ODD SEMESTER-2023

B. Tech Project-I Report

Framework for classifying authentic human faces and AI-generated human face images

Course Code: MC401

BACHELOR OF TECHNOLOGY

Submitted by:

Vyom Verma (2K21/MC/182)

Rudrakshi Sabharwal (2K21/MC/142)

Yamini (2K21/MC/183)

Under the supervision of

Dr. Aditya Kaushik

DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

We, Vyom Verma (2K21/MC/182), Rudrakshi Sabharwal (2K21/MC/142), Yamini (2K21/MC/183) students of B.Tech. Mathematics and Computing Engineering (MC), hereby declares that the project Dissertation titled "**Framework for classifying authentic human faces and AI-generated human faces**" which is submitted by us to the Department of Applied Mathematics, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship, or other similar title or recognition.

Place: New Delhi

Date: 14 Dec 2024

Vyom Verma **Rudrakshi Sabharwal** **Yamini**

DEPARTMENT OF APPLIED MATHEMATICS
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly, Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

We hereby certify that the Project Dissertation titled "**Framework for classifying authentic human faces and AI-generated human face images**" which is submitted by **Vyom Verma (2K21/MC/182)**, **Rudrakshi Sabharwal (2K21/MC/142)**, **Yamini (2K21/MC/183)**, from the Department of Mathematics and Computing Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the Degree of Bachelor of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: New Delhi
Date: 14 Dec2024

Dr. Aditya Kaushik
SUPERVISOR
(Professor)
Department of Applied Mathematics

DEPARTMENT OF APPLIED MATHEMATICS
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly, Delhi College of Engineering)
Bawana Road, Delhi-110042

ABSTRACT

This project explores the use of deep learning models, specifically Convolutional Neural Networks (CNNs), to differentiate between real human faces and those generated by artificial intelligence (AI). With the increasing prevalence of AI-generated media, such as deepfakes and synthetic images, it has become crucial to develop reliable methods for detecting AI-created content. The primary goal of this project is to design and train a model capable of accurately classifying images as either real or AI-generated.

The dataset used for training consists of 9.6k images, split between 5,000 real human faces and 4,630 AI-generated faces. These images come from various sources, ensuring a diverse representation of human faces and AI-generated content. The model is trained to recognize intricate differences in texture, lighting, and pixel-level details that differentiate real human faces from AI-generated ones. The system undergoes extensive testing and evaluation to determine its accuracy and robustness.

The preprocessing pipeline involved resizing, normalization, and data augmentation to ensure consistency and robustness. The trained model demonstrated a high level of accuracy in identifying subtle differences between real and AI-generated faces, achieving a test accuracy of 98.4%. Performance metrics such as precision and recall further validated the model's reliability.

To enhance the model's capabilities, data preprocessing techniques such as resizing, normalization, and augmentation are employed to standardize and diversify the training data. The model is evaluated using a range of performance metrics, including accuracy, precision, and recall, to assess its ability to generalize to new, unseen images.

The future scope of this project includes expanding the dataset to include more variations of AI-generated images, incorporating real-time image verification, and developing a user-interactive platform where users can upload images to verify their authenticity. This interactive platform would enable immediate detection of AI-generated faces, helping to mitigate the spread of misinformation in digital spaces.

Ultimately, the project lays the foundation for the development of advanced AI-based image detection tools that could be used across various industries, including media, cybersecurity, and social media, to ensure content authenticity and protect against digital manipulation

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to all the individuals who have supported and assisted us throughout our B. Tech Major Project. First and foremost, we would like to thank our project supervisor, **Dr. Aditya Kaushik**, *Professor*, Department of Applied Mathematics, Delhi Technological University for his constant guidance, support, and encouragement throughout the project. We are indebted to him for sharing his knowledge, expertise, and valuable feedback that helped us in shaping the project.

We would like to extend our sincere thanks to the **Vice Chancellor** of Delhi Technological University, "**Professor Prateek Sharma**" and the faculty members of the Department of Applied Mathematics for their support and encouragement throughout our academic journey.
We are grateful to "**Dr. Ramesh Srivastava**", Head of Department, Department of Applied Mathematics, Delhi Technological University, Delhi for his valuable suggestions and feedback on our project work.

We are also thankful to our parents for their constant support and motivation, which has been our driving force throughout our academic pursuits. We would like to express our gratitude to our friends and classmates who have been a constant source of inspiration and motivation.

Finally, I would like to thank all the participants who participated in the study, without whom this research would not have been possible. I express my sincere gratitude to all the individuals who have directly or indirectly contributed to the success of my project.

Vyom Verma (2K21/MC/182)

Rudrakshi Sabharwal (2K21/MC/142)

Yamini (2K21/MC/183)

CONTENTS

Candidate's Declaration 2
Certificate 3
Abstract 4

Acknowledgment… 5

1. INTRODUCTION
   - 1.1 Overview
   - 1.2 Importance of Classifying Authentic Human Faces and AI-generated Images.
   - 1.3 Project Objectives
     - 1.3.1 Specific Goals
   - 1.4 Ethical Considerations and Challenges
     - 1.4.1 Model Accuracy and Bias
     - 1.4.2 Data Privacy and Security
   - 1.5 Potential Applications and Future Work
     - 1.5.1 Practical Applications
     - 1.5.2 Future Research and Development

2. LITERATURE REVIEW
   - 2.1 Relevance of Deep Learning Models in AI-Generated Image Detection
   - 2.2 Related Works
     - 2.2.1 AI-Generated Image Detection Using Deep Learning
     - 2.2.2 Comparison with Traditional Identification Methods
   - 2.3 Database
   - 2.4 Privacy and Security Concerns

3. PROPOSED METHODOLOGY
   - 3.1 Data Acquisition
   - 3.2 Data Cleaning

3.3 Data Preprocessing
3.3.1. Resizing and Normalization
3.3.2. Image Augmentation
3.3.3. Batch Processing
3.3.4. Dataset Organization
3.3.5. Data Flow and Generation
3.3.6. Class Mode
3.4 Model Architecture
3.5 Training and Testing

4. RESULTS

5. FUTURE SCOPE
5.1 Enhancing the Dataset
5.2 Interactive User Platform for Image Verification
5.3 Real-time Image Verification

6. CONCLUSION

7. REFERENCES

8. PLAGIARISM REPORT

1.INTRODUCTION

1.1 OVERVIEW

This project explores the use of deep learning models, specifically Convolutional Neural Networks (CNNs), to differentiate between real human faces and those generated by artificial intelligence (AI). With the increasing prevalence of AI-generated media, such as deepfakes and synthetic images, it has become crucial to develop reliable methods for detecting AI-created content. The primary goal of this project is to design and train a model capable of accurately classifying images as either real or AI-generated.

The dataset used for training consists of 9.6k images, split between 5,000 real human faces and 4,630 AI-generated faces. These images come from various sources, ensuring a diverse representation of human faces and AI-generated content. The model is trained to recognize intricate differences in texture, lighting, and pixel-level details that differentiate real human faces from AI-generated ones. The system undergoes extensive testing and evaluation to determine its accuracy and robustness.

To enhance the model's capabilities, data preprocessing techniques such as resizing, normalization, and augmentation are employed to standardize and diversify the training data. The model is evaluated using a range of performance metrics, including accuracy, precision, and recall, to assess its ability to generalize to new, unseen images.

The future scope of this project includes expanding the dataset to include more variations of AI-generated images, incorporating real-time image verification, and developing a user-interactive platform where users can upload images to verify their authenticity. This interactive platform would enable immediate detection of AI-generated faces, helping to mitigate the spread of misinformation in digital spaces.

Ultimately, the project lays the foundation for the development of advanced AI-based image detection tools that could be used across various industries, including media, cybersecurity, and social media, to ensure content authenticity and protect against digital manipulation.

1.2 Importance of Classifying Authentic Human Faces and AI-generated Images.

The discernment of veritable human visages from algorithmically contrived counterparts has emerged as a pivotal necessity in the contemporary digital paradigm. As generative artificial intelligence systems attain unprecedented fidelity in synthesizing hyper-realistic imagery, the imperative to demarcate authenticity assumes critical importance. This endeavour underpins the integrity of numerous sectors, such as cybersecurity, where identifying deepfakes mitigates risks of fraudulent impersonation and propagandistic disinformation. In the domain of digital governance and content curation, it fortifies the mechanisms against the proliferation of manipulative visual artifacts. Moreover, within creative and

academic spheres, ensuring the provenance of imagery preserves the sanctity of intellectual rigor and originality. By advancing sophisticated classification methodologies, society is equipped to uphold ethical frameworks, shield individuals from digital subterfuge, and sustain the credibility of the virtual ecosystem.

The discernment of authentic human visages from algorithmically generated counterparts has become a cornerstone of digital integrity in an era dominated by artificial intelligence. As generative AI systems, particularly those leveraging deep learning architectures, achieve unparalleled realism in crafting synthetic imagery, the ability to classify such outputs has profound implications. This endeavour not only protects against deception but also fortifies trust across digital platforms.

In cybersecurity, the classification of AI-generated images plays a critical role in mitigating threats such as identity theft, phishing attacks, and the dissemination of deepfakes designed to manipulate public opinion. These malicious applications can destabilize institutions, disrupt elections, and erode societal trust. Furthermore, content moderation systems on social media platforms depend on distinguishing real from synthetic visuals to curtail the spread of misinformation and uphold community guidelines.

By advancing sophisticated and reliable classification frameworks, we not only enhance our ability to navigate the digital age responsibly but also foster a resilient ecosystem where authenticity and ethical practices are preserved. This critical undertaking ensures that the rapid advancements in AI technology are met with equally robust mechanisms for trust and accountability.

1.3 Project Objectives

Develop a Deep Learning Model for Image Classification

- Create a robust deep learning model capable of accurately distinguishing authentic human faces from AI-generated images.
- Train the model on a comprehensive dataset containing diverse real and AI-generated facial images to ensure adaptability across various demographics and synthetic image styles.
- Align the model's classification capabilities with established benchmarks for image authenticity verification and digital forensics standards.

Evaluate Model Performance

- Assess the model's efficacy using key performance metrics such as accuracy, precision, recall, and F1 score.
- Measure its ability to distinguish between real and AI-generated images, focusing on minimizing misclassification rates, including both false positives and false negatives.
- Perform extensive validation and testing to ensure the model's reliability across diverse datasets and conditions.

Explore Real-World Applications

- Investigate the integration of the deep learning model into practical applications such as social media moderation, content authentication systems, and cybersecurity frameworks.
- Test its utility for assisting digital content platforms in identifying manipulated visuals and enhancing trust among users.
- Evaluate how the model could support decision-making processes in domains like media integrity, misinformation control, and intellectual property management, ultimately fostering a safer and more transparent digital ecosystem.

1.3.1 Specific Goals

- **Data Collection and Preprocessing**:

  The initial step involves compiling a comprehensive dataset comprising real human face images and AI-generated counterparts. These images will be sourced from a variety of repositories, including public datasets, AI generation platforms, and real-world photography collections. The dataset will be designed to capture diversity in demographics, lighting conditions, and AI-generation techniques. Once collected, the images will undergo preprocessing to standardize them for deep learning applications. This includes resizing all images to a consistent resolution, normalizing pixel intensity values, and applying data augmentation techniques like rotations, flips, and colour adjustments. These transformations aim to enhance the model's ability to generalize across diverse scenarios while minimizing the risk of overfitting.

- **Model Development**:

  The project will utilize a CNN-based sequential model, designed specifically for the task of distinguishing real human faces from AI-generated images. The sequential architecture will include multiple convolutional layers for feature extraction, interspersed with max-pooling layers to reduce dimensionality while retaining critical features. Following the convolutional stages, fully connected dense layers will be added to integrate extracted features and make predictions.

Dropout layers will be integrated to prevent overfitting, ensuring the model generalizes well to unseen data. By designing the CNN sequentially, the architecture can be tailored specifically to capture the nuanced differences between real and AI-generated images, providing high classification accuracy while maintaining a straightforward and interpretable structure.

- **Model Training and Evaluation**:

  The training phase will involve exposing the model to the pre-processed dataset, enabling it to learn distinguishing features of real and AI-generated images. Advanced optimization techniques such as Adam or stochastic gradient descent will be employed, along with regularization methods like dropout and batch normalization, to prevent overfitting. Post-training, the model's performance will be assessed using key metrics, including accuracy, precision and recall. These metrics will provide insights into the model's effectiveness in correctly classifying images while minimizing false positives and false negatives. Iterative adjustments to the model's architecture, training parameters, and dataset composition will be made to optimize performance further.

- **Application and Integration**:

  The ultimate goal is to explore the practical applications of the trained model in real-world scenarios. Potential use cases include deploying the model in content moderation systems to identify synthetic media on social platforms, integrating it into digital forensics workflows for verifying image authenticity, and enhancing cybersecurity frameworks to detect manipulated visuals. Additionally, the model could serve as a valuable tool for educational and research purposes, helping researchers study the evolution and implications of synthetic imagery. By integrating this model into practical systems, the project aims to foster a more secure and trustworthy digital environment while mitigating the risks posed by the proliferation of AI-generated content.

1.4 Ethics Consideration and Challenges

The deployment of AI models for detecting real human faces versus AI-generated images presents several ethical considerations and challenges that need to be carefully addressed. One significant concern is the potential for bias in the model. If the training dataset primarily consists of a limited set of real human faces or AI-generated images from specific sources, the model may not generalize well across different demographics, such as varying ethnicities, genders, or ages. This could lead to skewed results and inequitable outcomes, where certain groups are either misidentified or overlooked. To mitigate this risk, it is essential to curate a diverse and representative dataset, ensuring that the model is capable of distinguishing AI-generated images from real faces across a wide range of scenarios and populations.

1.4.1 Model Accuracy and Bias

Accuracy and bias are paramount when developing AI models for face classification. High accuracy ensures that the model can reliably differentiate between real and AI-generated faces, which is crucial for applications in digital content moderation, security, and forensic investigations. However, accuracy alone is insufficient if the model exhibits bias, which can arise if the training data does not adequately represent the diversity of human faces. This bias could lead to unfair or discriminatory outcomes, such as higher rates of misclassification for certain demographics. Ensuring fairness requires the use of diverse datasets and continuous evaluation of the model's performance across various groups. Balancing accuracy and bias mitigation are essential to create an equitable system that can be trusted in real-world applications, such as media authentication and digital security.

1.4.2 Data Privacy and Security

The ethical handling of data, particularly images of real human faces, is critical in AI systems that detect AI-generated content. Privacy and security must be prioritized to protect individuals' sensitive data and ensure that it is used responsibly. Anonymization techniques, such as removing identifiable features or using synthetic datasets, can help prevent personal identification from the images. Secure data handling practices are also essential, including encryption during storage and transmission, to protect the data from unauthorized access. Moreover, robust access control mechanisms must be put in place to ensure that only authorized personnel can handle and access the data. These practices are crucial to safeguarding individual privacy and ensuring that the AI model adheres to data protection regulations such as GDPR, CCPA, or other relevant laws. By emphasizing privacy and security, the project ensures that the deployment of AI for detecting AI-generated versus real human images is both ethical and secure, fostering trust among users and stakeholders.

1.5 Potential Applications and Future Work

The successful implementation of this project has the potential to drive transformative applications in the field of digital content authenticity and AI-generated media detection. A key application is the development of a web application that integrates the trained model, providing users with a convenient and accessible tool for distinguishing real human faces from AI-generated images. This web application can be used by individuals, organizations, and digital platforms to verify the authenticity of media content, enabling quick and efficient analysis without requiring advanced technical expertise. Such a tool could be pivotal in combating the spread of misinformation, safeguarding digital identities, and ensuring content integrity in fields like journalism, social media, and legal investigations. Beyond individual use cases, the model could also be integrated into automated systems for content moderation and digital forensics, providing an additional layer of security and accountability in the digital ecosystem.

1.5.1 Practical Applications

The model can find widespread adoption across various sectors where media authenticity is critical. For instance, digital platforms can incorporate the model into their content moderation systems to flag potentially deceptive or AI-generated content before it is published. Similarly, law enforcement and forensic agencies could use the model to analyze digital evidence, aiding in the identification of manipulated images in criminal investigations. In journalism, the model can help verify the credibility of visual content, ensuring that news reporting is based on authentic media. Additionally, the model could be employed in brand protection, where companies can verify that their promotional materials or advertisements have not been tampered with by unauthorized entities. By automating the detection process, the model can significantly reduce the time and effort required for manual analysis, improving the efficiency and reliability of media verification workflows.

1.5.2 Future Research and Development

Future work on this project will focus on expanding the dataset to include a wider range of AI-generated images and real human faces, particularly from diverse sources and across different demographic groups. This will enhance the model's robustness and ensure it performs effectively in a variety of scenarios. Refining the model architecture will also be a priority, including exploring advanced neural network designs and transfer learning techniques to improve classification accuracy and computational efficiency.

Additionally, implementing real-world trials will be critical for validating the model’s performance under practical conditions. These trials could involve collaboration with content moderation teams, digital forensic experts, and other stakeholders to evaluate the model’s effectiveness in their specific use cases. Continuous updates and retraining of the model will be necessary to keep pace with the rapid advancements in generative AI technologies, ensuring it remains effective against emerging forms of AI-generated media. Engaging with experts in relevant fields, such as cybersecurity and media ethics, will help guide the project's evolution, ensuring that it aligns with societal needs and ethical standards. This iterative approach will pave the way for a scalable and impactful solution to the challenges posed by AI-generated media.

2. LITERATURE REVIEW

2.1 Relevance of Deep Learning Models in AI-Generated Image Detection

Deep learning models have emerged as a pivotal innovation in the field of artificial intelligence, particularly in their ability to analyse and interpret comp lex datasets with exceptional accuracy. These models, a subset of AI, are highly effective at recognizing patterns and making classifications, making them indispensable for tasks such as distinguishing between real human faces and AI-generated images.

In the context of identifying AI-generated content, deep learning models can process vast datasets comprising both authentic and synthetic images, learning to detect subtle differences that might elude human observation. This capability is particularly crucial as the sophistication of AI-generated images, including deepfakes and synthetic media, continues to evolve. By leveraging advanced neural network architectures, such as Convolutional Neural Networks (CNNs), these models can classify images with high precision, enabling timely and reliable verification of digital content.

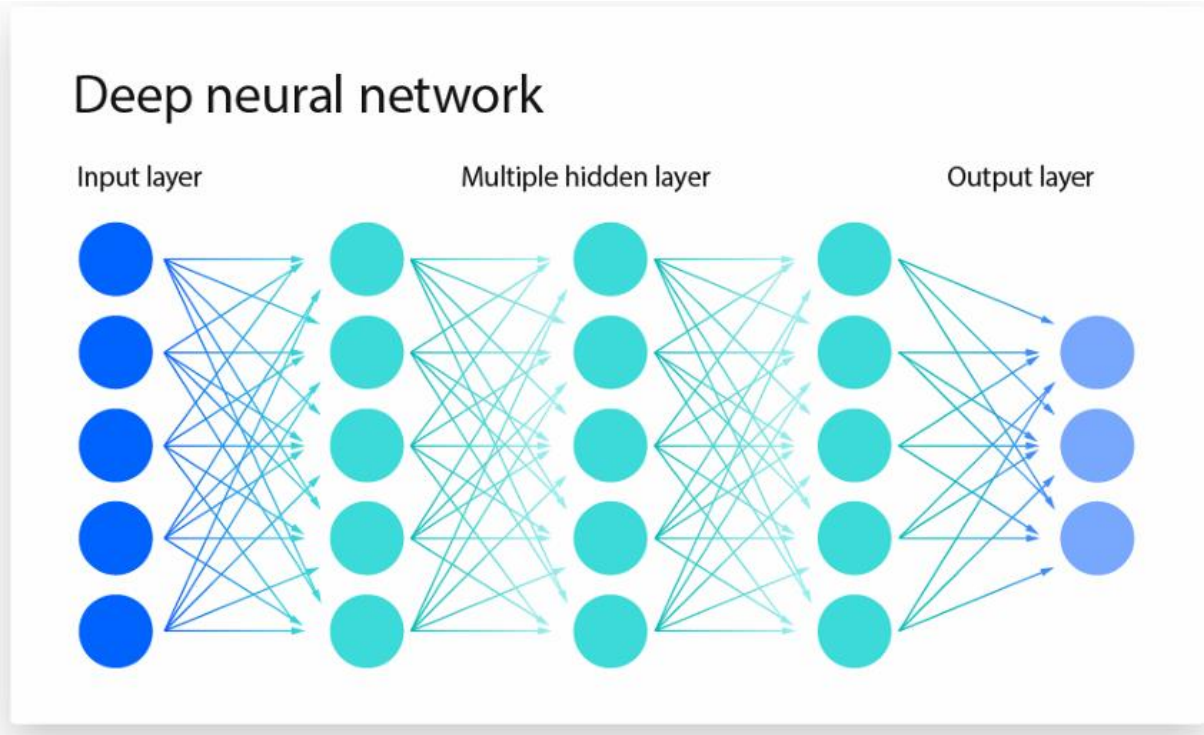

Fig. 1: Architecture of Deep Neural Network

Moreover, deep learning models thrive on large and diverse datasets, improving their accuracy and robustness over time. This adaptability is essential in the rapidly changing landscape of generative AI, where new techniques constantly emerge. By continuously learning from updated datasets, the models remain effective against evolving threats. Overall, deep learning models play a critical role in equipping individuals, organizations, and platforms with the tools to identify and address AI-generated content, ensuring authenticity and trust in digital interactions

2.2 Related Works

Numerous studies have investigated the application of deep learning in image analysis, particularly for tasks such as object detection, facial recognition, and content classification. In the realm of generative AI detection, researchers have developed models capable of identifying manipulated or synthetic images, including deepfakes and other AI-generated media. However, much of the existing work focuses on specific domains, such as video deepfake detection or the analysis of artistic styles in synthetic images.

Few studies have specifically explored the distinction between AI-generated human faces and authentic human faces across diverse datasets, leaving a significant gap in understanding and application. This project addresses this underexplored area by developing a robust classification model tailored to detecting subtle differences between real and synthetic human images, leveraging advancements in Convolutional Neural Networks (CNNs). This focus represents a critical step forward in enhancing content authentication and combating the misuse of generative AI technologies.

2.2.1 AI-Generated Image Detection Using Deep Learning

AI-generated image detection using deep learning harnesses the power of advanced AI models to distinguish between authentic human faces and AI-generated images. By analyzing visual data, such as pixel patterns, textures, and subtle inconsistencies, these models excel at identifying the unique characteristics of synthetic media that are often imperceptible to the human eye.

Deep learning architectures, particularly convolutional neural networks (CNNs), are well-suited for this task due to their ability to extract intricate features from images. When trained on a large and diverse dataset of real and AI-generated faces, these models achieve high accuracy in classification, providing a reliable solution for identifying manipulated content. This automated approach not only reduces the risk of human error but also ensures rapid and precise detection of synthetic media, addressing critical challenges such as misinformation, identity protection, and content authenticity.

The application of such models is particularly valuable in domains where authenticity verification is essential, including social media, journalism, and digital forensics. By leveraging deep learning for AI-generated image detection, stakeholders can maintain trust and integrity in the increasingly complex digital landscape.

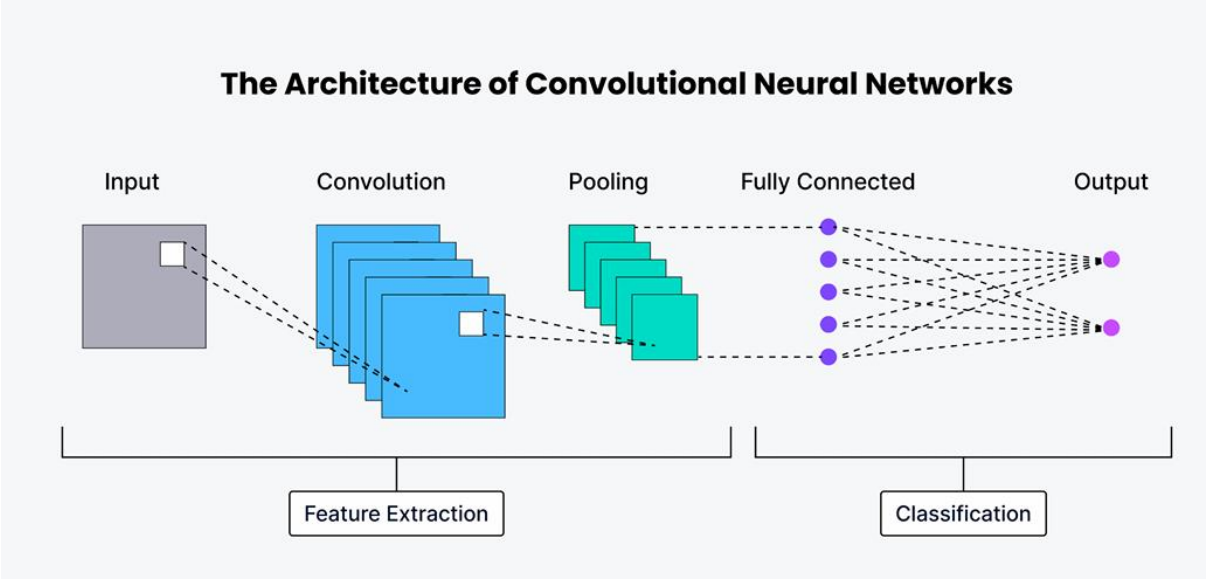

Fig. 2: Architecture of Convolutional Neural Networks

2.2.2 Comparison with Traditional Identification Methods

Traditional methods for identifying AI-generated images often rely on manual inspection or simple algorithmic techniques, such as detecting inconsistencies in lighting, reflections, or image artifacts. These approaches require individuals with significant expertise to evaluate images, which can be subjective, time-consuming, and prone to human error. Moreover, traditional techniques may struggle to keep pace with the rapidly evolving sophistication of generative AI models, making it increasingly difficult to identify subtle manipulations or high-quality synthetic content.

In contrast, deep learning models provide a more automated, objective, and scalable solution. By training on extensive datasets containing both real and AI-generated images, these models can efficiently analyze and classify images with a level of precision unattainable by traditional methods. For example, a convolutional neural network (CNN) can detect fine-grained patterns and inconsistencies that may go unnoticed by human observers, processing large volumes of data in a fraction of the time required by manual techniques.

Additionally, deep learning models continuously improve as they are exposed to more diverse and challenging datasets, ensuring they remain effective against newer and more advanced generative techniques. This adaptability, combined with their speed and accuracy, positions deep learning as a superior alternative to traditional methods for distinguishing between authentic and synthetic human faces in an era of rapidly advancing AI.

2.3 Database

The database for this project comprises a curated collection of images featuring authentic human faces and AI-generated counterparts. These images are sourced from diverse dataset, including publicly available repositories of AI-generated content and authentic human face datasets. The database is designed to represent a wide variety of scenarios, ensuring the inclusion of different lighting conditions, facial expressions, and demographic attributes to enhance the robustness of the model.

This dataset serves as the foundation for training the deep learning model, enabling it to distinguish between real and synthetic images with high accuracy. Each image is meticulously labelled to facilitate supervised learning, and the dataset undergoes preprocessing steps such as resizing, normalization, and augmentation to standardize the inputs and improve model performance. By leveraging a comprehensive and diverse database, the project aims to develop a reliable and adaptable solution for detecting AI-generated images across various applications.

2.4 Privacy and Security Concerns

Privacy and security are paramount considerations in the context of datasets containing human face images, particularly when they include AI-generated and authentic human images. Ensuring that these images do not inadvertently reveal personally identifiable information (PII) is critical. To maintain confidentiality, data anonymization techniques are employed to strip images of any identifiable features, such as metadata that could disclose the identity or location of individuals.

Additionally, robust data handling protocols are implemented to protect the dataset during storage and transmission. Encryption techniques are utilized to secure the images, ensuring that unauthorized access or tampering is prevented. Access control mechanisms are enforced, limiting data access strictly to authorized personnel, and regular system audits are conducted to identify and address any potential vulnerabilities.

Compliance with relevant data protection regulations, such as the General Data Protection Regulation (GDPR), underscores the project's commitment to ethical data usage. These measures ensure that the dataset is handled responsibly, preserving the privacy of individuals depicted in the real images while safeguarding the integrity of AI-generated counterparts.

Addressing privacy and security concerns is crucial for maintaining trust among stakeholders and ensuring that research involving such datasets adheres to the highest ethical standards. This focus on data security also establishes a strong foundation for the responsible development and deployment of AI systems, particularly in applications where human likenesses are central.

3. PROPOSED METHODOLOGY

3.1 Dataset Acquisition

For this project, the dataset is acquired from Kaggle, a platform that provides a wide variety of publicly available datasets. The focus is on gathering images of human faces, both real and AI-generated, to train the deep learning model for distinguishing between the two types. The dataset contains a diverse collection of approximately 9.6k images, consisting of 5,000 real human face images and 4,630 AI-generated face images. This dataset is sourced from Kaggle's repositories, ensuring it meets the quality standards necessary for training the model. Ethical considerations are adhered to, ensuring the data used complies with privacy regulations and does not infringe on copyright. This dataset serves as the foundation for developing the AI model, helping it accurately classify and differentiate between real human faces and AI-generated ones.

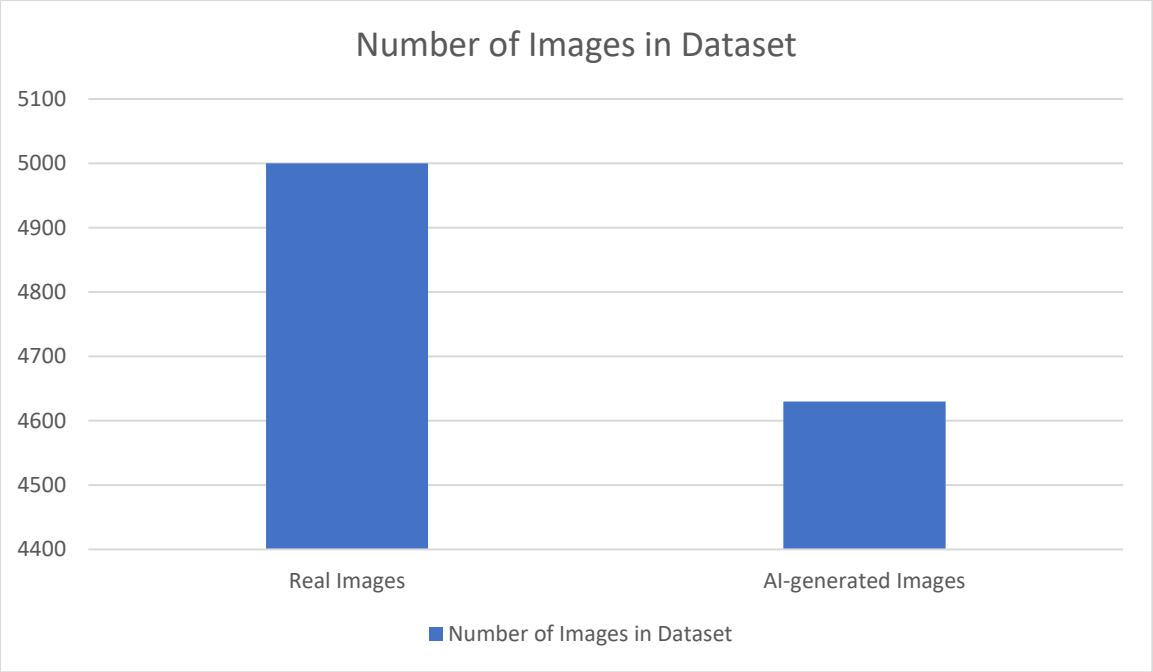

Fig. 3: Distribution of images in dataset

3.2 Data Cleaning

Data cleaning is a crucial phase in preparing the dataset of real human faces and AI-generated images for training the deep learning model. This process ensures that the dataset is of high quality, accurate, and consistent. The first step involves removing low-quality or irrelevant images, such as those with poor resolution, distorted features, or unnecessary artifacts. Duplicate images are identified and eliminated to prevent redundancy and reduce the risk of bias during model training.

Next, all images are standardized to maintain consistency in format and dimensions. They are resized to meet the input requirements of the convolutional neural network (CNN), ensuring uniformity across the dataset. The dataset is also reviewed to identify and address any missing or incorrect labels, with mislabelled images flagged for manual correction to maintain labelling accuracy.

Outliers, such as images that deviate significantly in appearance or content, are reviewed and either corrected or removed, as they could adversely impact the model's learning process. The dataset is then organized into distinct directories: one for real human faces and another for AI-generated images. This structured organization facilitates efficient training and validation of the model.

A meticulous data cleaning process is essential to ensure the integrity and reliability of the dataset. By training the model on well-curated data, the likelihood of accurate and robust classification is significantly enhanced.

3.3 Data Preprocessing

Data preprocessing is a vital step in preparing the dataset for training a deep learning model. It ensures that the input data is in a consistent format, which allows the model to process the images effectively and learn the underlying patterns. This section describes the preprocessing steps applied to AI-generated vs real human face images used in this project.

3.3.1. Resizing and Normalization

To maintain uniformity across the dataset, all images are resized to a consistent size of 256x256 pixels. This is important because deep learning models typically require input images of the same dimensions to process them efficiently. Additionally, the pixel values of the images are normalized by scaling them to a range of 0 to 1. This is done by dividing the pixel values by 255. This normalization helps the model converge faster and improves overall training performance.

3.3.2. Image Augmentation

Image augmentation is a technique used to artificially expand the size of the training dataset by generating new images through various transformations. Although not explicitly defined in the code provided, augmentation techniques such as rotations, flips, zooms, and translations are often applied to improve the model's ability to generalize and reduce overfitting. These transformations help expose the model to different variations of images, making it more robust when classifying unseen data.

3.3.3. Batch Processing

The images are processed in batches, where a specific number of images are passed through the model at a time. In this case, a batch size of 32 images is used. This approach speeds up the training process,

improves memory efficiency, and helps the model update its parameters more frequently, leading to better convergence.

3.3.4. Dataset Organization

The dataset is organized into separate directories for training, validation, and testing, with each category containing subdirectories for the two classes: real human faces and AI-generated faces. Each directory contains the images corresponding to the respective class. This organization ensures that the model can easily differentiate between the two categories during the training and evaluation phases.

3.3.5. Data Flow and Generation

The data is loaded into the model through data generators. These generators automatically load images from their respective directories during training, validation, and testing phases. The images are passed in batches through the generator, where they are resized, rescaled, and augmented as needed. The use of generators ensures that the data is processed efficiently and without the need to load the entire dataset into memory at once, which would be computationally expensive for large datasets.

3.3.6. Class Mode

The class mode is set to **binary**, indicating that the model is dealing with a binary classification problem, where there are two classes: real human faces and AI-generated faces. This setting ensures that the model understands it needs to classify each image as one of two possible categories.

3.4. Model Architecture

The architecture of the Convolutional Neural Network (CNN) designed to classify AI-generated versus real human face images consists of several layers aimed at extracting features from the input images and performing binary classification. The following provides a detailed explanation of each layer and its function in the overall model.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 128, 128, 32)	896
max_pooling2d (MaxPooling2D)	(None, 64, 64, 32)	0
conv2d_1 (Conv2D)	(None, 62, 62, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 31, 31, 64)	0
dropout (Dropout)	(None, 31, 31, 64)	0
flatten (Flatten)	(None, 61504)	0
dense (Dense)	(None, 128)	7,872,640
dropout_1 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 1)	129

Table 1: Model Structure of CNN

1. First Convolutional Layer

- **Conv2D(32, (3, 3), activation='relu', input_shape=(256, 256, 3))**:

  The first convolutional layer applies 32 filters of size 3x3 to the input images. The input images have dimensions of 256x256 pixels with 3 color channels (RGB). The ReLU activation function is used to introduce non-linearity, allowing the network to learn complex features such as edges, textures, and patterns found in human faces and AI-generated faces.

  Input Shape: The input shape is specified as (256, 256, 3), meaning the images are resized to 256x256 pixels with three color channels (RGB).

- **MaxPooling2D(2, 2)**:

  This max-pooling layer down-samples the feature maps by selecting the maximum value from each 2x2 block. Pooling helps reduce the spatial dimensions and computational load, while focusing on the most important features.

2. Second Convolutional Layer

- **Conv2D(64, (3, 3), activation='relu')**:

  The second convolutional layer applies 64 filters to the feature maps produced by the first layer. As the model deepens, it starts to capture more abstract features, such as facial structures, details of lighting, and patterns that distinguish real faces from AI-generated ones.

- **MaxPooling2D(2, 2)**:

  Another max-pooling layer reduces the spatial dimensions of the feature maps. This layer further helps in focusing on the most salient features, improving computational efficiency and model generalization.

3. Third Convolutional Layer

- **Conv2D(128, (3, 3), activation='relu')**:

  This layer applies 128 filters to capture more complex features from the images. The model starts learning high-level features that help distinguish subtle differences between AI-generated and real human faces, such as texture or shading anomalies in AI images.

- **MaxPooling2D(2, 2)**:

  The third max-pooling layer reduces the size of the feature maps, helping the model focus on the important abstract features learned so far.

4. Fourth Convolutional Layer

- **Conv2D(256, (3, 3), activation='relu')**:

  The fourth convolutional layer uses 256 filters to refine the feature extraction process. As the model deepens, it captures even more detailed and higher-level patterns in the faces, improving its ability to differentiate between real and AI-generated images.

- **MaxPooling2D(2, 2)**:

  Max-pooling is applied again to reduce the size of the feature maps, retaining the essential learned features while reducing computational complexity.

5. Flatten Layer

- **Flatten()**:

  The flatten layer reshapes the 2D feature maps into a one-dimensional vector that can be processed by the fully connected (dense) layers. This step is necessary for the transition from convolutional layers to dense layers, where the final classification decision will be made.

6. Fully Connected Dense Layer

- **Dense(128, activation='relu')**:

  The dense layer consists of 128 neurons, each connected to every neuron in the previous layer. The ReLU activation function is applied to introduce non-linearity and enable the model to learn complex relationships between the extracted features. This layer helps in capturing the intricate patterns that distinguish AI-generated faces from real ones.

- **Dropout(0.5)**:

  Dropout is applied with a rate of 50%, meaning that during training, half of the neurons in this layer are randomly deactivated to prevent overfitting. This improves the generalization of the model, making it more robust when evaluating unseen data.

7. Output Layer

- **Dense(1, activation='sigmoid')**:

  The output layer consists of a single neuron with a sigmoid activation function. The sigmoid function outputs a probability between 0 and 1, representing the likelihood that an image is real (1) or AI-generated (0). Since this is a binary classification task, the model's output is used to classify images into one of the two categories: real human faces or AI-generated human faces.

The model architecture employs a series of convolutional layers to extract low-level to high-level features from the input images. Max-pooling layers help reduce spatial dimensions and computational complexity, while dropout layers prevent overfitting. The final classification is made using a dense layer followed by a sigmoid output, which is ideal for binary classification tasks. This CNN architecture is well-suited for distinguishing between real and AI-generated human faces, learning complex features, and generalizing effectively to new images.

3.5 Training and Testing

The training and testing phases are essential for developing an effective deep learning model for classifying AI-generated versus human-generated images. In the training phase, the model is presented with a large set of labelled images that include both real human faces and AI-generated faces. The images are pre-processed to ensure consistency, with each image resized to a standard resolution and normalized for optimal model performance. These labelled images allow the model to learn to distinguish between real and AI-generated faces based on various features.

During training, the model undergoes several iterations (or epochs), where it learns to identify patterns in the images. The primary objective is to minimize the loss function, often binary cross-entropy, by adjusting the model's parameters using optimization algorithms such as Adam. The convolutional neural network (CNN) learns progressively from simple features like edges and textures to more complex features, such as facial structures and patterns that distinguish AI-generated faces from real ones. To enhance the model's robustness and prevent overfitting, data augmentation techniques—such as rotating, flipping, zooming, and cropping images—are applied, increasing the diversity of the training dataset.

Throughout the training process, the model's performance is monitored using a separate validation set. This subset is not used for training but helps assess the model's accuracy, providing insights into its generalization ability and allowing fine-tuning of hyperparameters, such as learning rate, batch size, and the number of layers.

Once training is complete, the model is tested using a previously unseen test set of images. The test set evaluates how well the model generalizes to new data, providing an accurate measure of its performance. Common evaluation metrics, such as accuracy, precision, recall, and F1 score, are used to assess how well the model distinguishes between AI-generated and real human faces. Based on the test results, the model is fine-tuned to optimize its performance, ensuring that it is ready for deployment in real-world applications, where accurate classification of AI versus human images is critical.

4.RESULT

In the final phase of model evaluation, we assessed the performance of our trained Convolutional Neural Network (CNN) model using an unseen test dataset. This dataset consisted of images labelled as either real human faces or AI-generated human faces. The primary metrics for evaluation were **Test Loss** and **Test Accuracy**, which provide insight into the model's ability to generalize to new, previously unseen data.

Test Loss: 0.0485

The **Test Loss** of 0.0485 represents the value of the loss function (categorical cross-entropy in this case) after evaluating the model on the test dataset. A lower loss value indicates that the model's predictions were close to the true labels. In our case, the relatively low-test loss suggests that the model performed well in terms of minimizing prediction errors across the test images. This outcome indicates that the model is effective at recognizing patterns in the test set and is capable of making accurate predictions.

Test Accuracy: 98.41%

The **Test Accuracy** of 98.41% demonstrates the high level of success the model achieved in correctly classifying the images as either real or AI-generated. With an accuracy close to 98.5%, the model exhibits exceptional performance, correctly identifying the majority of test images. This high accuracy suggests that the model has learned the distinguishing features between real human faces and AI-generated faces effectively during the training process and has maintained this ability when faced with new data.

Interpretation

The test results reflect the model's robustness and its capacity to generalize well to unseen data. A high accuracy and low loss are ideal outcomes for a machine learning model, indicating that it is both efficient and precise in its predictions. These results suggest that the model can reliably distinguish between real and AI-generated human faces, making it suitable for deployment in practical applications requiring such classification.

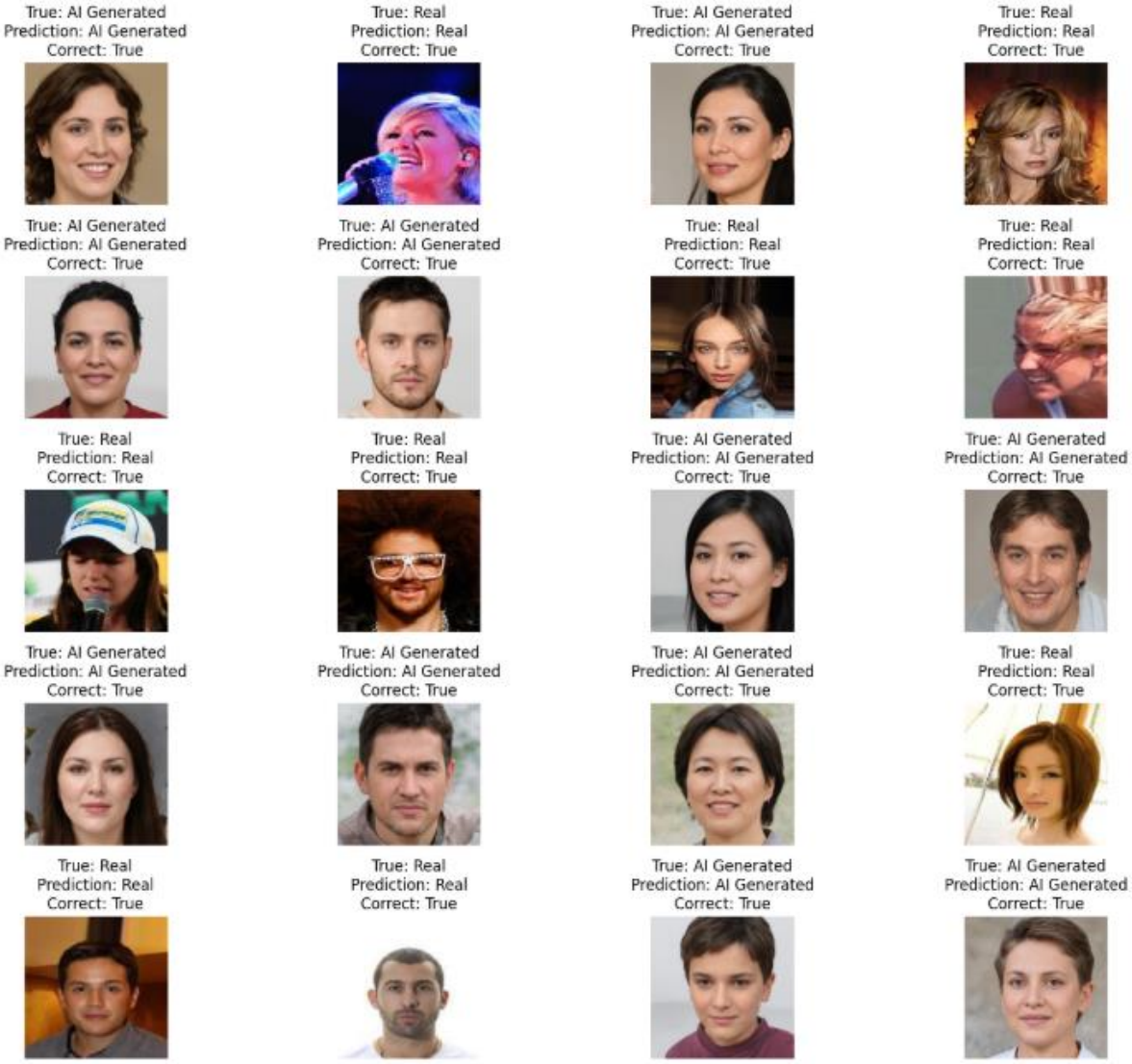

Fig. 4: Result of Tested CNN Model

5. FUTURE SCOPE

This project illustrates the successful implementation of deep learning models to distinguish between AI-generated and real human faces, marking an important step in AI's integration with image verification and authentication. By utilizing a Convolutional Neural Network (CNN), the system has demonstrated strong potential in accurately identifying whether an image is real or AI-generated. The model can be further developed to create a more robust solution for real-world applications in areas like security, media verification, and digital content creation.

5.1 Enhancing the Dataset

To enhance the model's performance and generalization ability, it would be beneficial to expand the dataset. Currently, the dataset includes real and AI-generated human faces, but it can be further diversified by adding images from different lighting conditions, camera angles, and varied backgrounds. Incorporating a wider range of AI-generated faces from different models, such as StyleGAN and others, would also provide the model with more examples to learn from, improving its accuracy and ability to detect subtle differences between real and AI-generated faces in more complex or realistic scenarios.

5.2 Interactive User Platform for Image Verification

To make the AI vs real image classification model more user-friendly, an interactive platform could be developed. This platform would allow users to upload images and receive immediate feedback on whether the image is real or AI-generated. Additionally, users could interact with the platform to provide feedback or correct the model's predictions, making it a dynamic system for improving accuracy over time. Incorporating a feature where users can flag images or suggest corrections would allow the model to learn from real-world data and continuously improve. This user-driven verification system could be beneficial for industries like content creation, online security, and media outlets, where verifying the authenticity of images is crucial.

5.3 Real-time Image Verification

Another exciting future direction involves incorporating real-time image verification. By integrating the trained model into platforms that process images in real-time, such as social media platforms, news outlets, or digital content tools, the system could detect whether an image is real or AI-generated instantly. This could be particularly valuable in areas like combating misinformation, verifying digital content, or securing identity verification systems. Real-time detection systems could assist in preventing the misuse of AI-generated faces, ensuring greater authenticity in digital media.

6.Conclusion

This project successfully demonstrates the application of deep learning models, particularly Convolutional Neural Networks (CNNs), in distinguishing between AI-generated and real human faces. Using a diverse dataset consisting of both real and AI-generated images, the model has proven capable of accurately identifying subtle differences between the two, showcasing its potential for reliable face image classification. With a strong accuracy rate achieved in this proof-of-concept, the model establishes a foundation for advancing the use of AI in detecting fake or manipulated images.

Looking forward, there are several avenues for enhancing the performance and scope of this model. Expanding the dataset to include a broader variety of lighting conditions, camera angles, and different AI-generation techniques will allow the model to generalize better, making it more robust and capable of handling a wide range of real-world scenarios. Additionally, exploring transfer learning using pre-trained models can reduce the training time and computational resources required, while simultaneously improving the accuracy and performance of the model.

Incorporating real-time image verification and developing an interactive user platform could significantly increase the practical applications of this technology. Such a platform would allow users to upload images and receive immediate feedback on whether the image is real or AI-generated, helping to identify and counteract misinformation. Furthermore, integration with existing media and security systems could offer real-time image monitoring, ensuring content authenticity in online platforms, media outlets, and social networks.

The advancements in AI-driven image classification not only have the potential to enhance security and verification systems but also play a crucial role in the growing need for combating digital manipulation in various fields, including journalism, social media, and online content creation. As the field continues to evolve, this technology can contribute to more transparent, trustworthy digital environments, ultimately empowering users and organizations to better distinguish between real and synthetic content.

7. REFERENCE

1. More Real than Real: A Study on Human Visual Perception of Synthetic Faces.Federica Lago (Department of Information Engineering and Computer Science, University of Trento) ,Cecilia Pasquini (Department of Information Engineering and Computer Science, University of Trento), Rainer Böhme (Department of Computer Science, University of InnsbruckHélène DumontInstitute of Neuroscience, Université Catholique de Louvain), Valérie Goffaux (Institute of Neuroscience, Université Catholique de Louvain) Giulia Boato (Department of Information Engineering and Computer Science, University of Trento) https://ar5iv.org/html/2106.07226

2. AI faces look more real than actual human faces. https://www.ucl.ac.uk/news/2023/nov/ai-faces-look-more-real-actual-human-faces

3. How to Distinguish AI-Generated Images from Authentic Photographs(Negar Kamali, Karyn Nakamura, Angelos Chatzimparmpas, Jessica Hullman, Matthew Groh) https://arxiv.org/abs/2406.08651

4. People Now See AI-Generated Faces as More Real Than Human Ones (Marlynn Wei M.D., J.D.) https://www.psychologytoday.com/us/blog/urban-survival/202311/people-now-see-ai-generated-faces-as-more-real-than-human-ones

5. More Real Than Real: A Study on Human Visual Perception of Synthetic Faces [Applications Corner] https://ieeexplore.ieee.org/abstract/document/9664582

6. AI faces look more real than actual human face www.sciencedaily.com/releases/2023/11/231113111717.htm