-B0 was competitive in model weight reduction and computational efficiency. As a result of XAI visualization using Grad-CAM and Attention Rollout, it was found that each model had a difference in the image area that was focused on when discriminating between deepfakes and actual images. In particular, ViT tends to pay wide attention to the entire image area, while XceptionNet and EfficientNet-B0 tended to focus on the core area of the face. The characteristics and limitations of each model were revealed simultaneously from various perspectives such as data diversity and quality, computational resources, and model interpretability. The comparative analysis of this study provides practical implications for several factors to be considered in the design of the actual deepfake detection system. As a result, it was confirmed that it is necessary to select a model that considers accuracy, efficiency, and interpretability in a balanced way depending on the purpose. This study is significant in that it comprehensively compared various models and XAI techniques under the same conditions. This conclusion can serve as a practical guideline for subsequent research and practical application in the field of deepfake detection.

7.3 Dissemination

The main achievements of this study will be spread through various academic channels inside and outside the university. The research results and analysis will be shared at in-school seminars and graduation thesis presentations. I would like to increase the completeness of the study through feedback from the supervisor and colleagues in the laboratory. In addition, the source code and key experimental results used in the study will be released through the GitHub repository. Through this, interested researchers or students can directly check and refer to the code or use it for further research. The body of the paper clearly describes the composition of the dataset, the analysis method, and the process of interpreting the results, increasing the reproducibility and transparency of the study. It will be actively used as academic discussions and follow-up research data in schools and affiliated laboratories. If the research results are further expanded in the future, external academic activities such as presenting papers at academic conferences and academic conferences will also be considered. As such, the achievements of this study are expected to contribute to various researchers through disclosure of GitHub and sharing within the school, focusing on submitting graduation papers.

7.4 Problems Encountered

In the course of the research, I faced a number of unexpected problems. In the training process of the deep learning model, the performance of the hyperparameter changes significantly, and repetitive experiments and detailed tuning were essential. In certain models, learning proceeded unstable or did not come out as accurate as expected. In this situation, various methods have been attempted, such as systematically combining multiple hyperparameters and applying early termination techniques. Even in the explainability analysis (XAI) stage, there was always a problem that the visualization results were not clear or inconsistent. By comparing various XAI techniques, I tried to choose a method that has high interpretability and matches the actual judgment basis. I wanted to test in several

hardware environments to evaluate computational efficiency, but due to the limitations of experimental resources and time, I was able to conduct only some experiments. Lack of memory and delay in execution time in the learning process often hindered research progress. These technical problems were intended to be overcome as much as possible by actively utilizing the Google Collab environment or optimizing the model structure. Although not all problems were completely solved, various attempts and efforts were continued to increase the reliability and completeness of the research.

Experiencing these problems once again reduced the importance of experimental design and data management. In particular, I found that even small errors occurring in the quality and preprocessing of datasets can have a significant impact on model performance. Repeating the experiment, the work of continuously checking and improving data duplication, label errors, and unbalanced sample problems was carried out in parallel. In addition, there have been numerous trials and errors in the process of carefully recording and analyzing performance changes, attempting various hyperparameter combinations. During model training, I have also experienced several phenomena in which learning stops unexpectedly or results diverge. Whenever this happened, the log was analyzed in detail, and the cause was tracked by repeatedly modifying key settings such as model structure or learning rate. During the experiment, the code suddenly stopped working due to external package updates or library compatibility issues. In this process, I felt that the know-how of maintaining consistency in code version management and environmental setting is important. Finally, I focused on recording all processes and problem solving processes in detail and building a reproducible experimental environment. These repetitive trial and error and improvement efforts will increase the reliability of the research results and serve as a great foundation for future research.

7.5 Outlook

In future research, it is essential to build a large-scale dataset reflecting more diverse deepfake generation technologies and the latest variations. It is necessary to develop a model that can analyze temporal information and inter-frame changes, including image data as well as image data that are actually used more often. In order to increase applicability in a real environment, the process of continuously securing and verifying data diversity and quality is also important. In order to increase the reliability and explanatory power of the model, research on new interpretation and visualization methods beyond the existing XAI techniques is also required. For real-time detection and large-scale system distribution, model weight reduction and improvement of computational efficiency are essential, and optimization considering hardware constraints is required. It is expected that detection performance can be maximized through various approaches such as ensemble learning, multimodal data combination, and domain adaptation. Research should be expanded so that the model can effectively respond to new types of deepfakes that are unexpected by simulating actual deepfake attack scenarios. In addition, a transparent service design that can clearly explain the deepfake discrimination results to users is also needed. The social acceptance and safety of deepfake detection technology should also be secured through research in connection with legal and ethical aspects. Finally, it is hoped that these studies can lead to practical solutions that can effectively reduce real social threats.

Future research also needs to focus on building an automated deepfake response system that can actually be utilized in the field. In particular, it is important to develop a customizable solution so that it can flexibly respond to user's various needs and environmental changes. A user-friendly interface and integrated management tool should also be developed so that it can be easily applied in various fields such as companies and public in-

stitutions. It also plans to actively seek ways to improve practical user experiences such as real time notification and visual reporting of fake detection results. It is also an important goal to promote collaboration and technological development with domestic and foreign researchers by disclosing research results and source codes as open sources. In addition, it is necessary to discover and improve various problems occurring in the field through continuous data collection and actual service application. Beyond the technical aspects, educational programs and campaigns can also be conducted in parallel to inform the general public of the risk of deepfake and how to respond. It is also necessary to contribute to effective technology diffusion and policy proposals by strengthening networks with various stakeholders such as industry, academia, and policy authorities. If such research and development are carried out from various angles considering both practicality and social impact, the actual value of deepfake response technology can be greatly increased. Ultimately, it is expected that these efforts will substantially contribute to the realization of a safe and reliable digital society.

Bibliography

[ABB+25] Amerini, Irene, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera et al.: *Deepfake media forensics: Status and future challenges.* Journal of Imaging, 11(3):73, 2025.

[AJ23] Ashok, V and Preetha Theresa Joy: *Deepfake Detection Using XceptionNet.* In *2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, pages 1–5. IEEE, 2023.

[AJP22] Ahmed, Imran, Gwanggil Jeon and Francesco Piccialli: *From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where.* IEEE Transactions on Industrial Informatics, 18(8):5031–5042, 2022.

[AKA+23] Abir, Wahidul Hasan, Faria Rahman Khanam, Kazi Nabiul Alam, Myriam Hadjouni, Hela Elmannai, Sami Bourouis, Rajesh Dey and Mohammad Monirujjaman Khan: *Detecting deepfake images using deep learning techniques and explainable AI methods.* Intelligent Automation & Soft Computing, 35(2):2151–2169, 2023.

[AKK23] Alam, M. N., M. Kaur and M. S. Kabir: *Explainable AI in healthcare: enhancing transparency and trust upon legal and ethical consideration.* International Research Journal of Engineering and Technology, 10(6):1–9, 2023.

[AM23] Afshari, N. and A. Mohammadi: *The Legal Implications of Deepfake Technology: Privacy, Defamation, and the Challenge of Regulating Synthetic Media.* Legal Studies in Digital Age, 2(2):13–23, 2023.

[AMG21] Ajoy, A., C. U. Mahindrakar and D. Gowrish: *DeepFake Detection using a frame based approach involving CNN.* In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1329–1333. IEEE, September 2021.

[ARuH+24] Ali, G., J. Rashid, M. R. ul Hussnain, M. U. Tariq, A. Ghani and D. Kwak: *Beyond the Illusion: Ensemble Learning for Effective Voice Deepfake Detection.* IEEE Access, 12:149940–149959, 2024.

[AS21] Arora, Tanvi and Rituraj Soni: *A review of techniques to detect the GAN-generated fake images.* Generative Adversarial Networks for Image-to-Image Translation, pages 125–159, 2021.

[AZH+21] Alzubaidi, L., J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie and L. Farhan: *Review of deep learning: concepts, CNN architectures, challenges, applications, future directions.* Journal of Big Data, 8:1–74, 2021.

[BCDZ25] Babaei, Reza, Samuel Cheng, Rui Duan and Shangqing Zhao: *Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis.* Journal of Sensor and Actuator Networks, 14(1):17, 2025.

[Boe25] Boediman, Eko Putra: *Exploring the impact of deepfake technology on public trust and media manipulation: A scoping review.* Jurnal Komunikasi, 19(2):131–152, 2025.

[BRS+24] Bozkir, Efe, Clara Riedmiller, Athanassios N Skodras, Gjergji Kasneci and Enkelejda Kasneci: *Can You Tell Real from Fake Face Images? Perception of Computer-Generated Faces by Humans.* ACM Transactions on Applied Perception, 22(2):1–23, 2024.

[BYDA25] Banerjee, Sayan, Sumit Kumar Yadav, Ankit Dhara and Md Ajij: *A Survey: Deepfake and Current Technologies for Solutions.* 2025.

[Chi25] Chinnaraju, A.: *Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability.* World Journal of Advanced Engineering Technology and Sciences, 14(3):170–207, 2025.

[DKT+25] Das, Subhranil, Rashmi Kumari, Utkarsh Tripathi, Shivi Parashar, Adetya Dubey, Akanksha Yadav and Raghwendra Kishore Singh: *Enhanced DeepFake Detection Using CNN and EfficientNet-Based Ensemble Models for Robust Facial Manipulation Analysis.* In *2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT)*, pages 677–681. IEEE, 2025.

[Dur21] Durgadevi, M.: *Generative Adversarial Network (GAN): A general review on different variants of GAN and applications.* In *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, pages 1–8. IEEE, July 2021.

[EM25] Ennab, M. and H. Mcheick: *Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models.* Machine Learning and Knowledge Extraction, 7(1):12, 2025.

[Ess24] Essa, Ehab: *Feature fusion Vision Transformers using MLP-Mixer for enhanced deepfake detection.* Neurocomputing, 598:128128, 2024.

[FN24] Fantozzi, P. and M. Naldi: *The explainability of transformers: Current status and directions.* Computers, 13(4):92, 2024.

[GDS+24] Gupta, S., A. K. Dubey, R. Singh, M. K. Kalra, A. Abraham, V. Kumari, J. R. Laird, M. Al-Maini, N. Gupta, I. Singh and K. Viskovic: *Four transformer-based deep learning classifiers embedded with an attention U-Net-based lung segmenter and layer-wise relevance propagation-based heatmaps for COVID-19 X-ray scans.* Diagnostics, 14(14):1534, 2024.

[GGM+22] Ganguly, Shreyan, Aditya Ganguly, Sk Mohiuddin, Samir Malakar and Ram Sarkar: *ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection.* Expert Systems with Applications, 210:118423, 2022.

[GSA+24] Ghous, U., M. S. Sarfraz, M. Ahmad, C. Li and D. Hong: *EXNet: (2+1)D extreme xception net for hyperspectral image classification.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17:5159–5172, 2024.

[GT24] Gowrisankar, Balachandar and Vrizlynn L.L. Thing: *An adversarial attack approach for eXplainable AI evaluation on deepfake detection models.* Computers & Security, 139:103684, 2024.

[GYV+24] Gambín, Ángel Fernández, Anis Yazidi, Athanasios Vasilakos, Hårek Haugerud and Youcef Djenouri: *Deepfakes: Current and future trends.* Artificial Intelligence Review, 57(3):64, 2024.

[HB21] Hancock, Jeffrey T and Jeremy N Bailenson: *The social impact of deepfakes*, 2021.

[HGL+23] He, K., C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang and D. Shen: *Transformers in medical image analysis.* Intelligent Medicine, 3(1):59–78, 2023.

[HJC+23] Huo, Y., K. Jin, J. Cai, H. Xiong and J. Pang: *Vision transformer (ViT)-based applications in image classification.* In *2023 IEEE 9th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pages 135–140. IEEE, May 2023.

[HJMK24] Hosain, Md Tanzib, Jamin Rahman Jim, MF Mridha and Md Mohsin Kabir: *Explainable AI approaches in deep learning: Advancements, applications and challenges.* Computers and electrical engineering, 117:109246, 2024.

[HJNDU24] Heidari, Arash, Nima Jafari Navimipour, Hasan Dag and Mehmet Unal: *Deepfake detection using deep learning methods: A systematic and comprehensive review.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(2):e1520, 2024.

[HL22] Huang, M. L. and Y. C. Liao: *A lightweight CNN-based network on COVID-19 detection using X-ray and CT images.* Computers in Biology and Medicine, 146:105604, 2022.

[HS23] Hussain, T. and H. Shouno: *Explainable deep learning approach for multi-class brain magnetic resonance imaging tumor classification and localization using gradient-weighted class activation mapping.* Information, 14(12):642, 2023.

[HSC22] Hall, Stuart W, Amin Sakzad and Kim-Kwang Raymond Choo: *Explainable artificial intelligence for digital forensics.* Wiley Interdisciplinary Reviews: Forensic Science, 4(2):e1434, 2022.

[HZMK+22] Hasanpour Zaryabi, Elham, Laleh Moradi, Bahram Kalantar, Naoki Ueda and Agus Arip Halin: *Unboxing the black box of attention mechanisms in remote sensing big data using XAI.* Remote Sensing, 14(24):6254, 2022.

[IHB+24] Islam, Masabah Bint E., Muhammad Haseeb, Hina Batool, Nasir Ahtasham and Zia Muhammad: *AI Threats to Politics, Elections, and Democracy: A Blockchain-Based Deepfake Authenticity Verification Framework.* Blockchains, 2(4):458–481, 2024.

[JNT+22] Jahmunah, Vicneswary, Eddie YK Ng, Ru-San Tan, Shu Lih Oh and U Rajendra Acharya: *Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals.* Computers in Biology and Medicine, 146:105550, 2022.

[JWZ24] Jiang, X., S. Wang and Y. Zhang: *Vision transformer promotes cancer diagnosis: A comprehensive review.* Expert Systems with Applications, 252:124113, 2024.

[JZDL24] Javed, M., Z. Zhang, F. H. Dahri and A. A. Laghari: *Real-time deepfake video detection using eye movement analysis with a hybrid deep learning approach.* Electronics, 13(15):2947, 2024.

[KBMB24] Kumar, T., R. Brennan, A. Mileo and M. Bendechache: *Image data augmentation approaches: A comprehensive survey and future directions.* IEEE Access, 2024.

[KCP+24] Koritala, Sai Pragna, Mahitha Chimata, Sai Naren Polavarapu, Bhavya Sri Vangapandu, Tarun Krishna Gogineni and VM Manikandan: *A Deepfake detection technique using Recurrent Neural Network and EfficientNet.* In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2024.

[KDN23] Khan, S. A. and D. T. Dang-Nguyen: *Deepfake Detection: Analyzing Model Generalization Across Architectures, Datasets, and Pre-Training Paradigms.* IEEE Access, 12:1880–1908, 2023.

[Kha24] Kharvi, Prakash L.: *Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media* . IEEE Security & Privacy, 22(04):115–122, July 2024.

[KKK24] Kaushal, Anukriti, Sanjay Kumar and Rajeev Kumar: *A review on deepfake generation and detection: bibliometric analysis.* Multimedia Tools and Applications, pages 1–41, 2024.

[KLI+25] Khan, Abdullah Ayub, Asif Ali Laghari, Syed Azeem Inam, Sajid Ullah, Muhammad Shahzad and Darakhshan Syed: *A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions.* Discover Computing, 28(1):48, 2025.

[KPL22] Khoo, Brandon, Raphaël C-W Phan and Chern-Hong Lim: *Deepfake attribution: On the source identification of artificially generated images.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(3):e1438, 2022.

[KRS+23] Khan, Asifullah, Zunaira Rauf, Anabia Sohail, Abdul Rehman Khan, Hifsa Asif, Aqsa Asif and Umair Farooq: *A survey of the vision*

transformers and their CNN-transformer based variants. Artificial Intelligence Review, 56(Suppl 3):2917–2970, 2023.

[LAP+11] Lukin, Vladimir, Sergey Abramov, Nikolay Ponomarenko, Karen Egiazarian and Jaakko Astola: *Image filtering: potential efficiency and current problems.* In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1433–1436. IEEE, 2011.

[LHLL23] Lin, Hao, Wenmin Huang, Weiqi Luo and Wei Lu: *DeepFake detection with multi-scale convolution and vision transformer.* Digital Signal Processing, 134:103895, 2023.

[Lim21] Lim, Shaun: *Judicial decision-making and explainable artificial intelligence: a reckoning from first principles.* Singapore Academy of Law Journal, 33:280–314, 2021.

[LLS+23] Li, Sinan, Tianfu Li, Chuang Sun, Ruqiang Yan and Xuefeng Chen: *Multilayer Grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis.* Journal of manufacturing systems, 69:20–30, 2023.

[LSM21] Liu, K., R. Shuai and L. Ma: *Cells image generation method based on VAE-SGAN.* Procedia Computer Science, 183:589–595, 2021.

[Lu24] Lu, Yuhang: *Enhancing Explainability and Performance in AI-Driven Face Recognition and Deepfake Detection.* PhD thesis, EPFL, 2024.

[LWD+23] Lin, X., S. Wang, J. Deng, Y. Fu, X. Bai, X. Chen, X. Qu and W. Tang: *Image manipulation detection by multiple tampering traces and edge artifact enhancement.* Pattern Recognition, 133:109026, 2023.

[LZG+25] Li, S., S. Zhao, P. Gao, S. Ge, R. Huang and L. Shu: *HP-MobileNetV3: a lightweight neural network for non-intrusive load monitoring.* Engineering Research Express, 7(1):015223, 2025.

[MAKM22] Mukhlif, A. A., B. Al-Khateeb and M. A. Mohammed: *An extensive review of state-of-the-art transfer learning techniques used in medical imaging: Open issues and challenges.* Journal of Intelligent Systems, 31(1):1085–1111, 2022.

[MPK20] Malolan, Badhrinarayan, Ankit Parekh and Faruk Kazi: *Explainable deep-fake detection using visual interpretability methods.* In *2020 3rd International conference on Information and Computer Technologies (ICICT)*, pages 289–293. IEEE, 2020.

[MWL+22] Minh, Dinh, Hui Xin Wang, Yu Feng Li et al.: *Explainable artificial intelligence: a comprehensive review.* Artificial Intelligence Review, 55:3503–3568, 2022.

[Mya24] Myakala, P. K.: *Beyond Accuracy: A Multi-faceted Evaluation Framework for Real-World AI Agents.* Available at SSRN 5089870, 2024.

[NM16] Nugrahaeni, R. A. and K. Mutijarsa: *Comparative analysis of machine learning KNN, SVM, and random forests algorithm for facial expression classification.* In *2016 International Seminar on Application for Technology of*

Information and Communication (ISemantic), pages 163–168. IEEE, August 2016.

[NNN+22] Nguyen, Thanh Thi, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham and Cuong M Nguyen: *Deep learning for deepfakes creation and detection: A survey.* Computer Vision and Image Understanding, 223:103525, 2022.

[QCNS22] Qi, Z., W. Chen, R. A. Naqvi and K. Siddique: *Designing deep learning hardware accelerator and efficiency evaluation.* Computational Intelligence and Neuroscience, 2022(1):1291103, 2022.

[RMU24] Ramadhani, Kurniawan Nur, Rinaldi Munir and Nugraha Priya Utama: *Improving video vision transformer for deepfake video detection using facial landmark, depthwise separable convolution and self attention.* IEEE Access, 12:8932–8939, 2024.

[ŞAO25] Şahin, E., N. N. Arslan and D. Özdemir: *Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning.* Neural Computing and Applications, 37(2):859–965, 2025.

[SCMS23] Stassin, S., V. Corduant, S. A. Mahmoudi and X. Siebert: *Explainability and evaluation of vision transformers: An in-depth experimental study.* Electronics, 13(1):175, 2023.

[SD23] Singh, Pawan and Dr Bharat Dhiman: *Exploding AI-Generated Deepfakes and misinformation: A threat to global concern in the 21st century.* Available at SSRN 4651093, 2023.

[SICM22] Shuvo, M. M. H., S. K. Islam, J. Cheng and B. I. Morshed: *Efficient acceleration of deep learning inference on resource-constrained edge devices: A review.* Proceedings of the IEEE, 111(1):42–91, 2022.

[SSAT24] Sarker, Md. Abdullah Al, Bharanidharan Shanmugam, Shoriful Azam and Suresh Thennadil: *Enhancing smart grid load forecasting: An attention-based deep learning model integrated with federated learning and XAI for security and interpretability.* Intelligent Systems with Applications, 23:200422, 2024.

[SSTE+24] Soudy, Ahmed Hatem, Omnia Sayed, Hala Tag-Elser, Rewaa Ragab, Sohaila Mohsen, Tarek Mostafa, Amr A Abohany and Salwa O Slim: *Deepfake detection using convolutional vision transformers and convolutional neural networks.* Neural Computing and Applications, 36(31):19759–19775, 2024.

[SW22] Solaiyappan, Siddharth and Yuxin Wen: *Machine learning based medical image deepfake detection: A comparative study.* Machine Learning with Applications, 8:100298, 2022.

[SYG+23] Saxena, Akash, Dharmendra Yadav, Manish Gupta, Sunil Phulre, Tripti Arjariya, Varshali Jaiswal and Rakesh Kumar Bhujade: *Detecting Deepfakes: A Novel Framework Employing XceptionNet-Based Convolutional Neural Networks.* Traitement du Signal, 40(3), 2023.

[SYK+22] Saleem, R., B. Yuan, F. Kurugollu, A. Anjum and L. Liu: *Explaining deep neural networks: A survey on the global interpretation methods.* Neurocomputing, 513:165–180, 2022.

[TC22] Taeb, Maryam and Hongmei Chi: *Comparison of deepfake detection techniques through deep learning.* Journal of Cybersecurity and Privacy, 2(1):89–106, 2022.

[Tiw25] Tiwari, Sudhakar: *The Impact of Deepfake Technology on Cybersecurity: Threats and Mitigation Strategies for Digital Trust.* Available at SSRN 5259359, 2025.

[TLN+22] To, Tuan-An, Hoang-Chau Luong, Nham-Tan Nguyen, Trong-Tin Nguyen, Minh-Triet Tran and Trong-Le Do: *Deepfake detection using efficientnet: Working towards dense sampling and frames selection.* In *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 612–617. IEEE, 2022.

[VE21] Von Eschenbach, W. J.: *Transparency and the black box problem: Why we do not trust AI.* Philosophy & Technology, 34(4):1607–1622, 2021.

[VGD+24] Vashishtha, Srishti, Harshit Gaur, Uttirna Das, Sreejan Sourav, Eshanika Bhattacharjee and Tarun Kumar: *Optifake: Optical flow extraction for deepfake detection using ensemble learning technique.* Multimedia Tools and Applications, 83(32):77509–77527, 2024.

[WHHJ22] Wanner, Jonas, Lukas-Valentin Herm, Kai Heinrich and Christian Janiesch: *The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study.* Electronic Markets, 32(4):2079–2102, 2022.

[WKKG21] Walia, S., K. Kumar, M. Kumar and X. Z. Gao: *Fusion of handcrafted and deep features for forgery detection in digital images.* IEEE Access, 9:99742–99755, 2021.

[WMMD24] Wazid, Mohammad, Amit Kumar Mishra, Noor Mohd and Ashok Kumar Das: *A secure deepfake mitigation framework: Architecture, issues, challenges, and societal impact.* Cyber Security and Applications, 2:100040, 2024.

[WYA+24] Wang, Yuxuan, Zhan Yang, Iman Azimi, Amir M. Rahmani and Pasi Liljeberg: *Attention-Based Explainable AI for Wearable Multivariate Data: A Case Study on Affect Status Prediction.* In *2024 IEEE 20th International Conference on Body Sensor Networks (BSN)*, pages 1–4. IEEE, October 2024.

[YHEG+23] Yasser, Basma, Jumana Hani, Salma El-Gayar, Omar Amgad, Nourhan Ahmed, Hala M Ebied, Habiba Amr and Mohamed Salah: *Deepfake Detection Using EfficientNet and XceptionNet.* In *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 598–603. IEEE, 2023.

[YZF23] Yuan, F., Z. Zhang and Z. Fang: *An effective CNN and Transformer complementary network for medical image segmentation.* Pattern Recognition, 136:109228, 2023.

[YZJ+22] Yang, H., X. Zhao, T. Jiang, J. Zhang, P. Zhao, A. Chen, M. Grzegorzek, S. Qi, Y. Teng and C. Li: *Comparative study for patch-level and pixel-level segmentation of deep learning methods on transparent images of environmental microorganisms: from convolutional neural networks to visual transformers.* Applied Sciences, 12(18):9321, 2022.

[Zha22] Zhang, T.: *Deepfake generation and detection, a survey.* Multimedia Tools and Applications, 81:6259–6276, 2022.

[ZO23] Zhang, H. and K. Ogasawara: *Grad-CAM-based explainable artificial intelligence related to medical text processing.* Bioengineering, 10(9):1070, 2023.

[ZWZ+24] Zhao, X., L. Wang, Y. Zhang, X. Han, M. Deveci and M. Parmar: *A review of convolutional neural networks in computer vision.* Artificial Intelligence Review, 57(4):99, 2024.

[ZZL+24] Zhang, Li, Dezong Zhao, Chee Peng Lim, Houshyar Asadi, Haoqian Huang, Yonghong Yu and Rong Gao: *Video Deepfake Classification Using Particle Swarm Optimization-based Evolving Ensemble Models.* Knowledge-Based Systems, 289:111461, 2024. Published online 8 April 2024.

[ZZS+22] Zhang, Haowen, Shengyuan Zhao, Yifei Song, Shishuai Ge, Dazhong Liu, Xianming Yang and Kongming Wu: *A deep learning and Grad-Cam-based approach for accurate identification of the fall armyworm (Spodoptera frugiperda) in maize fields.* Computers and Electronics in Agriculture, 202:107440, 2022.