Report for the Degree of Master of Computer Science

DeepFake Face Detection using Dual Input CNN with Explainable AI (XAI)

Ujjwol Kayastha
(LC0003001469)

Phoenix College of Management
Computer Science and Multimedia Department
Lincoln University, Malaysia

July, 2024

Research Project for the Degree of master of Computer Science

DeepFake Face Detection using Dual Input CNN with Explainable AI (XAI)

Supervised by Prof. Dr. ..., PhD

A report submitted in partial fulfilment of the requirements for the degree of Master of Computer Science

Ujjwol Kayastha
(LC0003001469)

Phoenix College of Management
Computer Science and Multimedia Department
Lincoln University, Malaysia

July, 2024

Declaration

I hereby declare that this research project entitled **DeepFake Face Detection using Dual Input CNN with Explainable AI (XAI)** is based on my original research work. Related works on the topic by other researchers have been duly acknowledged. I owe all the liabilities relating to accuracy and authenticity of the data and any other information included hereunder.

Signature
Name of the Student: Ujjwol Kayastha
Registration Number: LC0003001469
Date:

Recommendation

This is to certify that this report entitled **DeepFake Face Detection using Dual Input CNN with Explainable AI (XAI)** prepared and submitted by **Ujjwol Kayastha**, in partial fulfilment of the requirements of the degree of Master of Computer Science (MCS) awarded by Lincoln University, has been completed under my supervision. I recommend the same for acceptance by Lincoln University.

Signature
Name of the Supervisor: Mr. Santosh Dhungana
Organization: Phoenix College of Management
Date:

Certificate

The report entitled **DeepFake Face Detection using Dual Input CNN with Explainable AI (XAI)** prepared and submitted by **Ujjwol Kayastha** has been examined by us and is accepted for the award of the degree of Master of Computer Science (MCS.) in Postgraduate by Lincoln University.

Name of the external examiner in Bold [Signature] [Date signed]
External examiner

Name of the report supervisor [Signature] [Date signed]
Supervisor

Name of Head of Department or Principal [Signature] [Date signed]

Acknowledgements

First of all, I would like to thank Lincoln University for giving me a chance to prepare the project in partial fulfilment of the requirements of the degree of Master of Computer Science (MCS) awarded by Lincoln University. After many months of hard work and sincere effort from my side, this research has been conducted. I would like to acknowledge the following notable personalities who have contributed their valuable efforts in different ways in creation of this research.

I would like to give special thanks to Mr. Santosh Dhungana, my research project supervisor for his professional guidance and valuable support and for his useful and constructive recommendations on this project. I also owe deep gratitude to all reputed authors whose writings have provided me with the necessary guidance and invaluable materials for the enrichment of my research papers in all possible ways.

My special appreciation goes to my colleague and to all my family members, teachers and friends for their continuous encouragement and help to complete this work directly or indirectly. I wish to thank various people for their contribution to this project: Mr. Bibek Ale Magar and Mr. Dipesh Dulal for their valuable technical support on this project: Mr. Prakash Devkota for their help in collecting data necessary for the project .

Signature
Name of the Student: Ujjwol Kayastha
Registration Number: LC0003001469
Date:

Abstract

Deepfakes represent a significant challenge in today's digital landscape, creating hyper-realistic fake images and videos using advanced AI techniques like Generative Adversarial Networks (GANs). This research aims to develop a robust model for detecting DeepFake images utilizing a Dual Input Convolutional Neural Network (DICNN) combined with Explainable AI (XAI) techniques. By employing LIME (Local Interpretable Model-agnostic Explanations) and opting for better optimizer RMSProp (Root Mean Square Propagation), this study enhances the interpretability and reliability of the detection process. The DICNN model achieved an impressive accuracy rate, demonstrating its efficacy in identifying deepfakes. This research contributes to the ongoing efforts to combat digital misinformation and underscores the importance of explainable AI in building trust and transparency in machine learning models. In literature review, the research highlights the accuracy of DICNN with other state-of-the-art models and the significance of combining deep learning with explainable AI to ensure the reliability of AI systems. This study also highlights the differences in accuracy between two optimizers, RMSProp and Adam.

Keywords: Deepfakes, Convolutional Neural Networks, Explainable AI, SHAP, LIME, Deep Learning, Face Images, Face Detection

Table of Content

Title	Page
Declaration	i
Recommendation	ii
Certificate	iii
Acknowledgements	iv
Abstract	v
Table of Content	vii
List of figures	viii
List of tables	ix
CHAPTER 1	**1**
INTRODUCTION	
1.1 Background	1
1.1.1 Deepfake detection	2
1.1.2 Explainable AI (XAI)	2
1.2 Statement of the problem	3
1.3 Research questions	3
1.4 Research objectives	4
1.5 Significance/rationale of the study	4
1.6 Limitation and scope of the study	4
1.7 Ethical considerations	5
CHAPTER 2	**6**
LITERATURE REVIEW	
2.1 Convolutional Neural Networks (CNNs)	6
2.1.1 Single Input CNN Models	7
2.1.2 Dual Input CNN Models	7
2.2 Explainable AI (XAI)	9
2.2.1 SHAP	9
CHAPTER 3	**11**
METHODOLOGY	
3.1 Dataset Description	11
3.2 Development of Proposed Model	11
3.2.1 Proposed DICNN Model	12
3.3 Data Preprocessing	13

3.4 Training the Model 13
3.5 Evaluation of the Model 13
3.6 Explainable AI (XAI) 14

CHAPTER 4 **15**
RESULTS AND DISCUSSION
4.1 Training Factors A 15
4.1.1 Training and Validation Results 15
4.2 Training Factors B 17
4.2.1 Training and Validation Results 17
4.3 Discussion 19
4.3.1 Performance Validation with XAI 19

CHAPTER 5 **20**
CONCLUSION AND RECOMMENDATION
5.1 Conclusion 20
5.2 Recommendation 20

REFERENCES **21**

List of figures

	Title	Page
Fig. 2.1	Simple representation of CNN	6
Fig. 2.2	Single input CNN architecture	7
Fig. 2.3	Proposed DICNN model	8
Fig. 2.4	Summary details for DICNN model	8
Fig. 2.5	DICNN accuracy	9
Fig. 2.6	SHAP results for (a) fake image and (b) real image	10
Fig. 3.1	Fake and Real faces	12
Fig. 3.2	Model Test with original Dataset with training and validation accuracy and losses	13
Fig. 4.1	Training and Validation Results for Adam Optimizer	17
Fig. 4.2	Training and Validation Results for RMSProp Optimizer	19
Fig. 4.3	LIME Explanation for Fake Image	19
Fig. 4.4	LIME Explanation for Real Image	19

List of table

	Title	Page
Table 3.1	Summary details of proposed DICNN model	12
Table 3.2	Comparison of model with other state-of-art methods	14
Table 4.1	List of materials and tools used for training the model	15
Table 4.2	Epoch-wise Training and Validation Results for Adam Optimizer	16
Table 4.3	List of materials and tools used for training the model	17
Table 4.4	Epoch-wise Training and Validation Metrics for RMSProp Optimizer	18

1 INTRODUCTION

1.1 Background

In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have rapid advancements which has led to development of Large Language Models (LLMs) like GPT-3, BERT, etc. On the other hand, it has also led to proliferation of deepfake (DF) techniques. Deepfake is a technique that uses AI to create fake images, videos, and audio recordings that appear real. Deepfake technology has been used to create fake news, hoaxes, and other forms of misinformation. It has been used to create fake images and videos of celebrities, politicians, and other public figures that can be used to blackmail, defame them or manipulate public opinion and influence elections.. It can also be used to commit fraud or other forms of financial crimes, espionage or other forms of national security threats ot other forms of crimes or unethical behavior. [1] Advancement in Deep Learning models has enabled computer vision [2], Natural Language Processing (NLP) [3], image processing [4], image steganography [5] and smart transportation systems [1] to name a few. Photo editing softwares with inbuilt AI algorithms like Adobe Photoshop, GIMP, etc. have made it easier to create realistic and sophisticated deepfakes even to those who do not have photo editing experience. [6] There are many easily available applications that can swap faces in videos and images like FaceApp, Zao, etc. that can be used to create deepfakes. When (General Adversarial Networks) GANs are used to create deepfakes, the researches done earlier that depended on deciphering the metadata of the images or videos to detect deepfakes are no longer effective along with splicing or copy-move detection techniques. Some of the researches are conducted to detect deepfakes produced by GANs. [7] NVIDIA has developed a deep learning model called StyleGAN that can generate high-quality deepfake images and videos that are almost indistinguishable from real images and videos which has made it possible for malicious actors to create deepfakes that can be used to deceive people and spread misinformation. [8] We have been using biometric authentication systems like face recognition, fingerprint recognition, iris recognition, etc. on daily basis for financial transactions, unlocking smartphones, access management etc. [9] Deepfake technology is evolving rapidly to create deepfakes that even makes us question the authenticity and integrity of information in digital world [10], [11]. StyleGAN offers data-driven simulation relevant for deepfakes creation process optimization [12] which this research aims to explore.

The ease of use and availability of deepfake technology has made it easier to generate hyper-realistic deepfakes. This has lessened the value of human integrity and dignity. Social media platforms like Facebook and Twitter (now X) have started to take steps to detect and remove deepfakes from their platforms. Many cases of deepfakes have been reported: some of the notable incidents include the deepfake video of Barack Obama, Mark Zuckerberg, and Nancy Pelosi. In 2019, a United Kingdom-based energy company was scammed of 243,000

euros by a deepfake audio of the CEO's voice. [13] In a report published back in 2020, more than 85,000 deepfake contents were which was doubled since initial observation in 2018. [14]

1.1.1 Deepfake detection

The term "deepfake" is developed from the technology "deep learning," a form of Artificial Intelligence (AI). Deep Learning (DL) and neural network technologies used to generate fake photos and videos that are difficult to distinguish from real ones are called "deepfake." Deepfake detection is a challenging problem because deepfake images and videos are often very realistic and difficult to distinguish from real images and videos. DF techniques are evolving rapidly and becoming more sophisticated, making it harder to detect deepfakes using traditional methods as it uses Deep Learning (DL) algorithms like Generative Adversarial Networks (GANs). Even though it can be used for good purposes like in the entertainment industry, it can also be used to create fake news, hoaxes, and other forms of misinformation or even security threats and privacy invasion. To maintain the trust and integrity of the digital world, it is important to develop effective deepfake detection techniques to detect and prevent its malicious usage. [15] A plethora of researches have been conducted and published regarding deepfakes classification and detection using Machine Learning (ML) algorithms like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), etc. despite the fact that the technology is relatively new and evolving.

Since the deepfake technology is evolving rapidly, it is important to develop deepfake detection models that can keep up with the evolving deepfake technology and mitigate the threat and impact of deepfakes. Not all algorithms are effective but some has shown promising results. Even though some ML algoriths perfoms well, it cannot be devoid of mistakes. The question arisis regarding trust, safety, accuracy and reliability. European Union has proposed a regulation to regulate the use of deepfake technology to prevent its misuse and protect the privacy, security of individuals and right to explanation under General Data Protection Regulation (GDPR). [16]

1.1.2 Explainable AI (XAI)

ML algorithms are often considered as blackbox models as they are difficult to interpret and explain. This is where Explainable AI (XAI) comes into play. XAI is a set of techniques and methods that can be used to explain and interpret the decisions made by ML models. [17] All of the mentioned work has achieved great accuracy, and DL algorithms have shown excellent performance. But because of the incomprehensible behavior of DL algorithms, there is a lack of liability and trust in the outcomes. Sometimes, this risk of making a wrong decision may outweigh the benefits of precision, speed, and decision-making efficacy. That is why XAI can

help to understand and explain DL and neural networks better. Transparency of results and model improvement justify the adoption of XAIs in the proposed method. This research uses XAI to showcase the image concentration of the sample images, which is novel in detecting deepfake images with high precision. The usage of XAI makes the proposed method very reliable for detecting deepfake.

LIME [18] was one of the first two most significant efforts in the history of XAI. A tool called LIME may identify features from an image or text that are accountable for an ML model's predictions. It is not model-specific. It can be applied to a large range of ML and DL algorithms. By feeding the comparable model inputs and watching how the predictions change, LIME tries to figure out the model's most important features, or the major components that drive any given choice. This method provides easy explanations, such as whether the model's predictions are driven by a specific word in a document or a feature in an image.

1.2 Statement of the problem

Deep Learning (DL) models like Convolutional Neural Networks (CNNs) have been used to detect deepfakes by analyzing the patterns and features in the images and videos to lessen the impact of deepfakes. [19] However, the performance of the CNN models can be improved by using dual input CNN (DICNN) models that can take advantage of the dual input images to improve the detection accuracy of deepfakes. [20] Despite advances in DeepFake detection, current methods face limitations in accuracy and interpretability. Traditional single input CNNs struggle with complex DeepFake patterns, and their decision-making processes remain opaque and lead to false negatives. There is a need for a more robust and explainable approach to detect and understand DeepFakes. Explainable AI (XAI) techniques aim to address the opacity issue by providing insights into how AI models make decisions. Furthermore, while dual input CNN models that can process multiple sources of information such as temporal and spatial features, have demonstrated improved performance. The proposed research aims to address these limitations by developing a dual input CNN model for DeepFake detection and enhancing its interpretability using XAI techniques and better optimizer.

1.3 Research questions

Q1 How effective is a dual input CNN in detecting DeepFake faces compared to traditional single input CNN models with different optimizers?

Q2 How can Explainable AI (XAI) techniques enhance the interpretability and reliability of DeepFake detection models?

1.4 Research objectives

The research aims to explore use of DICNN with XAI to develop DeepFake detection model by leveraging technologies. This research enhances the implementation of DICNN with XAI to improve the interpretability and reliability of DeepFake detection models with different datasets and advanced optimizer. The main objective of the proposed study is to anticipate and understand fraudulent images, and the major contributions are outlined in the points that follow:

O1 A dual branch CNN architecture is proposed to enlarge the view of the network with more prominent performance in auguring the fake faces.

O2 The study explores the blackbox approach of the DICNN model using LIME to construct explanation-driven findings.

1.5 Significance/rationale of the study

The growing challenges of deepfake technology have raised concerns about the authenticity and integrity of digital content. The proposed research aims to develop a dual input CNN model for DeepFake detection and enhance its interpretability using XAI techniques. The research will contribute to the development of more effective and reliable DeepFake detection models that can help mitigate the threat and impact of deepfakes and help improve the trust and integrity of the digital world. The research will also help raise awareness about the importance of developing ethical AI and ML technologies that respect human rights and values. The research will also help inspire future research and innovation in the field of AI and ML and contribute to the development of more advanced and reliable AI and ML technologies that can help address the challenges and opportunities of the digital age.

1.6 Limitation and scope of the study

Limitations

1. **Evolving Techniques**: Deepfake technology is rapidly evolving, and new methods may emerge that could potentially bypass the detection techniques developed in this study. Continuous updates and adaptations are necessary to keep the detection models effective.

2. **Computational Resources**: The proposed dual input CNN and XAI techniques require substantial computational resources for training and evaluation, which may not be accessible to all researchers or practitioners.

3. **Interpretability vs. Accuracy**: While Explainable AI aims to enhance the interpretability of the models, there may be a trade-off between achieving high accuracy and maintaining explainability, which needs careful consideration.

Scope

1. **Dual Input CNN Model**: This study focuses on developing and evaluating a dual input CNN model that utilizes both spatial and temporal features for improved deepfake detection. The scope includes experimenting with different datasets and configurations to optimize performance.

2. **Explainable AI Techniques**: The research integrates Explainable AI techniques, such as SHAP values, to interpret and explain the decision-making process of the CNN model. This will help in understanding how the model differentiates between real and fake content.

3. **Comparative Analysis**: The study includes a comparative analysis of the proposed dual input's result generated with few datasets and different optimizers to evaluate the performance and interpretability of the model.

1.7 Ethical considerations

The research is conducted in accordance with ethical guidelines and principles to ensure the protection of human subjects and data privacy. Use of publicly available datasets and adherence to data protection regulations is done. Informed consent will be obtained for any data collection involving human subjects, and data anonymization techniques are used to protect privacy. The research also ensures transparency and accountability in reporting the results and interpretations to maintain the integrity and trustworthiness of the research findings. The references and citations are properly acknowledged to give credit to the original authors and sources. The research adheres to academic integrity and ethical standards in conducting and reporting the research.

The research adheres to the ethical standards set forth by my academic institution and any relevant legal requirements, especially regarding data usage and privacy.

2 LITERATURE REVIEW

Plethora of researches has been conducted to detect deepfakes using Deep Learning in various fields, especially in detection and recognition. From medical imaging, disease detections and classifications [21] to sentiment analysis [22]. Image recognition has created huge hype in field of DL. Many approaches like LSTM (Long Short-Term Memory), convolutional traces. Some have used GANs (Generative Adversarial Networks) for the deepfakes detection. [23] Targeting specific features sets limitation as they can sometimes miss vital points.

2.1 Convolutional Neural Networks (CNNs)

The CNN (Convolutional Neural Networks) [24] is powerful and efficient field in image classification and recognition. The DL algorithms takes image inputs and weights attributes and assign them to distinguish between real and deepfakse. Due to its less pre-processing and ability to learn characteristics, filters after training and its efficiency, many researches have been carried out. One of the researches [25] proposes an idea to extract face from videos and classify as real or fake using various CNN methods, such as ResNet, Inception and VGG. Their most efficient architecture resulted 90.2% validation accuracy. Researches has been carried out to develop system that can understand visual data which gave birth to Computer Vision. In 2012, AlexNet, an AI model, developed by researchers from University of Toronto, won ImageNet contest with 85% accuracy that was driven by CNNs. [26]

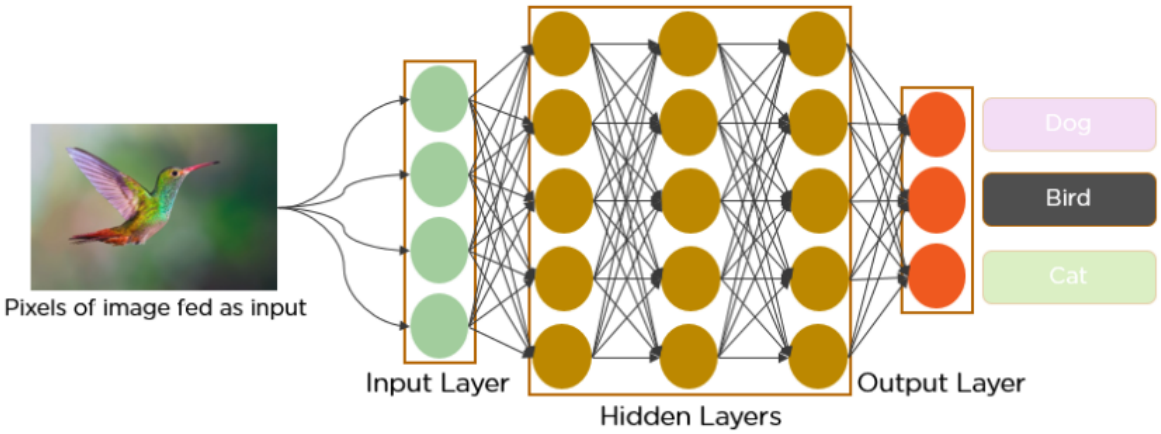

Fig. 2.1: Simple representation of CNN

CNNs are vital in computer vision tasks that includes object detection, image classification and segmentation. Modern CNNs uses Python to leverage advance techniques to extract and learn featurs from images. To train the models effectively, methos like hyperparameters, regularization and optimization techniques are crucial.

2.1.1 Single Input CNN Models

Deepfake detection on images, videos, or other form of media mostly uses traditional single input CNN that extract features from images using convolutional layers, pooling and fully connected layers for image classification.

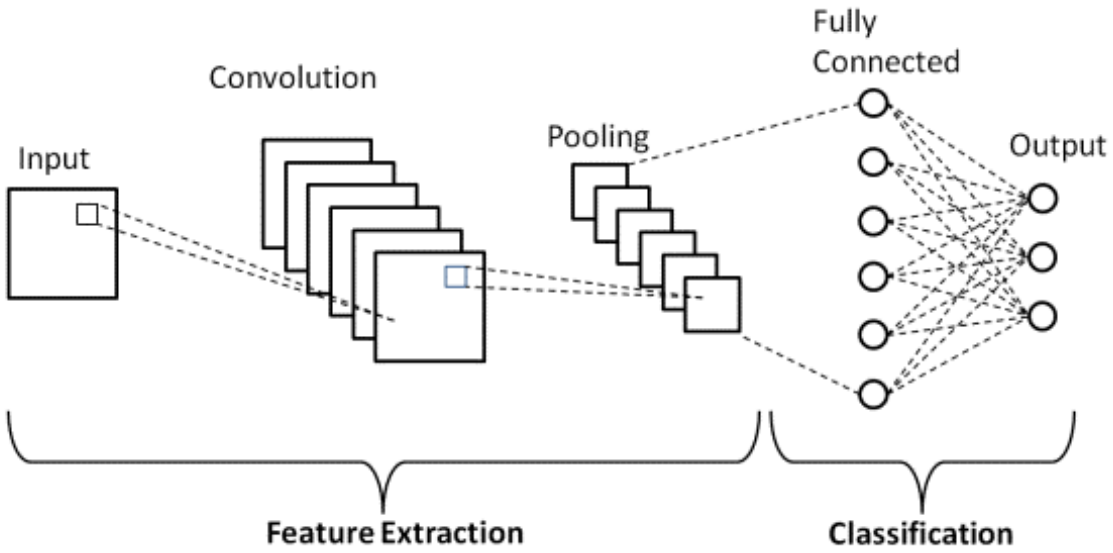

Fig. 2.2: Single input CNN architecture

Fig. 2.2 depicts the main parts of CNN architecture that are stacked to generate output.

2.1.2 Dual Input CNN Models

DICNN (Dual Input Convolutional Neural Network) is based on base model of CNN. From numerous inputs it can update the parameters adaptively and identify deep patterns.

In the base research paper [20] referred, two input layers of size (224 x 224 x 3) were defined. The researcher processed one branch to continue with single convolution layer whose output was then flattened to concatenate flattened results of another branch. Two dense layers and dropout layers were added on top of that. The integration of DICNN-XAI to augur fake face images and SHAP based explanation is depicted in the 2.3. To explore blackbox approach of DICNN, after analysis it was fed to SHAP, an explainable AI. The research was carried out from the dataset of fake and real face images from Kaggle. The dataset was divided into training, validation and testing sets. The model was trained on training set and validated on validation set. The model was tested on testing set and the results were evaluated using accuracy, F1-score, precision and recall. The results were compared with other models and the proposed model achieved better results.

The model was developed in python using Keras and TensorFlow framework and the summary details for proposed DICNN architecture is depicted in 3.1.
Here, the researcher used the model and for evaluation, training accuracy, training loss, validation accuracy, validation loss, test accuracy, test loss, F1-score were used. The DICNN

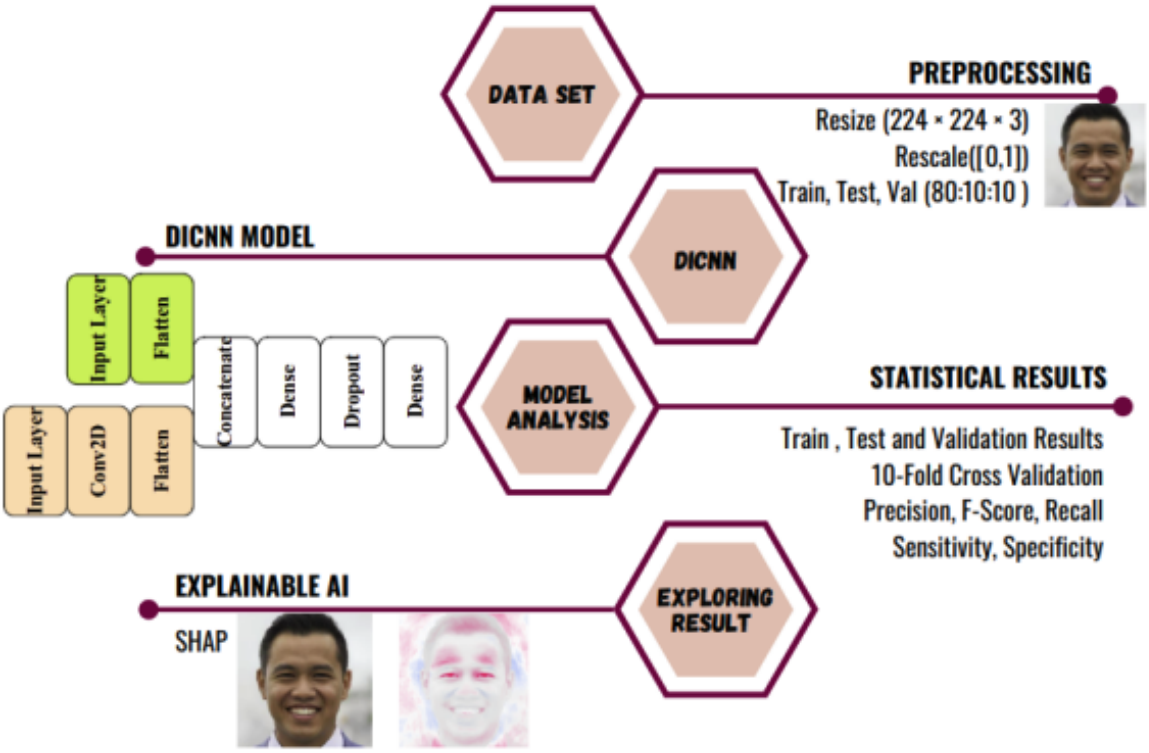

Fig. 2.3: Proposed DICNN model

Layer Name	Shape of Output	Param #	Connected to
Input 1	(None, 224, 224, 3)	0	-
Input 2	(None, 224, 224, 3)	0	-
Conv2D	(None, 222, 222, 32)	896	Input 1
Flatten 1	(None, 150,528)	0	Input 2
Flatten 2	(None, 1,577,088)	0	Conv2D
Concatenate Layer	(None, 1,727,616)	0	[Flatten 1, Flatten 2]
Dense 1	(None, 224)	386,986,208	Concatenate Layer
Dropout	(None, 224)	(None, 224)	Dense 1
Dense 2	(None, 2)	450	Dropout
Total params: 386,987,554 Trainable params: 386,987,554 Non-trainable params: 0			

Fig. 2.4: Summary details for DICNN model

achieved average training accuracy of $99.36 \pm 0.62\%$ and average validation accuracy of $99.30 \pm 0.94\%$ as illustrated in the 2.5.

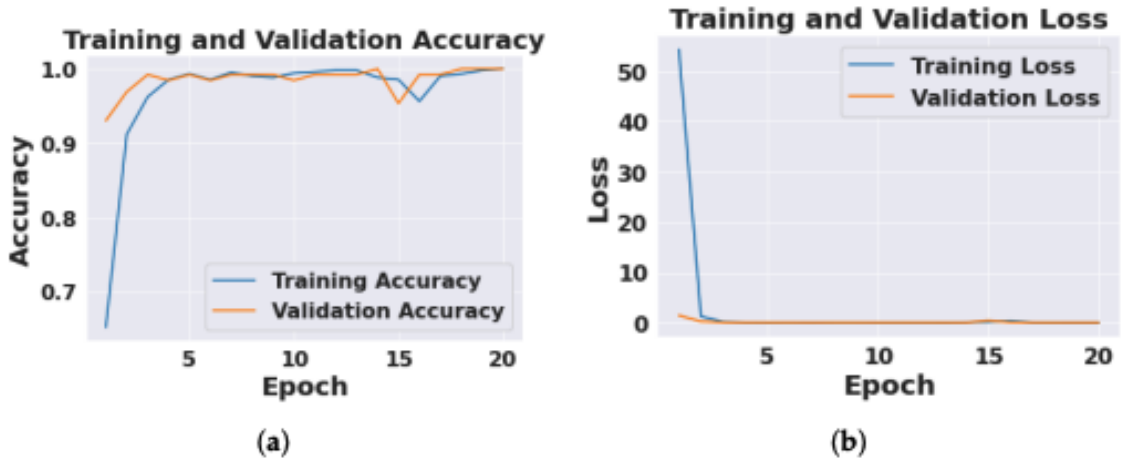

Fig. 2.5: DICNN accuracy

2.2 Explainable AI (XAI)

Deep learning algorithms are blackbox in nautre and it is difficult to understand the decision making process. XAI (Explainable AI) is a field of AI that focuses on making decisions of AI models interpretable. It is crucial to understand the decision making process of AI models. So, to increase the readability and interpretability of AI models, XAI is used, especiallly in image processing [5], computer vision, forensics [27] and criminal investigation [28].

2.2.1 SHAP

SHAP (SHapley Additive exPlanations) is a method that uses game theory to explain the output of any machine learning model. It is used to explain the output of the model by computing the contribution of each feature to the prediction. [29] How each pixel in the image contributes to the prediction is explained by SHAP. The SHAP values are calculated by taking the difference between the prediction of the model with the feature and without the feature. The SHAP values are then used to explain the prediction of the model. Red color pixels contribute to prediction of class while blue pixels make class predictions less likely correct. [30]

Shapley values are computed using Equation 1

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \tag{1}$$

For a particular attribute i, f_x is the switch of results subsumed by values from SHAP. S is the member of all features from feature N, with the deviation of feature i. The weighting

factor $\frac{|S|!(M-|S|-1)!}{M!}$ sums up the numerous ways, and the subset S can be permuted. For the attributes with subset S, the results are denoted by $f_x(S)$ and are a result of Equation 2.

$$f_x(S) = \mathbb{E}[f(x)|x_S] \tag{2}$$

With each original trait replaced, (x_i), SHAP replaces a binary variable (z_i') that represents whether x_i is absent or present as per Equation 3.

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' = \text{bias} + \sum \text{featureContribution} \tag{3}$$

In Equation 3, for model $f(x)$, the confined surrogate model is $g(z')$.

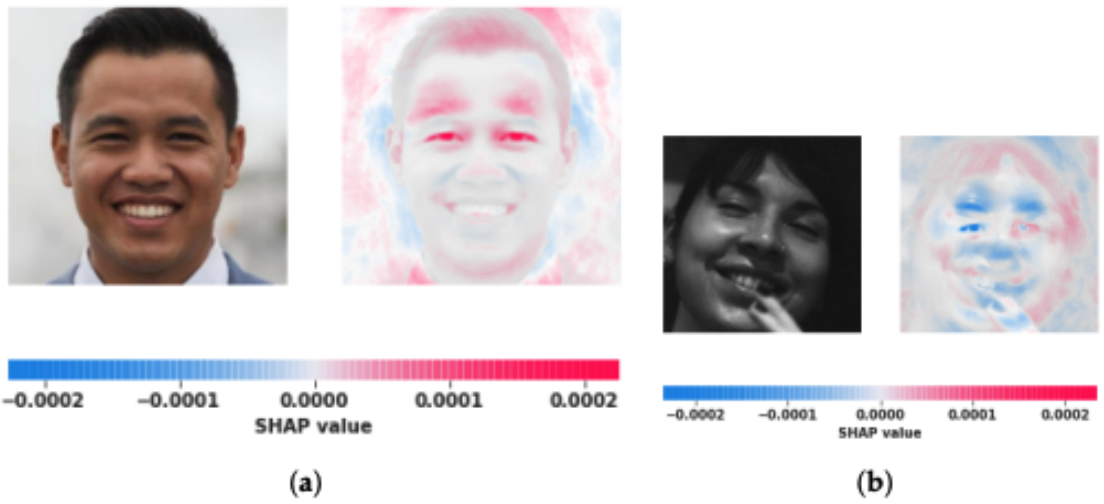

Fig. 2.6: SHAP results for (a) fake image and (b) real image

As showin in Fig. 2.6, SHAP results for fake and real images are depicted. The red pixels intensity on fake image is high, where as on real image, the blue pixels intensity is high. The SHAP values are calculated for each pixel in the image and the results are shown in the image.

3 METHODOLOGY

Deepfake detection has become a crucial part of the digital world. The research has been conducted to detect deepfakes using Deep Learning in various fields, especially in detection and recognition. From medical imaging, disease detections and classifications to sentiment analysis. Image recognition has created huge hype in field of DL. Many approaches like LSTM (Long Short-Term Memory), convolutional traces, GANs (Generative Adversarial Networks) are used for the deepfakes detection. Targeting specific features sets limitation as they can sometimes miss vital points. For this research combined approach is opted. Research began with the primary research by consulting the expertise regarding the topic and research questions. Then alongside it, Literature Reviews, observation played a vital role for the research to take its shape. Since plethora of research works have been carried out in this field, to review the previous research works and their findings to gain broad and better understanding of the subject was vital, it was resourceful and helped in identifying the research gap.

3.1 Dataset Description

The dataset used for this research is from the open-source website Kaggle [31]. The dataset consists of fake and real face images. The dataset is divided into training, validation and testing sets. The model will be trained on training set and validated on validation set. The model will be tested on testing set and the results will be evaluated using accuracy, F1-score, precision and recall. The results will be compared with other models and the proposed model will be evaluated. The dataset consists of 140,000 images of faces. The dataset is divided into 70,000 images of fake faces and 70,000 images of real faces. Fake images were generated using NVIDIA's Style-Based Generator Architecture. Dataset from: https://www.kaggle.com/code/dima806/deepfake-vs-real-faces-detection-vit/input

Figure 3.1 shows the fake and real faces from the dataset.

3.2 Development of Proposed Model

1. **Python using Keras**: As the primary programming language, due to its robust support for data analysis and machine learning libraries.

2. **TensorFlow**: These are the leading machine learning frameworks that are used for developing and training AI models, particularly for their support in building and deploying complex neural networks like LSTM.

The model is develop in Google Colab, a cloud-based platform that is short for Google Colaboratory, is a platform offered free of charge by Google that lets you write and run python code in your browser. In particular, it lets you run Jupyter notebooks without having

Fig. 3.1: Fake and Real faces

to worry about your hardware or the software installed on your computer. [32] Google Colab is a tool that also facilitates access to computing resources and common machine learning libraries.

3.2.1 Proposed DICNN Model

To develop the model with same trainable params as developed in the base research paper [20], it was developed in python using Keras and TensorFlow framework. Images were split into train, test and validation sets after preprocessing. Same layers and architecture are used as in the base research paper to test the validity of the model as documented in the paper. The model is then trained with with better optimizer and different datasets to test it's efficiency.

Layer Name	**Shape of Output**	**Param #**	**Connected to**
Input 1	(None, 224, 224, 3)	0	-
Input 2	(None, 224, 224, 3)	0	-
Conv2D	(None, 222, 222, 32)	896	Input 1
Flatten 1	(None, 150,528)	0	Input 2
Flatten 2	(None, 1,577,088)	0	Conv2D
Concatenate Layer	(None, 1,727,616)	0	[Flatten 1, Flatten 2]
Dense 1	(None, 224)	386,986,208	Concatenate Layer
Dropout	(None, 224)	(None, 224)	Dense 1
Dense 2	(None, 2)	450	Dropout

Table 3.1: Summary details of proposed DICNN model

Fig. 3.2: Model Test with original Dataset with training and validation accuracy and losses

Table reference: 3.1 [20]

Total params: 386,987,554

Trainable params: 386,987,554

Non-trainable params: 0

3.3 Data Preprocessing

To ensure the images were properly processed, they have to be checked by plotting them using Matplotlib. The images were cropped, and the face was centered. The images were rescaled to match the RGB channel.

3.4 Training the Model

Before training the model with the chosen dataset, the model was compiled with Adam optimizer and categorical crossentropy loss function. The model was trained for 10 epochs with batch size of 10. The model was trained with the original dataset used in the base research paper and the results were evaluated 3.2. The result output came as similar as depicted in the research paper. The model was then trained with different datasets to test the efficiency of the model. The model was trained with the training datasets with exact same parameters in base paper with the intention to test the validity of the model. Image size of 224 x 224 x 3 was used for the model to improve computing performance after shuffling. Early stopping callbacks were used to stop the model from overfitting/underfitting. The imaages were rescaled to [0,1].

Since TensorFlow provides different Deep Learning (DL) models, every model has its own unique density and layer count. For comparison purposes, all the hyperparameters were kept the same.

3.5 Evaluation of the Model

On comparing the developed DICNN model with other state-of-art methods, the proposed model achieved good performance as depicted in the table 3.2.

Ref	Category	Method	Dataset	Performance (%)	XAI
[33]	DL	Xception Network	150,000 images	Acc: 83.99%	No
[34]	DL	CNN	60,000 images	Acc: 97.97%	No
[35]	DL	dual-channel CNN	9000 images	Acc: 100%	No
[36]	DL	CNN	321,378 face images	Acc: 92%	No
[37]	DL	Naive classifiers	Faces-HQ	Acc: 100%	No
[38]	DL	VGG	10,000 real and fake image	Acc: 99.9%	No
[38]	DL	ResNet	10,000 real and fake image	Acc: 94.75%	No
[39]	DL	Two Stream CNN	30,000 images	Acc: 88.80%	No
[40]	Physical	Corneal specular highlight	1000 images	Acc: 94%	No
[41]	Human	Visual	400 images	Acc: 50-60%	No
DICNN	DL	DICNN	1289 images	Acc: 99.36 ± 0.62	SHAP

Table 3.2: Comparison of model with other state-of-art methods

The model was validated with the original data used in the base research paper. The model was able to achieve the same level of accuracy as mentioned in the paper. The intention of the model was to gain the same accuracy as depicted in the base research paper, which was achieved. The result in depicted in the Figure. 3.2.

After achieving the desired results, the model was then trained with different datasets to test the efficiency.

3.6 Explainable AI (XAI)

DL models are often considered as black-box models as they are complex and difficult to interpret. To make the model more interpretable, Explainable AI (XAI) is used. SHAP (SHapley Additive exPlanations) is a popular XAI technique. AI predicts and accepts results without explanation, this is where Explainable AI came into play. It gives clarification through analysis, masking, weight of features, maksing, numeric values and visualization. The base model has incorporated SHAP, whereas in this research, the Local Interpretable Model-Agnostic Explanations (LIME) was used to explain the model in addition to SHAP as it can be used for DL models. It uses surrogate models to output the outcome.

4 RESULTS AND DISCUSSION

4.1 Training Factors A

After augumentation of the data, the model was trained for 20 epochs with a train generator and validation generator, resulting in good model accuracy for each model. The training factors have been summarized in the table 4.1.

Training factor	Values
Platform	Google Colaboratory
TPU	Python 3 Google Compute Engine backend
Optimizer	Adam
Loss function	Categorical cross-entropy
Learning rate	0.001
Epoch	20
Batch size	10

Table 4.1: List of materials and tools used for training the model

4.1.1 Training and Validation Results

The table below presents the training and validation loss and accuracy metrics for each epoch during the training process of the dual-input Convolutional Neural Network (CNN) model. The training was conducted over 20 epochs. The performance metrics demonstrate the following key observations:

- **Training Accuracy:** The training accuracy exhibits a high starting point at approximately 90.27% in the first epoch, stabilizing around 85-90% in subsequent epochs. This indicates the model's consistent learning pattern.
- **Validation Accuracy:** The validation accuracy is remarkably high, starting at 98.31% in the first epoch and achieving a perfect accuracy of 100% in the 18th epoch. The consistently high validation accuracy suggests that the model generalizes well to unseen data.
- **Training Loss:** The training loss shows fluctuations throughout the epochs, starting at 0.3247 and reaching 0.1859 by the 20th epoch. This variability could indicate the model's ongoing adjustments during the learning process.
- **Validation Loss:** The validation loss demonstrates significant variability, with values ranging from 0.0036 to 0.2476. The lowest validation loss occurs in the 18th epoch, at 0.0036, corresponding to the highest validation accuracy.

These results underscore the model's high performance and generalization capabilities. The low validation loss in several epochs highlights the model's effectiveness in minimizing error on unseen data, whereas the variability in training loss suggests continuous learning and optimization during training.

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	0.3247	0.9027	0.0354	0.9831
2	0.2831	0.8749	0.0755	0.9746
3	0.3369	0.8315	0.2476	1.0000
4	0.3365	0.8627	0.0219	0.9915
5	0.2225	0.8766	0.2333	0.9153
6	0.3035	0.8401	0.0577	0.9661
7	0.2589	0.8532	0.0219	0.9915
8	0.2451	0.8636	0.0958	0.9492
9	0.1923	0.8871	0.0184	0.9915
10	0.2542	0.8566	0.0231	0.9831
11	0.2476	0.8610	0.0102	0.9915
12	0.2438	0.8471	0.0354	0.9831
13	0.2718	0.8514	0.1603	0.8898
14	0.2670	0.8332	0.0903	0.9576
15	0.2376	0.8784	0.0242	0.9831
16	0.2470	0.8566	0.0181	0.9915
17	0.1699	0.8454	0.0117	0.9915
18	0.1808	0.8766	0.0036	1.0000
19	0.2671	0.8714	0.0322	0.9915
20	0.1859	0.8940	0.0207	0.9915

Table 4.2: Epoch-wise Training and Validation Results for Adam Optimizer

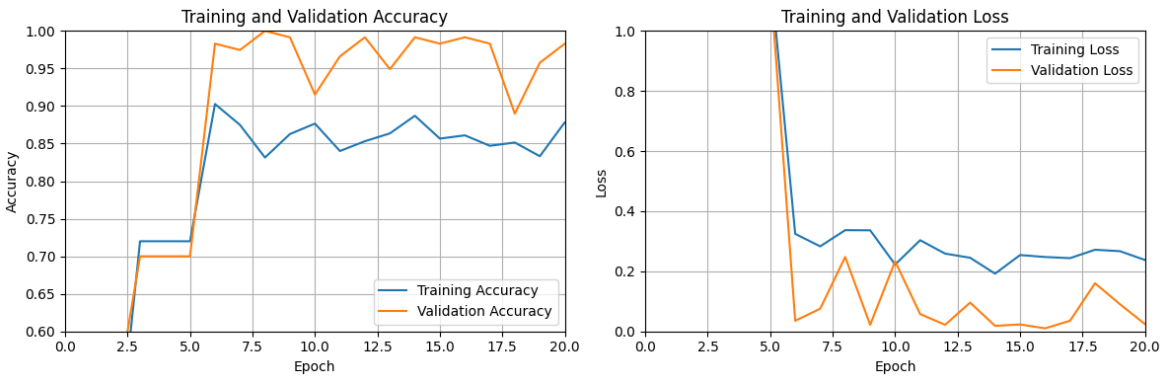

Fig. 4.1: Training and Validation Results for Adam Optimizer

4.2 Training Factors B

The model was compiled with RMSProp optimizer and categorical cross-entropy loss function. RMSProp (Root Mean Square Propagation) is an optimization algorithm that uses the moving average of the squared gradient to normalize the gradient. It is an adaptive learning rate method that divides the learning rate by the moving average of the squared gradient. This helps to adjust the learning rate based on the gradient's magnitude, making it more stable and efficient. The training factors have been summarized in the table 4.3.

Training factor	**Values**
Platform	Google Colaboratory
TPU	Python 3 Google Compute Engine backend
Optimizer	RMSProp
Loss function	Categorical cross-entropy
Learning rate	0.001
Epoch	20
Batch size	10

Table 4.3: List of materials and tools used for training the model

4.2.1 Training and Validation Results

The table below presents the training and validation loss and accuracy metrics for each epoch.

- **Training Accuracy**: The training accuracy starts at approximately 77.76% in the first epoch and shows a general upward trend, reaching 94.70% by the 20th epoch. This indicates that the model is effectively learning and improving its performance.
- **Validation Accuracy**: The validation accuracy starts at a high 94.92% in the first epoch, achieving perfect accuracy (100%) in several epochs, including the second and final epochs. This suggests strong generalization ability to unseen data.

- **Training Loss**: The training loss starts high at 36.7045 in the first epoch but decreases significantly over the epochs, indicating the model's improved ability to minimize error during training.
- **Validation Loss**: The validation loss exhibits fluctuations, starting at 0.3475 in the first epoch, reaching as low as 0.0007 in the second epoch, and ending at 0.0063 in the final epoch. The low validation loss in multiple epochs highlights the model's effectiveness.

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	36.7045	0.7776	0.3475	0.9492
2	2.3207	0.8888	0.0007	1.0000
3	1.0619	0.9079	2.5053	0.8136
4	0.8817	0.9123	0.2394	0.9068
5	0.7055	0.9331	0.1060	0.9746
6	0.9905	0.9209	0.0614	0.9746
7	0.7818	0.9357	0.2718	0.9661
8	0.4616	0.9288	0.0055	1.0000
9	0.3379	0.9348	0.0353	0.9831
10	1.2596	0.8871	0.0474	0.9915
11	0.3767	0.9453	0.0691	0.9915
12	0.3609	0.9409	0.1279	0.9746
13	0.8072	0.9070	0.0652	0.9831
14	0.3556	0.9322	0.0606	0.9746
15	0.6149	0.9253	0.0292	0.9915
16	0.4864	0.9288	0.0336	0.9746
17	0.5084	0.9331	0.0493	0.9661
18	0.3350	0.9513	0.0325	0.9746
19	0.3296	0.9288	0.1307	0.9068
20	0.1756	0.9470	0.0063	1.0000

Table 4.4: Epoch-wise Training and Validation Metrics for RMSProp Optimizer

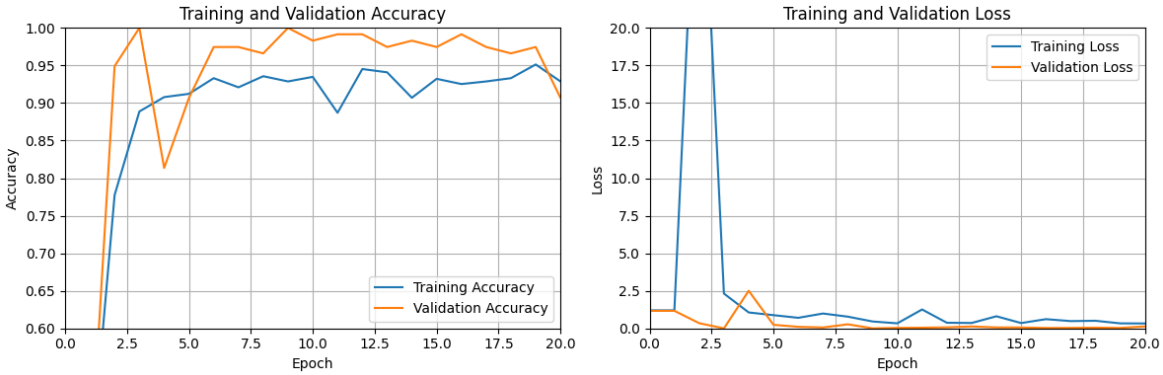

Fig. 4.2: Training and Validation Results for RMSProp Optimizer

4.3 Discussion

4.3.1 Performance Validation with XAI

To analyze the deefake, it is crucial to identify real from fakes and that the model is not biased. SHAP was employed in the base paper, whereas LIME was added to the research to provide more explanation. The Local Interpretable Model-Agnostic Explanations (LIME) algorithm was applied to the proposed model. The LIME algorithm is a popular XAI algorithm that can be used to explain image samples with visual representation. LIME is a model-agnostic algorithm that approximates the local linear behavior of the model. LIME helps us understand why the model made a particular decision by showing which parts of the image were important. The highlighted regions are crucial for the model's prediction. For a real image, these are the areas that convinced the model it is real. For a fake image, these are the areas that convinced the model it is fake. This helps humans understand the model's decision-making process, making it more transparent and trustworthy.

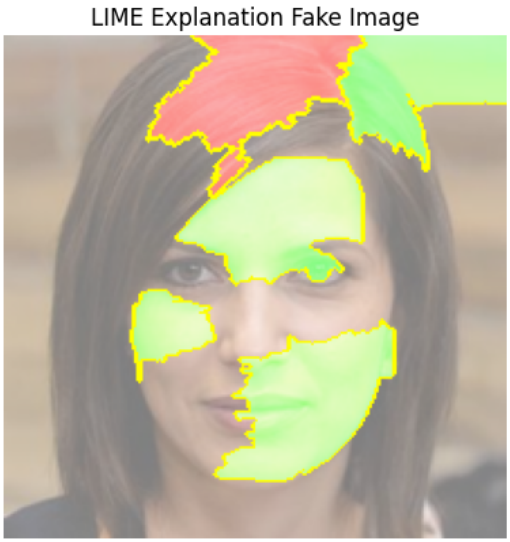

Fig. 4.3: LIME Explanation for Fake Image

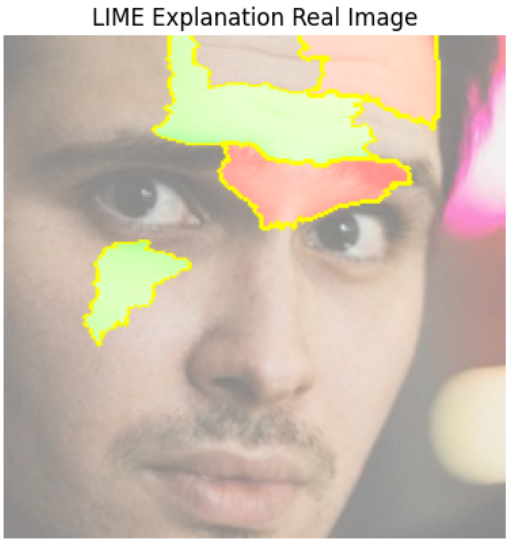

Fig. 4.4: LIME Explanation for Real Image

5 CONCLUSION AND RECOMMENDATION

5.1 Conclusion

The main objective of the project are to detect deepfakes with the help of Dual Input Convolutional Neural Network (DICNN) and to evaluate the model. The research has been conducted to detect deepfakes using Deep Learning in various fields, especially in detection and recognition. The research will shed light on the use of and help in day-to-day life. The use of XAI in deepfake image detection is novel and is exclusively present in this research paper. The proposed system provides 99.17% accuracy in detecting deepfake images from real images. The results indicate that the proposed method is very robust at detecting deepfake images and is also reliable and trustworthy because of the verification and integration from XAI. The SHAP & LIME explainable AI are also used in this research to describe which component of the image from the dataset caused the model to create specific classifications, ensuring the model's validity and reliability. If this system was available in public, it would save hassle of dealing with fake content and fake information. Eventhough, XAI is limited to images in present context, in near future, we can expect it to be used in videos and other forms of media. Moreover, the research demonstrates the value of combining deep learning with explainable AI to ensure the transparency and trustworthiness of AI systems.

5.2 Recommendation

The trust in the developed model and its predictions are based on how well the model is trained and how well the model is explained. Further research can be carried out to improve the model and to make it more efficient and reliable. The findings show that the DICNN-XAI model is very efficient in detecting deepfake images from real images with high accuracy. This enhances the trust in DL systems enabling better understanding of system behavior. In addition to SHAP used in base research paper, LIME was added to the research to provide more explanation to the model with 10x times the dataset. The results were compared and the model was evaluated. This study can be extended to videos and other forms of media to detect deepfakes. It can be used for other XAI algorithms like GradCAM, that can improve auguring problems. Also more diverse data can be used to train the model to make it more robust and reliable. More heterogeneous data can ensure that the model developed is efficient. This current model can be extended to detect deepfakes in videos, which pose a more complex challenge due to the temporal dynamics involved.

REFERENCES

[1] L. Gaur, S. Mallik, and N. Z. Jhanjhi, "Introduction to deepfake technologies," 2022.

[2] M.-H. Guo, T.-X. Xu, J.-J. Liu, *et al.*, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, pp. 331–368, 2022. DOI: 10.1007/s41095-022-0271-y.

[3] T. B. Shahi and C. Sitaula, "Natural language processing for nepali text: A review," *Artificial Intelligence Review*, vol. 55, pp. 3401–3429, 2021. DOI: 10.1007/s10462-021-10093-1.

[4] C. Sitaula and T. B. Shahi, "Monkeypox virus detection using pre-trained deep learning-based approaches," *Journal of Medical Systems*, vol. 46, pp. 1–9, 2022. DOI: 10.1007/s10916-022-01868-2.

[5] M. Bhandari, S. Panday, C. P. Bhatta, and S. P. Panday, "Image steganography approach based ant colony optimization with triangular chaotic map," in *Proceedings of the 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, vol. 2, Gautam Buddha Nagar, India, 2022, pp. 429–434.

[6] D. Wang, Y. Arzhaeva, L. Devnath, *et al.*, "Automated pneumoconiosis detection on chest x-rays using cascaded learning with real and synthetic radiographs," in *Proceedings of the 2020 Digital Image Computing: Techniques and Applications (DICTA)*, Melbourne, Australia, 2020, pp. 1–6.

[7] S. Li, V. Dutta, X. He, and T. Matsumaru, "Deep learning based one-class detection system for fake faces generated by gan network," *Sensors*, vol. 22, no. 20, p. 7767, 2022. DOI: 10.3390/s22207767.

[8] A. D. Wong, *Bladerunner: Rapid countermeasure for synthetic (ai-generated) stylegan faces*, Accessed on 31 October 2022, 2022. arXiv: 2210.06587 [cs.CV]. [Online]. Available: https://doi.org/10.48550/ARXIV.2210.06587.

[9] L. Tran, X. Yin, and X. Liu, *Representation learning by rotating your faces*, 2017. arXiv: 1705.11136 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1705.11136.

[10] H. H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2020, pp. 5781–5790.

[11] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Republic of Korea, 2019, pp. 1–11.

[12] E. Zotov, "Stylegan-based machining digital twin for smart manufacturing," Ph.D. dissertation, University of Sheffield, Sheffield, UK, 2022.

[13] J. Damiani, "A voice deepfake was used to scam a ceo out of $243,000," *Forbes*, 2019, [Online]. Available: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=66e1686a2241.

[14] CyberNews, "Report: Number of deepfakes double every six months," 2021, [Online]. Available: https://cybernews.com/privacy/report-number-of-expert-crafted-video-deepfakes-double-every-six-months.

[15] P. Korshunov and S. Marcel, *Deepfakes: A new threat to face recognition? assessment and detection*, arXiv preprint arXiv:1812.08685, 2019. arXiv: 1812.08685 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1812.08685.

[16] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a 'right to explanation'," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[17] IBM, *Explainable ai*, [Online]. Available: https://www.ibm.com/watson/explainable-ai.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, New York, NY, USA, 2016, pp. 1114–1135. DOI: 10.1145/2939672.2939778.

[19] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020. DOI: 10.1016/j.inffus.2020.07.001.

[20] M. Bhandari, A. Neupane, S. Mallik, L. Gaur, and H. Qin, "Auguring fake face images using dual input convolution neural network," *Journal of Imaging*, vol. 9, no. 1, p. 3, 2023. DOI: 10.3390/jimaging9010003. [Online]. Available: https://doi.org/10.3390/jimaging9010003.

[21] M. H. M. Khan, N. B. Jahangeer, W. Dullull, S. Nathire, X. Gao, *et al.*, "Class classification of breast cancer abnormalities using deep convolutional neural network (cnn)," *PLOS ONE*, vol. 16, no. 8, pp. 1–15, 2021. DOI: 10.1371/journal.pone.0257426.

[22] K. N. Alam, M. S. Khan, A. R. Dhruba, M. M. Khan, J. F. Al-Amri, *et al.*, "Deep learning-based sentiment analysis of covid-19 vaccination responses from twitter data," *Computational and Mathematical Methods in Medicine*, vol. 2021, no. 4321131, pp. 1–15, 2021. DOI: 10.1155/2021/4321131.

[23] L. Guarnera, O. Giudice, and S. Battiato, "Deepfake detection by analyzing convolutional traces," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, 2020, pp. 2841–2850.

[24] S. Saha, *A comprehensive guide to convolutional neural networks—the eli5 way*, https://tinyurl.com/towards-data-science-cnn, [Online]. Available: Accessed on 20 July 2024, 2018.

[25] M. Patel, A. Gupta, S. Tanwar, and M. S. Obaidat, "Trans-df: A transfer learning-based end-to-end deepfake detector," in *Proceedings of the IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2020, pp. 796–801.

[26] A. Vidhya, *Convolutional neural networks (cnn)*, https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/, [Online]. Available: Accessed on 20 July 2024, 2021.

[27] S. W. Hall, A. Sakzad, and K.-K. R. Choo, "Explainable artificial intelligence for digital forensics," *Wiley Interdisciplinary Reviews: Forensic Science*, vol. 4, no. 4, e1434, 2022. DOI: 10.1002/wfs2.1434.

[28] L. B. Winter, "Criminal investigation, technological development, and digital tools: Where are we heading?" In *Investigating and Preventing Crime in the Digital Era*, Berlin, Germany: Springer, 2022, pp. 3–17.

[29] S. M. Lundberg and S.-I. Lee, *Shap (shapley additive explanations) documentation*, Accessed: 20 July 2024, 2021. [Online]. Available: https://shap.readthedocs.io/en/latest/.

[30] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., pp. 4765–4774, 2017.

[31] Kaggle, *140k real and fake faces*, https://www.kaggle.com/xhlulu/140k-real-and-fake-faces, [Online]. Available: Accessed on 20 July 2024.

[32] DataScientest, *Google colab: The power of the cloud for machine learning*, https://datascientest.com/en/google-colab-the-power-of-the-cloud-for-machine-learning, [Online]. Available: Accessed on 20 July 2024, 2021.

[33] Y. Xu, K. Raja, and M. Pedersen, "Supervised contrastive learning for generalizable and explainable deepfakes detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, Waikoloa, HI, USA, 2022, pp. 379–389.

[34] Y. Fu, T. Sun, X. Jiang, K. Xu, and P. He, "Robust gan-face detection based on dual-channel cnn network," in *Proceedings of the 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Suzhou, China, 2019, pp. 1–5. DOI: 10.1109/CISP-BMEI48845.2019.8965937.

[35] F. M. Salman and S. S. Abu-Naser, "Classification of real and fake human faces using deep learning," *International Journal of Academic Engineering Research (IJAER)*, vol. 6, pp. 1–14, 2022.

[36] Y. Zhang, L. Zheng, and V. L. L. Thing, "Automated face swapping and its detection," in *Proceedings of the 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, Singapore, 2017, pp. 15–19. DOI: 10.1109/SIPROCESS.2017.8124533.

[37] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, *Unmasking deepfakes with simple features*, 2019. arXiv: 1911.00686 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1911.00686.

[38] A. Gandhi and S. Jain, "Adversarial perturbations fool deepfake detectors," in *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, 2020, pp. 1–8.

[39] B. Yousaf, M. Usama, W. Sultani, A. Mahmood, and J. Qadir, "Fake visual content detection using two-stream convolutional neural networks," *Neural Computing and Applications*, vol. 34, pp. 7991–8004, 2022. DOI: 10.1007/s00521-022-06817-2.

[40] S. Hu, Y. Li, and S. Lyu, "Exposing gan-generated faces using inconsistent corneal specular highlights," in *Proceedings of the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 2500–2504. DOI: 10.1109/ICASSP39728.2021.9413521.

[41] S. Nightingale, S. Agarwal, E. Härkönen, J. Lehtinen, and H. Farid, "Synthetic faces: How perceptually convincing are they?" *Journal of Vision*, vol. 21, no. 9, p. 2015, 2021. DOI: 10.1167/jov.21.9.2015.