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

Yujin Jeon
Software Engineering M.Sc.
University of Europe for Applied Sciences
Potsdam, 14469, Germany
yujin.jeon.developer@gmail.com

***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 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. 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. 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. 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.**

***Index Terms*—Deepfake, Deep Learning, Vision Transformers (ViT), Explainable AI (XAI), Feature Extraction, Convolutional Neural Networks (CNN)**

I. Introduction

Due to the development of deepfake technology, it has become difficult to distinguish manipulated images from reality, causing social problems such as personal information infringement and dissemination of false information. In particular, it has emerged as a serious problem that threatens social trust and public safety because of the increased likelihood of being abused to synthesize celebrity faces or distort politically and socially sensitive issues. The improvement of deepfake has reached a level that is difficult to respond with existing passive verification methods. Accordingly, the need for quick and accurate automatic detection technology is increasing. Deepfake detection technology can play a key role in maintaining the reliability of media content and preventing the spread of false information. In particular, the introduction of an automatic detection system is essential because deepfake content can rapidly spread on social media platforms and cause enormous damage. Financial institutions and law enforcement agencies can also actively use deepfake detection technology to maintain security and reliability. It also plays an important role in maintaining the soundness of democracy by preventing the spread of false images in the political election process. Recently, academia and industry are actively conducting research using various deep learning-based detection models and explainable artificial intelligence (XAI) techniques to cope with the development of deepfake technology. Therefore, this study focuses on evaluating and improving deepfake detection performance by applying the latest deep learning models and XAI techniques.

Until now, research has mainly focused on deepfake detection using CNN based models, and comparative analysis with state-of-the-art structures such as Vision Transformer (ViT) has been limited. In particular, studies that systematically analyze and evaluate differences in performance between different architectures are rare. In addition, the use of XAI techniques that explain which visual basis the prediction results of the model were based is still in its infancy. Most existing studies rely only on accuracy indicators, and the basis for determining the model has not been sufficiently analyzed. As a result, interpretability of whether the model's prediction is reliable is low, and there are limitations in its application in practical applications. This study performs comparative analysis by selecting three different structures of models (ViT, XceptionNet, and EfficientNet-B0) to fill the gap in these studies. In particular, for each model, XAI based visual analysis is performed by applying Grad-CAM (XceptionNet, EfficientNet-B0) and Attention Rollout (ViT). This allows us to visually explain how each model recognizes and judges deepfake images. In addition, not only the performance of the model, but also the balance between computational efficiency

and visual explanation is included in the evaluation target. Ultimately, this study aims to present the criteria for model selection and design an interpretable and efficient deepfake detection framework for actual system application.

In recent years, deepfake technology has been widely used in the fields of social media, politics, and entertainment, and social concerns about its potential for abuse are also growing. In particular, when combined with a GPT based generative model, text-to-video integration manipulation is possible, further increasing the risk of fact distortion. In this context, it is very urgent to develop a technology to quickly and accurately identify deepfake content. Recent studies have attempted to improve detection performance by utilizing CNN based ResNet, Inception, and EfficientNet. In addition, the Transformer based ViT model is attracting attention with different characteristics from CNN through global attention to the entire image. There is also an increasing number of studies to visually analyze the judgment basis of models through XAI techniques such as Grad-CAM and Attention Rollout. However, most of the existing studies focus on only one or two models, and studies that cover comparisons and interpretability evaluations between various architectures are rare. In particular, a visual explanation of what features the model pays attention to is a key factor in securing reliability in practical applications. Therefore, this study aims to contribute to technological progress and practical applicability in this field by integrating the three different architecture models. It has a high application value in various applications such as real-time detection systems, SNS content filtering, and digital evidence analysis.

II. Literature Review

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 I for further details and assessment of previous studies.

A. 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.* [1] 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.* [2] 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.* [3] 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.* [4] 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.

B. 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.* [4] 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.* [5] 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.* [6] 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 multi-modal detection. Koritala *et al.* [7] 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.

C. Vision Transformer (ViT)

Deepfake detection research based on Vision Transformer (ViT) has recently shown strong performance on various datasets and transform structures. Ganguly *et al.* [2] 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.* [8] 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.* [9] 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.* [10] 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.

D. Explainable AI (XAI)

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.* [11] 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.* [12] 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.* [13] 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.* [14] 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.

III. Our Contribution

A. Gap Analysis

Although the field of deepfake detection is developing rapidly, there are still several unresolved research gaps. First, there are few studies that quantitatively compare the performance between various deep learning models under balanced conditions. In particular, there are extremely limited cases of evaluating the difference in performance between Transformer-based models such as ViT and CNN based models in the same environment. Second, studies comparing interpretability and reliability by applying XAI techniques to different architectures are very scarce. Third, most of the existing studies focus only on quantitative indicators such as accuracy, making it difficult to guarantee the reliability of judgment in real applications. Fourth, an integrated analysis considering the trade-off between computational efficiency and visual explanatory power is also insufficient. Fifth, existing studies tend not to fully consider the possibility of bias according to the composition of the learning dataset. Sixth, there is a lack of consistency analysis between visual explanation results and model prediction in many cases. This gap can be a practical obstacle to model selection and application. This study aims to increase the reliability and applicability of deepfake detection technology through an experimental approach to solve these problems.

B. Research Questions

1) RQ1: How do the architectural differences between ViT, XceptionNet, and EfficientNet-B0 impact their accuracy

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

C. 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.

D. Novelty of this study

This study is differentiated from existing studies in that it deals with the balance between performance, efficiency, and explainability in deepfake detection. In particular, it makes the following research contributions by comparing several architecture based models in an integrated manner and visually analyzing the judgment basis.

- Quantitative comparison of the performance and computational efficiency of ViT, XceptionNet, and EfficientNet-B0 in the same environment.
- Visual analysis of explainability by applying the appropriate XAI technique to each model.
- The reliability of the model judgment is verified by analyzing the consistency between the prediction and visualization results.
- This study present evaluation criteria that consider not only performance but also efficiency and interpretability when selecting deep learning models.

E. 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.

IV. 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 1 for the workflow.)

A. Dataset

In this study, a Kaggle public image dataset for machine learning based fake image detection was used. The dataset consists of two classes, each clearly divided into a real image and a fake image.(Please refer to Fig 2) The entire dataset is designed to minimize learning bias as it is organized in a balanced way. The data were used by dividing into training, validation, and test sets, and each set was divided so that there

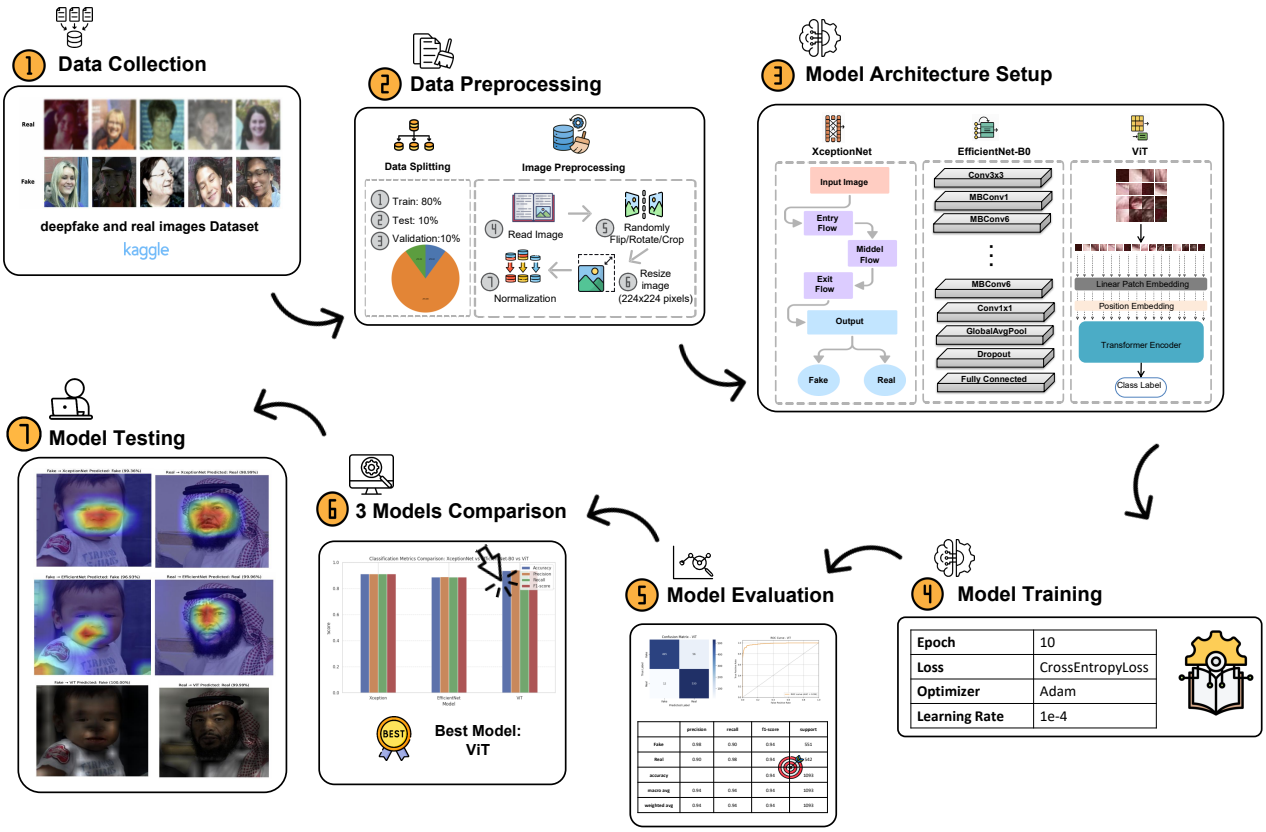

Fig. 1: This picture is a visualization of the entire process of the deepfake image detection experiment performed in this study. After the dataset is structured in a balanced way, it is divided into training, validation, and test sets and goes through a normalization process. After that, learning is performed on three models, ViT, XceptionNet, and EfficientNet-B0, and the performance of each model is evaluated. As an XAI technique, Grad-CAM is applied to the CNN-based model and Attention Rollout is applied to ViT to extract the basis for visual judgment.

was no difference in quantity between classes. 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 3. The training set was focused on improving the performance of the model, and the verification 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.

B. 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 imbalance problem, the class ratio remained the same even within each set. In addition, various transformed images were generated by applying techniques such as horizontal flipping, rotation, and random cropping for data augmentation. This preprocessing process contributed to improving the generalization performance of the model and reducing the possibility of overfitting. The processed data is used as model input for later stages and plays an important role in ensuring the consistency and stability of the entire experiment.

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, show-

Fig. 2: This figure shows some samples of the Kaggle image dataset used in this study. The first row consists of real images and the second row consists of fake images, and visual differences between each class can be compared. The real images have been blurred to protect personal information. The fake image contains distorted information or artificial traces synthesized on the face of a real person, which acts as a major feature that the detection model can learn. The dataset contains images with various races, lighting, facial expressions, and backgrounds, contributing to improving the model's generalization ability. (Dataset source: https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images)

ing 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.

C. 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

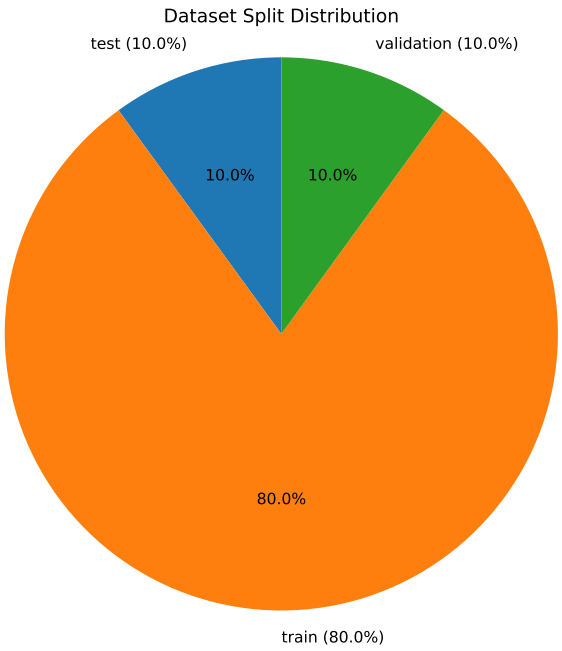

Fig. 3: This figure shows the entire segmentation structure of the Kaggle image dataset used in this study. The data were divided into a ratio of 80% for training, 10% for validation, and 10% for testing. Each segmentation set is configured to maintain balance between classes and provides a basis for quantitatively measuring the performance evaluation and generalization ability of the model.

samples per class) and macro average (simple average between classes), respectively, and both methods are included in the table II.

D. 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 verification and test batch size was set to 32. Each model was trained for a total of 10 epics. 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 III 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 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.(Fig 4)

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. (Fig 5)

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

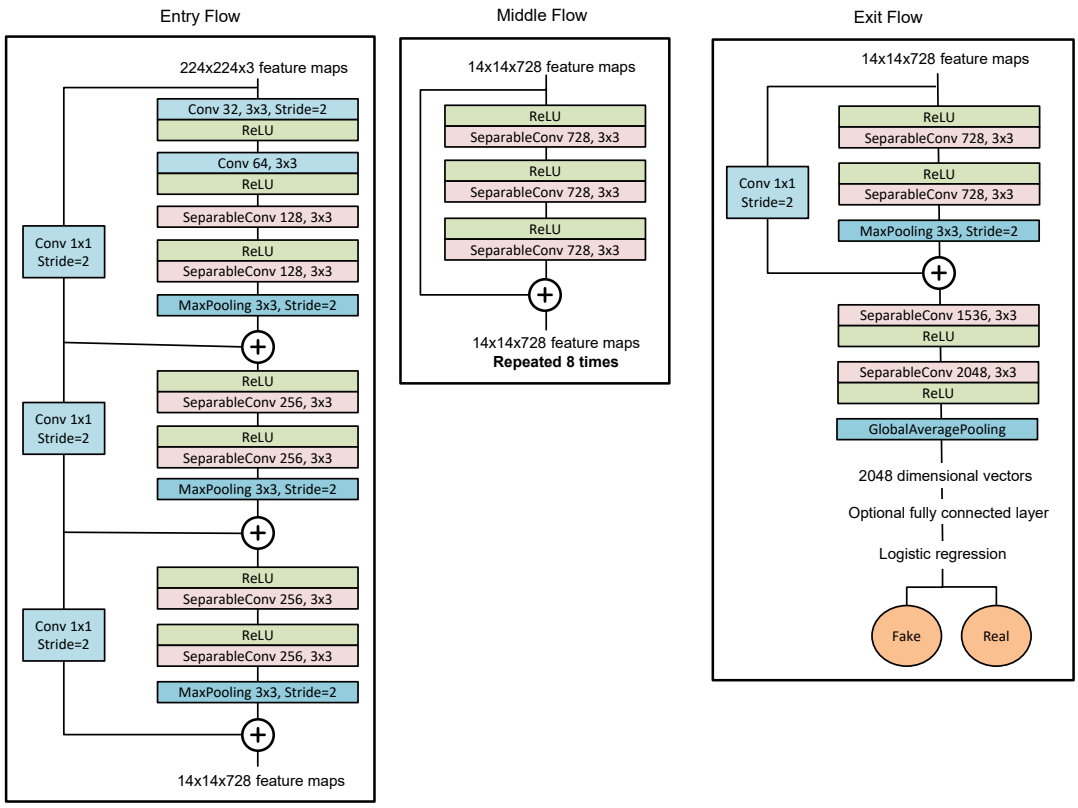

Fig. 4: As a structure of the XceptionNet architecture, we demonstrate the use of depthwise separate convolution across the stage from Entry Flow to Exit Flow and across the network.

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. (Fig 6)

V. Results

Structural differences between the three models (ViT, XceptionNet, and EfficientNet-B0) have shown distinct differences in deepfake image detection performance. Experimental 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 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 7 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.

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

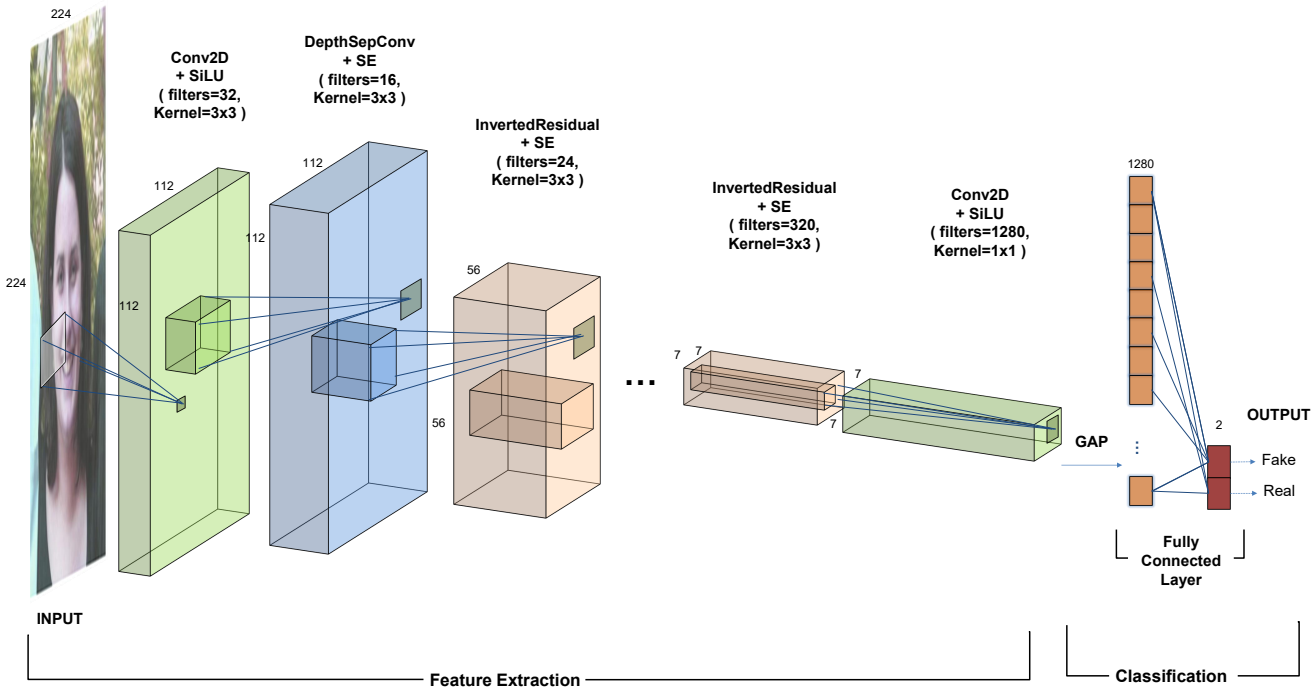

Fig. 5: It is the complete structure of the EfficientNet-B0 architecture, showing the main steps from Stem to MBConv blocks, Head, Pooling, and Classifier, as well as the components of each block.

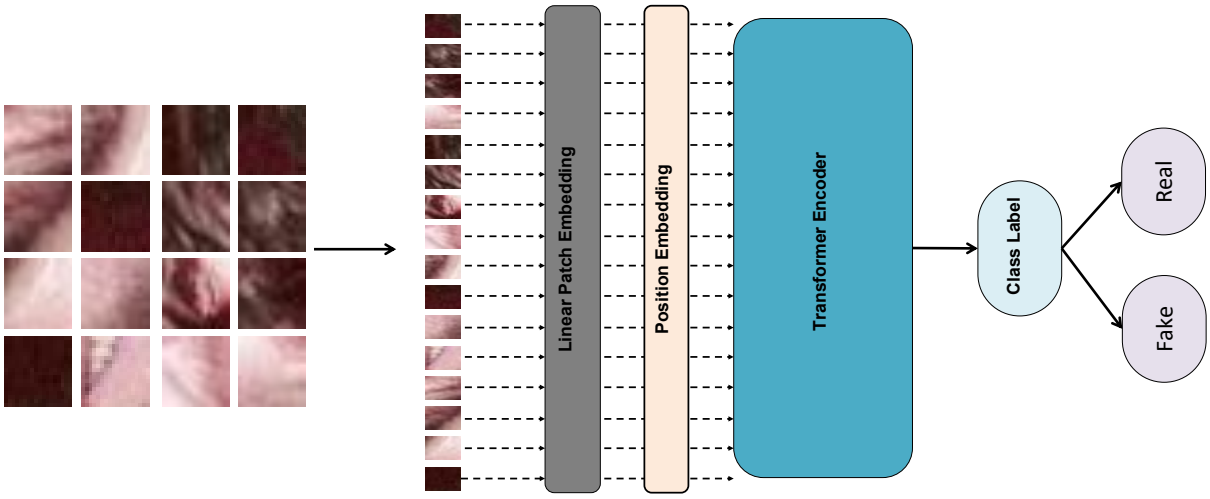

Fig. 6: It is the entire structure of the Vision Transformer (ViT) architecture, which shows the process of segmenting the input image in patch units, processing it with a transformer encoder through each patch embedding and positional encoding.

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 IV

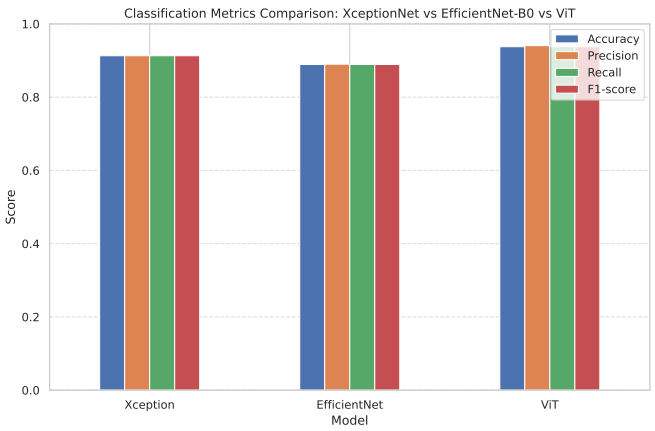

Fig. 7: A bar graph comparing the deepfake image detection performance of three models (XceptionNet, EfficientNet-B0, and ViT.

The feature extraction method of each model was visually analyzed through Grad-CAM (Fig 8, Fig 9) and Attention Rollout (Fig 10) techniques. XceptionNet and EfficientNet-B0 showed high attention to the main facial regions (eyes, nose, mouth, etc.) of images as a result of Grad-CAM visualization. At this time, Grad-CAM visualization is mainly represented by color heatmaps (red, yellow series), allowing intuitive confirmation of modulated areas. These patterns were quite consistent with the real deepfake modulation. On the other hand, ViT's attention rollout visualization results showed a wide distribution over the entire face area, which is mainly expressed in the form of a gray scale heatmap. In particular, ViTs exhibited interest dispersed simultaneously across multiple sites rather than at a single site. All three models showed high attention near the deepfake modulation region, but ViTs tended to incorporate information over a wider region. EfficientNet-B0 and XceptionNet generated clearer activation maps in the strongly modulated region. As a result of XAI visualization, it was found that the internal feature extraction method of each model was distinctly different, and this difference was also evident in the visualization form of the XAI technique (Color or Black and White).

VI. Discussion

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

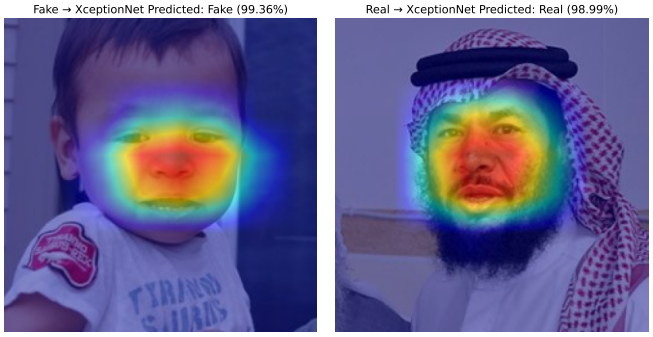

Fig. 8: This figure is Grad-CAM visualization result of XceptionNet model. The left is a deepfake image and the right is a result of a real image. Closer to red in each image means an area that the model considers more important for classification, and closer to blue indicates an area of relatively low importance.

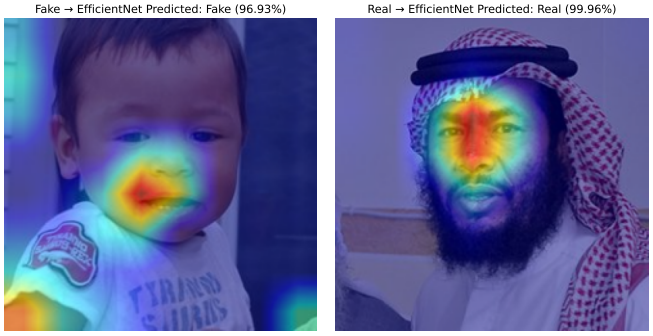

Fig. 9: This figure shows Grad-CAM visualization results of the EfficientNet-B0 model. The left side shows the deepfake image, and the right side shows the area of attention for the real image.

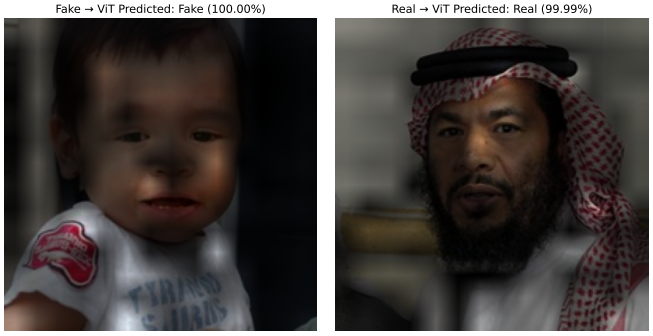

Fig. 10: The visualization result of the Attention Rollout of the Vision Transformer (ViT) model. The left side represents the deepfake image, and the right side represents the area of interest for the real image. The brighter the area, the greater the contribution of the model to classification, and the darker the area, the lower the importance.

(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.

A. 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.

B. Future Directions

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.

VII. 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.

References

[1] A. Saxena, D. Yadav, M. Gupta, S. Phulre, T. Arjariya, V. Jaiswal, and R. K. Bhujade, "Detecting deepfakes: A novel framework employing xceptionnet-based convolutional neural networks." *Traitement du Signal*, vol. 40, no. 3, 2023.

[2] S. Ganguly, A. Ganguly, S. Mohiuddin, S. Malakar, and R. Sarkar, "Vixnet: Vision transformer with xception network for deepfakes based video and image forgery detection," *Expert Systems with Applications*, vol. 210, p. 118423, 2022.

[3] V. Ashok and P. T. Joy, "Deepfake detection using xceptionnet," in *2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*. IEEE, 2023, pp. 1–5.

[4] B. Yasser, J. Hani, S. El-Gayar, O. Amgad, N. Ahmed, H. M. Ebied, H. Amr, and M. Salah, "Deepfake detection using efficientnet and xceptionnet," in *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE, 2023, pp. 598–603.

[5] T.-A. To, H.-C. Luong, N.-T. Nguyen, T.-T. Nguyen, M.-T. Tran, and T.-L. Do, "Deepfake detection using efficientnet: Working towards dense sampling and frames selection," in *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE, 2022, pp. 612–617.

[6] S. Das, R. Kumari, U. Tripathi, S. Parashar, A. Dubey, A. Yadav, and R. K. 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)*. IEEE, 2025, pp. 677–681.

[7] S. P. Koritala, M. Chimata, S. N. Polavarapu, B. S. Vangapandu, T. K. Gogineni, and V. Manikandan, "A deepfake detection technique using recurrent neural network and efficientnet," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2024, pp. 1–6.

[8] H. Lin, W. Huang, W. Luo, and W. Lu, "Deepfake detection with multi-scale convolution and vision transformer," *Digital Signal Processing*, vol. 134, p. 103895, 2023.

[9] E. Essa, "Feature fusion vision transformers using mlp-mixer for enhanced deepfake detection," *Neurocomputing*, vol. 598, p. 128128, 2024.

[10] K. N. Ramadhani, R. Munir, and N. P. Utama, "Improving video vision transformer for deepfake video detection using facial landmark, depth-wise separable convolution and self attention," *IEEE Access*, vol. 12, pp. 8932–8939, 2024.

[11] W. H. Abir, F. R. Khanam, K. N. Alam, M. Hadjouni, H. Elmannai, S. Bourouis, R. Dey, and M. M. Khan, "Detecting deepfake images using deep learning techniques and explainable ai methods," *Intelligent Automation & Soft Computing*, vol. 35, no. 2, pp. 2151–2169, 2023.

[12] V. Jahmunah, E. Y. Ng, R.-S. Tan, S. L. Oh, and U. R. Acharya, "Explainable detection of myocardial infarction using deep learning models with grad-cam technique on ecg signals," *Computers in Biology and Medicine*, vol. 146, p. 105550, 2022.

[13] H. Zhang, S. Zhao, Y. Song, S. Ge, D. Liu, X. Yang, and K. 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*, vol. 202, p. 107440, 2022.

[14] S. Li, T. Li, C. Sun, R. Yan, and X. Chen, "Multilayer grad-cam: An effective tool towards explainable deep neural networks for intelligent fault diagnosis," *Journal of manufacturing systems*, vol. 69, pp. 20–30, 2023.

TABLE I: This table is a literature review table that comparing and organizing 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(s)	Drawback/ Limitations
2023	Saxena *et al.* [1]	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
2022	Ganguly *et al.* [2]	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.* [3]	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.* [4]	Deepfake Detection Using EfficientNet and XceptionNet	FF++, Celeb-DF	EfficientNet-B4, XceptionNet	High acc, evaluated by log loss/AUC	Comparative study of two CNNs	Details/results limited, 2 datasets
2022	To *et al.* [5]	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.* [6]	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
2024	Koritala *et al.* [7]	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.* [8]	DeepFake detection with multi-scale convolution and vision transformer	Celeb-DF, DFDC, FF++, WildDeepfake	Multi-scale CNN, ViT	Outperforms most methods, good generalization	Hybrid multi-scale & global learning	Datasets/results not fully detailed
2024	Essa *et al.* [9]	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.* [10]	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
2023	Abir *et al.* [11]	Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods	Kaggle (Flickr/StyleGAN)	InceptionResNetV2, DenseNet201, LIME	Acc up to 99.87%	CNN/XAI comparison, LIME explainability	Only still images, specific datasets
2022	Jahmunah *et al.* [12]	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.* [13]	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
		Multilayer Grad-CAM: An effective tool towards	Simulated,		Better		

TABLE II: 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.

Model	**Accuracy**	**Precision (Weighted)**	**Recall (Weighted)**	**F1-score (Weighted)**	**Precision (Macro)**	**Recall (Macro)**	**F1-score (Macro)**
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

TABLE III: This table shows the Summary of network learning environments and key hyperparameter settings used in this study.

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

TABLE IV: 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