

Deepfake Face Extraction and Detection using MTCNN-Vision Transformers

1st Richa Singhdept electronics and communication
vardhaman college of engineering

Hyderabad, India

richasinghricha13@gmail.com

3rd B. Chandu Priyadept electronics and communication
vardhaman college of engineering

Hyderabad, India

kripachandu5@gmail.com

2nd K. Ashwinidept electronics and communication
vardhaman college of engineering

Hyderabad, India

kashwini78@gmail.com

4th K. Pavan Kumardept electronics and communication
vardhaman college of engineering

Hyderabad, India

kanchalapavankumar20eccc@vardhaman.com

Abstract—Deepfake detection is a major problem nowadays. The deepfake detection can be done using face extraction and detection. Strong solutions for face extraction and detection are required in light of the growing prevalence of deepfake technology. This paper combines Vision Transformers with Multi-Task Cascaded Convolutional Networks (MTCNN) to propose a novel method. This approach leverages the real-time face identification skills of MTCNN and the long-range dependency capture ability of Vision Transformers to improve the accuracy of detecting manipulated faces in deepfake footage. We carry out extensive experiments on several deepfake datasets, demonstrating the efficacy of the suggested hybrid strategy. The proposed model findings show that this approach performs better than conventional face identification techniques, particularly when dealing with situations where minor facial alterations are involved. This integration strikes a good compromise between computing efficiency and precision, which makes it a viable option for practical uses. The proposed MT-VIT (Multi-Task Vision Transformer) model provides good accuracy as compared to other state-of-the-art like Residual Networks, Mobile Net, CNN, and Meso-Net.

Keywords—multitask cascaded convolutional neural network (MTCNN), vision transformer (VIT), deepfake detection.

I. INTRODUCTION

Deepfake technology is at the forefront of both the possibilities and limitations of artificial intelligence applications, thanks to the rapid growth of deep learning techniques. A combination of "deep learning" and "fake," the term "deepfake" describes the process of manipulating or producing realistic-looking multimedia information, frequently with the use of advanced machine learning models. In this case, a state-of-the-art strategy in the battle against the spread of deepfakes is the combination of multi-task cascaded convolutional networks (MTCNN) for face extraction and vision transformers (VITs) for detection. For face alignment and detection, Multi-task Cascaded Convolutional Networks (MTCNN) is a reliable and popular framework. Its face detection, landmark localization, and bounding box regression three-stage architecture work incredibly well at extracting facial features from complicated images. The MTCNN algorithm is a dependable option for face extraction from a variety of datasets because it has demonstrated efficacy in managing differences in position, lighting conditions, and facial emotions. Particularly in image classification applications, vision transformers (VITs) have become a ground-breaking paradigm change in computer vision. VITs use self-attention mechanisms instead of conventional

convolutional neural networks (CNNs) to collect pixel dependencies and global context. They can learn intricate patterns and long-range relationships as a result, which makes them ideal for identifying the altered face traits found in deepfakes. The suggested framework becomes more adept at identifying minute details and anomalies that might elude conventional detection techniques by utilizing VITs. This work's main contribution is the combination of VIT-based detection and MTCNN-based face extraction, which results in a reliable pipeline for detecting deep fake content. Even in the presence of complex deepfake generation techniques, the suggested framework provides greater accuracy in identifying modified facial features by combining the benefits of these two cutting-edge technologies. A thorough strategy that addresses both the extraction and detection phases of the deepfake production process is made possible by the combination of MTCNN and VITs. This work is significant because it has the ability to lessen the social problems brought about by the fast spread of fake content. Content authentication is made possible by the potent technique of accurately extracting faces using MTCNN and then detecting deepfakes using VITs. This has ramifications for a number of industries, including cybersecurity, journalism, and entertainment, where the ubiquity of false information and online dangers necessitates sophisticated detection systems. Real content can no longer be easily distinguished from modified information thanks to the development of deepfake technology. The increasing complexity of deepfake generation techniques is often beyond the capabilities of current deepfake detection tools. In order to close this gap, this work suggests a novel approach that combines the advantages of MTCNN for face extraction with Vision Transformers for precise and dependable deepfake detection. The main challenge is creating a comprehensive framework that can recognize faces that have been altered in multimedia content, supporting further initiatives to stop the dissemination of false information, and safeguarding the integrity of digital media.

II. RELATED WORK

This research presents a lightweight convolutional neural network (CNN) for real-time face expression detection, achieving 67% accuracy and 3.1% of 16GB of memory use on the FER-2013 dataset. This model overcomes the complexity of deep learning models and manual methods for face detection in various industries [1]. The study explores the rise of deepfakes, artificial intelligence techniques used to create false content, posing a threat to political stability, information integrity, and public trust. It advocates for unified, real-time

solutions and a comprehensive approach that integrates technical solutions with public education [2]. This study examines deepfakes in text, audio, and visual formats, highlighting their growing threat to political stability, public trust, and information integrity. It calls for unified, real-time solutions, public awareness campaigns, and legislative action to combat these issues [3]. The paper discusses the threat of DeepFake, an artificial intelligence technique, and proposes a GAN model to enhance image quality by attacking DeepFake detectors, highlighting the importance of forensic technologies in preventing such attacks [4]. This study aims to identify audio splicing in various acquisition equipment models using clustering, convolutional neural networks, and a distance-measuring technique. It detects speech audio splicing, temporal splicing instants, and changes in voice recordings using unique audio clip features [5]. The study presents a convolutional neural network for deepfake detection using Gabor filters and back-propagation learning. It reduces model size by 64.9% compared to adaptive weighted filters. The architecture addresses existing models' issues with temporal information and self-attention [6]. Chang, Yan, Yamagishi, and Echizen's study proposes cyber vaccination as a solution for deepfake immunity, enabling face-containing media to self-heal after manipulation using AI-driven deepfake technology. The system includes a vaccine, neutralizer, and deepfake imitation, demonstrating effective immunity against face substitution and corruption [7]. Researchers propose a deep learning strategy using convolutional neural networks for facial gender classification using artificial deepfake corpora. The method, which removes data collection barriers and prioritizes privacy, produces low EER scores and high accuracy rates, resembling genuine faces [8]. This study uses artificial intelligence techniques to detect fake audio using machine and deep learning methods. The Mel-frequency cepstral coefficients method is used, and support vector machine and VGG-16 models are used for accuracy [9]. This study uses machine learning and deep learning techniques to detect deepfake audio using the Mel-frequency cepstral coefficients method. The project aims to separate real and fake audio using transfer learning-based techniques, achieving an equal error rate of less than 1.5% [10]. Deepfakes have grown increasingly realistic and challenging to identify in recent years. In this part, we will go over the various deep learning techniques used to produce and identify deepfakes.

A. Deepfake Generation

Deepfake generation, a technique using artificial intelligence to create realistic, phony content, raises concerns about potential misuse for dishonest purposes like impersonation and false information. [7]. Efforts are being made to develop countermeasures and detection techniques to mitigate the negative impacts of deepfake content on various sectors, including politics, entertainment, and cybersecurity [8].

B. Face Swapping

Face swapping is a technique for manipulating digital images in which the face of one person is smoothly replaced with the face of another in a picture or video. Deep learning and computer vision are used to accurately map and transpose faces, gaining popularity in entertainment for light-hearted content. However, concerns about misinformation and privacy necessitate strong detection techniques to prevent misuse and maintain accuracy.[10],[11].

C. Deepfake Detection

Deepfake detection, a process involving advanced AI systems, identifies altered media, including videos, and requires adaptation to prevent false information spread and abuse [12]. [13] Strong detection technologies are crucial for maintaining media integrity, preventing misinformation from deepfake, and preserving trust in digital content across various domains.

D. Deep Learning

Using neural networks, deep learning algorithms are at the forefront of deepfake face detection, able to recognize modified faces. Deepfake face detection uses Convolutional Neural Networks (CNNs) to analyze facial features, distinguishing between real and fake information based on spatial hierarchies [14],[15],[16],[17]. Recurrent Neural Networks (RNNs) simulate facial features' sequential dependencies, detecting inconsistencies generated by deepfake techniques [18],[19],[20]. Mobile Net is a lightweight convolutional neural network architecture suitable for real-time processing on mobile and edge devices [21]. Inception ResNetV2, an advanced design combining residual networks and Inception, enhances deepfake face identification by capturing nuanced facial traits and promoting effective gradient flow during training [22],[23],[24]. Meso-Net is used for discriminating between altered facial material and meaningful micro-expression patterns [25].

III. PROPOSED METHODS

The proposed MT-VIT model integrates vision transformers for real and false detection with MTCNN for precise face extraction, enhancing security, social media authenticity, and other applications requiring accurate face extraction and manipulation detection. The whole model process is displayed in Fig. 1, as shown below. In which step-by-step the proposed model process is explained.

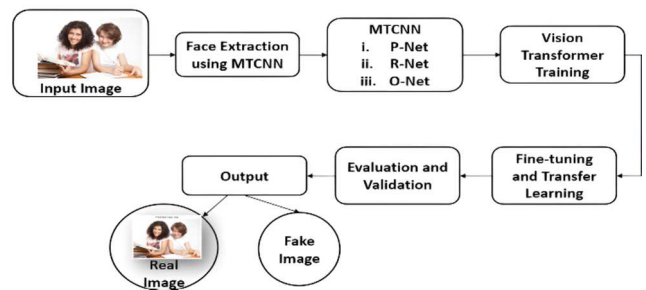


Fig. 1. Block diagram of the proposed MT-VIT model.

A. MTCNN

Because of their strong and versatile design, multi-task cascaded convolutional networks (MTCNN) are very beneficial for face extraction. With its three stages—face detection, bounding box regression, and facial landmark localization—MTCNN is an excellent tool for accurately localizing faces in complicated pictures with a range of scales and orientations. Its multi-stage architecture guarantees flexibility in a range of situations, such as partial occlusion and position changes. The efficiency of MTCNN lies in its precise identification of facial landmarks and areas, offering a complete face extraction solution. MTCNN is an essential component of computer vision tasks, facilitating the accurate and quick extraction of facial information from images, which in turn supports tasks involving analysis or classification. Its

applications include security systems, emotion analysis, and facial identification. From Fig. 2, we can easily understand the architecture of the MTCNN algorithm. From the below equations (1), (2), and (3), we understand the process of face extraction in a given image by using face classification, bounding box regression, and facial landmark localization.

1) *Face Classification*: This is a cross-entropy loss binary classification problem.

$$L_i^{\text{det}} = -(Y_i^{\text{det}} \log(P_i) + (1 - Y_i^{\text{det}})(1 - \log(p_i))) \quad (1)$$

2) *Bounding Box Regression*: A regression issue is the learning objective. The offset between a candidate and the closest ground truth is computed for every candidate window. For this work, Euclidean loss is used.

$$L_i^{\text{box}} = \|\hat{Y}_i^{\text{box}} - Y_i^{\text{box}}\|_2^2 \quad (2)$$

3) *Facial Landmark Localization*: The challenge of localizing facial landmarks is expressed as a regression with Euclidean distance serving as the loss function.

$$L_i^{\text{landmark}} = \|\hat{Y}_i^{\text{landmark}} - Y_i^{\text{landmark}}\|_2^2 \quad (3)$$

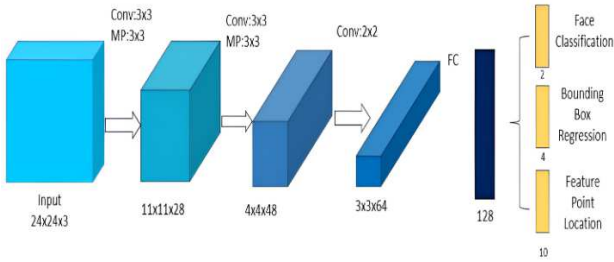


Fig. 2. Architecture of MTCNN which is used in the proposed MT-VIT model.

B. Vision Transformers

Vision transformers (ViTs) are a valuable tool for distinguishing between real and fraudulent images because of their exceptional capacity to extract long-range dependencies and global context from visual input. ViTs interpret images as a sequence of patches, as opposed to standard convolutional neural networks, which enable them to identify complex patterns and correlations throughout the entire image. Because of this, ViTs are excellent at identifying tiny clues that point to picture modification or deepfakes. Through training on a variety of datasets that include both real and altered images, ViTs develop the ability to distinguish between the two groups using global features, allowing for reliable detection. ViTs are an effective tool for real or fraudulent image recognition in a variety of applications, including security, forensics, and media verification. They also help fight disinformation by ensuring the validity of visual content. From Fig. 3, we can analyze the vision transformer process step by step.

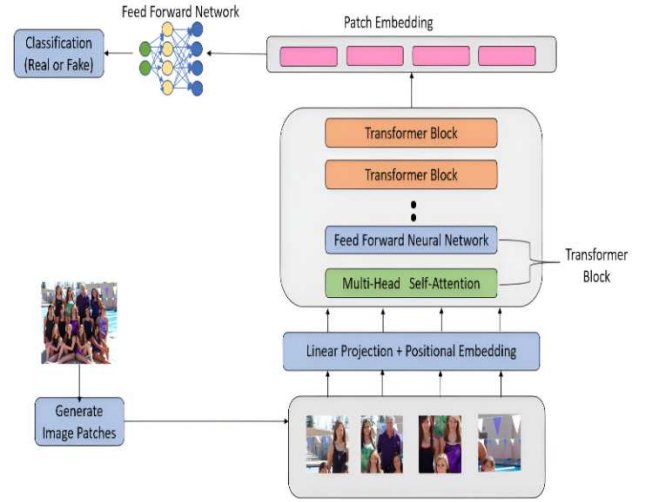


Fig. 3. Architecture of Vision Transformer which has been used in the proposed MT-VIT model.

1) *Multi-Head*: The multi-head attention technique in Vision Transformer (ViT) models uses parallel attention heads, each picking up unique patterns and connections in visual patches. To improve the model's ability to collect various visual aspects and help with image classification and other computer vision tasks, the outputs are concatenated and linearly projected.

2) *Self Attention*: The Vision Transformer (ViT) model uses linear projections of input embeddings into query (Q), key (K), and value (V) vectors to implement self-attention. By calculating the dot product of Q and K, scaling the results, applying a softmax to get weights, and finally computing a weighted sum of the values, attention scores are calculated. The attention formula is displayed in the below equation (5), using Softmax. Through this procedure, the model is able to assign relative weights to various input sequence elements.

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})V \quad (4)$$

IV. RESULTS AND DISCUSSION

A. Dataset

For the proposed MT-VIT model, the paper has used the 140k real and fake datasets for the detection of real and fake images. Firstly, select an image from the provided dataset, extract the individual faces from a group of faces in a single image, and represent it, which is done by the MTCNN algorithm. For the extracted faces, again check whether the extracted face is real or fake using the vision transformer. The model is trained on the training set, its performance is tracked during training on the validation set, and its ultimate performance is assessed on the test set. The ViT model's accuracy and quality can be raised by developers by appropriately preprocessing the input photos. For the proposed model, they have used 1,40,000 images, and for training, they have used 1,00,000 images. For validation, they have used 20,000 images, and testing is done on the 20,000 images from the provided dataset.

The final output of the proposed model is represented in Fig. 4 for the 140k dataset, which shows the faces extracted from the dataset using the MTCNN algorithm. In this process, first they classified the different faces and then created a

bounding box for the exact face using the localization process. Then, finally, facial landmarks from the before-stage extracted output were presented, presenting the final extracted face image. From Fig. 5, you can see the output of the fake images detected by the vision transformer algorithm, and from Fig. 6, you can see the output of the real images detected by the proposed model. For a better understanding of the proposed model, we can see Fig. 8, which illustrates the performance of the proposed model. In this, you can see two graphs, i.e., one graph represents the accuracy between the training set and the validating set, and the other graph represents the loss between the training set and the validating set. Along with this, they have also plotted the confusion matrix between the predicted value and the actual value. Fig. 7 illustrates the comparison of different existing models with the proposed MT-VIT model in bar graph format. In this graph, you can see the accuracy, recall, and precision values of the models used. This graphical representation helps others understand the model more efficiently.

In the below two tables, they are doing a quantitative analysis of the proposed model and the other state-of-the-art models. From the table, you can see the different existing datasets used in the model, and along with that, you can also see the number of real and fake images present in that particular dataset, i.e. DFFD [23] is the dataset that contains 58,708 real images and 240,336 fake images. real and fake face extraction [24] The dataset has 1000 fake and 1200 real images in it. The dataset has 67,000 fake and 20,000 real images in it, or 100k fake. The [21] dataset has 65,000 fake and 35,000 real images in it, and the dataset has 70,000 fake and 70,000 real images in it. The 140k dataset is used for the proposed model to extract and detect the images. In Table 1, they are analyzing different datasets with the proposed dataset for the better accuracy of the model. Using Table 1, you can do a quantitative analysis of the number of real and fake images present in each dataset. To understand the proposed model in a better way, you can see Table 2, which represents the different models used in this proposed model for comparison with the proposed MT-VIT model. In these tables, they are doing a quantitative analysis of the accuracy, recall, and precision values for the proposed model and the other state-of-the-art models. The existing models used in this are CNN [17], RNN [11], InceptionResNetv1 [23], and Mobile Net [21]. Table 2 also illustrates the real and fake images detected from the existing models and the proposed MT-VIT model. In Table 2, each deep learning model is trained and analyzed with the proposed model to check the accuracy and efficiency of the proposed MT-VIT model. The quantitative analysis done for the detection of real and fake for the other state-of-the-art model with the proposed model is displayed in Table 2. The quantitative analysis is done based on the accuracy, recall, and precision values.



Fig. 4. Illustrates how we are detecting faces from different input images and then extracting each face individually.



Fig. 5. Illustrate the fake images of the vision transformer in the proposed MT-VIT model.



Fig. 6. Illustrate the real images of the vision transformer in the proposed MT-VIT model.

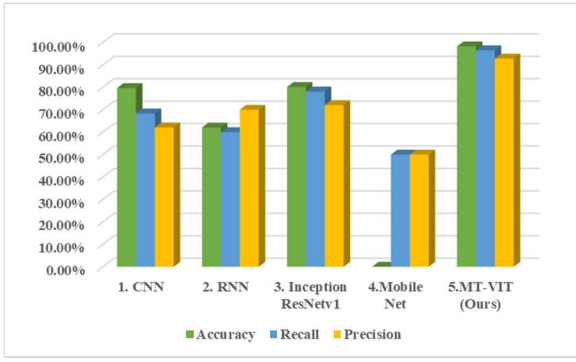


Fig. 7. Illustrates the comparison of the proposed MT-VIT model with the other state-of-the-art model.

TABLE I. QUANTITATIVE ANALYSIS OF DIFFERENT DATASETS

Model	Real Images	Fake Images
DFFD	58,708	240,366
real and fake face extraction	1000	1200
ifake face	67,000	20,000
100k fake	65,000	35,000
140k	70,000	70,000

B. Quantitative Analysis of Dataset

Fig. 9 illustrates the quantitative analysis of the fake image output of the four datasets used in the proposed model. From Fig. 9, we can analyze the output of the four datasets and compare them with the selected 140k dataset for the detection of real and fake images. and Fig. 10 illustrates the quantitative analysis of the real image output of the datasets used for this model. The represented output images in figs. 9 and 10 use the DFFD [23], real and fake face extraction [24], fake face extraction, and 140k. By performing the quantitative analysis, you can see the output images of fake and real faces separately in the below figures. By doing this quantitative analysis, you can see the difference among the output images of each dataset used in the model. Table I illustrates the quantitative analysis of the different dataset used in the model for the comparison and for performing this proposed model. Table II. illustrates the quantitative analysis of the accuracy, recall, and precision values of our proposed model with state-of-the-art of other models. In this table we are analyzing the different deep learning model with the proposed model.

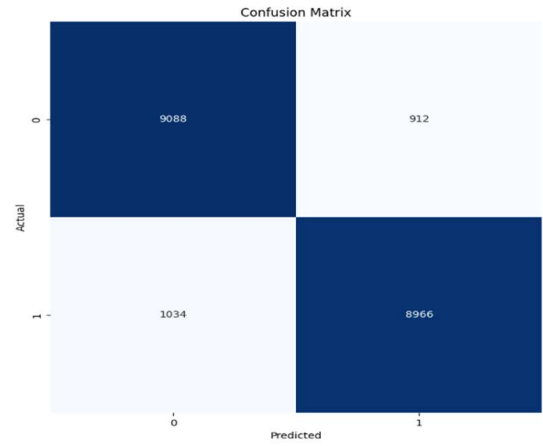
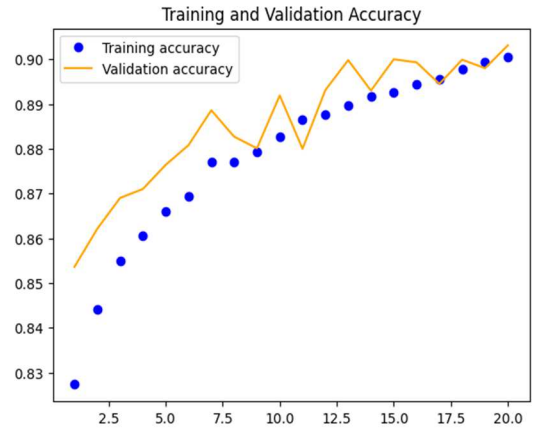
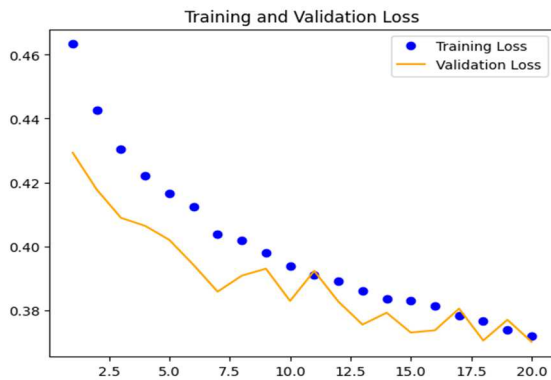
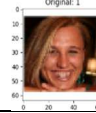

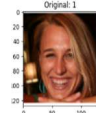









Fig. 8. Illustrates the performance graph of the proposed model. Here we are having accuracy and loss graph and along with that we also have the confusion matrix.

TABLE II. QUANTITATIVE ANALYSIS OF DIFFERENT MODELS

Models	Accuracy	Recall	Precision	Real Images	Fake Images
CNN [17]	79.66%	68.44 %	62.07%		
RNN [11]	62.07%	60.00 %	70.00%		
Inception ResNet [23]	80.12%	78.13 %	72.05%		
Mobile Net [21]	60.00%	50.00 %	50.00%		
MT-VIT	98.22%	96.48 %	92.79%		

C. Performance Metrics

The evaluation of models involves assessing their performance using several widely utilized measures to

classify drone images. These measures include accuracy, recall, and precision.

- Accuracy = True Negative + True Positive / Total no. of samples
- Recall = True Positive / True Positive + False Negative
- Precision = True Positive / True Positive + False Positives

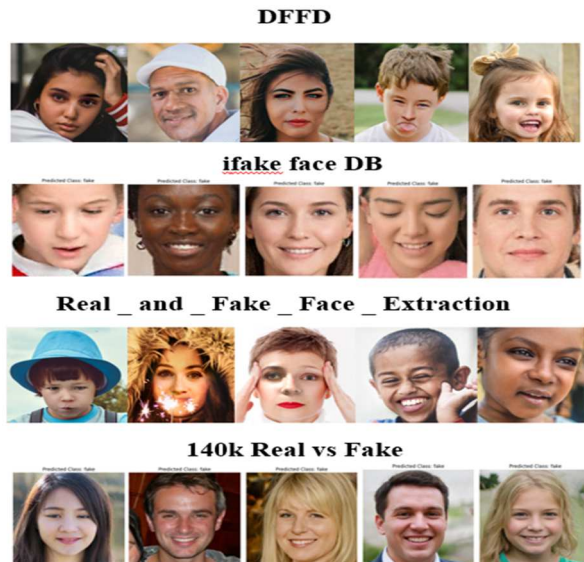


Fig. 9. Illustrates the quantitative analysis of the fake output images of the taken four datasets for finding the best dataset to detect real and fake images for the proposed model.

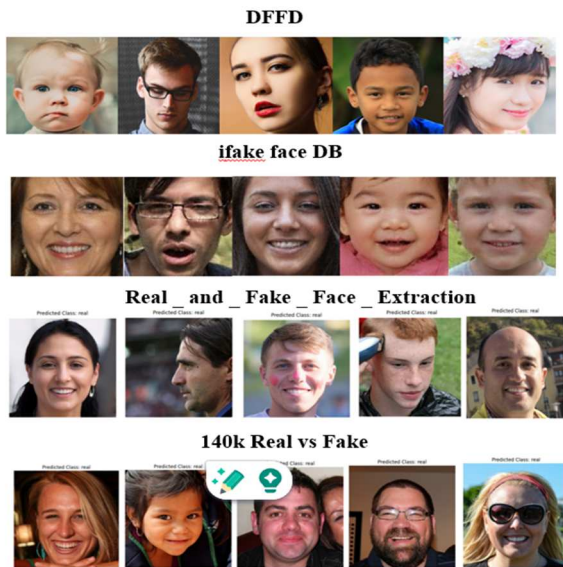


Fig. 10. Illustrates the quantitative analysis of the real output images of the taken four datasets for finding the best dataset to detect real and fake images for the proposed model.

D. System and Software Required

For the proposed model in the system requirement, we have used mainly three components. They are hardware, memory, and storage.

1) *Hardware*: In the hardware, we have used a high-performance environment, i.e., a graphical processing unit

(GPU), to accelerate the deep learning computations involved in face extraction and detection. 7

2) *Memory (RAM)*: A sufficient amount of RAM is used to support the simultaneous execution of the deep learning model and for the processing of high-resolution images.

3) *Storage*: We have an adequate storage capacity to hold the pre-trained model, intermediate outcomes produced during the face extraction and detection stages, and the datasets.

In the software requirement, we have focused on the operating system, deep learning frameworks, programming language, libraries and dependencies, model pre-trained weights, and development of the environment.

4) *Operating System*: For this proposed model, we have used the Windows system. The overall process is done on this operating system only.

5) *Deep Learning Frameworks*: In this, we have installed the required PyTorch and Tensorflow frameworks.

6) *Programming Language*: The complete proposed model is done using the Python language, as its proficiency is high for implementing and integrating MTCNN and Vision Transformer models in the project.

7) *Libraries and Dependencies*: We have installed all the required libraries and dependencies, like OpenCV, Numpy, and Scikit-Learn.

8) *Model pre-trained weights*: Due to the availability of pre-trained weights for MTCNN and VIT models, it is easy to transfer learning and enhance the accuracy of the model.

9) *Development of Environment*: For the development of this project, we have used IDEs such as Jupyter Notebook and Google Colab for the development of code, debugging, and experimentation.

E. Discussion

In this study, we explored the complex field of deepfake face extraction and recognition using a clever combination of Vision Transformer (ViT) and MTCNN (Multi-task Cascaded Convolutional Networks) models. The accurate extraction of facial features, which is an essential first step in the detection pipeline, was made possible by the use of MTCNN. The multi-stage architecture of MTCNN made it possible to precisely localize facial landmarks, which advanced our understanding of facial features. By leveraging attention techniques for comprehensive feature representation, the inclusion of the Vision Transformer greatly improved detection accuracy. ViT's capacity to extract contextual information and long-range dependencies from facial photos was crucial in identifying the minute changes included in deepfake content. The issues presented by complicated manipulations inherent in deepfake generation, lighting circumstances, and variable facial expressions were handled by the synergistic combination of MTCNN and Vision Transformer. Our results highlight how well this hybrid strategy works to strengthen defenses against dishonest, deep-fake activities. MTCNN and Vision Transformer have shown strong face extraction and identification skills, which show promise in reducing the threats to society that arise from the spread of deceptive visual information. This study adds to the body of knowledge regarding deepfake countermeasures by highlighting the need to use cutting-edge methods for improved face manipulation detection accuracy and dependability. From tables 1 and 2, you can see the quantitative analysis of different datasets used

in the proposed model, and we have also done the quantitative analysis of the other state-of-the-art models with our proposed model to show the difference and accuracy rate of the proposed model is better than other methods. The proposed model outputs are represented in Fig. 5, Fig. 6, and Fig. 7. Fig. 5 illustrates the output images of the extracted faces from the provided 140k dataset. Figures 6 and 7 illustrate the fake and real faces, respectively, which are detected using the vision transformer algorithm. Figs. 9 and 10 represent the quantitative study and analysis of the different datasets used in the model and display all the detected fake and real images from each dataset. The overall performance of the proposed model is good, as it is a combination of MTCNN and vision transformers. The MTCNN algorithm has high accuracy in extracting faces compared to other face extraction methods. The main reason for using the vision transformer rather than other deep learning or machine learning algorithms is that the vision transformer has high computational speed and accuracy levels in detecting images.

V. CONCLUSION

In conclusion, a complete and practical solution in the field of computer vision is provided by the combined use of multi-task cascaded convolutional networks (MTCNN) for face extraction and vision transformers (ViTs) for the detection of real or fraudulent images. MTCNN is an important component in the early phases, exhibiting impressive localization accuracy for faces in intricate pictures. Its three-stage architecture makes it a strong option for face extraction tasks by guaranteeing adaptability to different scales, orientations, and partial occlusions. By adding a potent method for determining the veracity of facial photographs, the integration of Vision Transformers enhances the capabilities of the system. ViTs are excellent at identifying minute clues that point to picture tampering since they are built to record complex patterns and global context. ViTs can be trained on a variety of datasets that contain both real and modified images. This helps the model learn and generalize, improving its efficaciousness in distinguishing between real and fake images. The combination of ViTs for genuine or false identification with MTCNN for accurate face extraction creates a strong basis as we move toward a time where altered images represent ever-greater risks. The ongoing development of deep learning methods and the cooperative application of various architectures represent the advancement toward increasingly complex, dependable, and adaptable computer vision systems that can handle the complexities of image processing in practical applications.

VI. FUTURE SCOPE

Promising developments lie ahead for face extraction with MTCNN and fraudulent picture detection with Vision Transformers (ViTs). Refinement of the model, integration with explainable AI, transfer learning for domain-specific applications, resilience to adversarial assaults, real-time implementation, diversity expansion of the dataset, and cooperation with multimedia analysis techniques can be the primary areas of future research. These developments may result in improved face extraction accuracy and a deeper comprehension of intricate patterns. Furthermore, ethical considerations such as fairness in face extraction, deepfake detection, and bias reduction should be given top priority in future studies. Computer vision systems in the future can be more ethical, flexible, and enhanced for a range of uses in the digital world by concentrating on these areas.

REFERENCE

- [1] V. N. Tran, S. G. Kwon, S. H. Lee, H. S. Le, and K. R. Kwon, "Generalization of Forgery Detection With Meta Deepfake Detection Model," *IEEE Access*, vol. 11, no. November 2022, pp. 535–546, 2023, doi: 10.1109/ACCESS.2022.3232290.
- [2] F. Ding, G. Zhu, Y. Li, X. Zhang, P. K. Atrey, and S. Lyu, "Anti-Forensics for Face Swapping Videos via Adversarial Training," *IEEE Trans. Multimed.*, vol. 24, no. c, pp. 3429–3441, 2022, doi: 10.1109/TMM.2021.3098422.
- [3] A. H. Khalifa, N. A. Zaher, A. S. Abdallah, and M. W. Fakhr, "Convolutional Neural Network Based on Diverse Gabor Filters for Deepfake Recognition," *IEEE Access*, vol. 10, pp. 22678–22686, 2022, doi: 10.1109/ACCESS.2022.3152029.
- [4] J. Lan et al., "Expression Recognition Based on Multi-Regional Coordinate Attention Residuals," *IEEE Access*, vol. 11, no. May, pp. 63863–63873, 2023, doi: 10.1109/ACCESS.2023.3285781.
- [5] K. N. Ramadhani, R. Munir, and N. P. Utama, "Improving Video Vision Transformer for Deepfake Video Detection using Facial Landmark, Depthwise Separable Convolution and Self Attention," *IEEE Access*, vol. 12, no. January, pp. 8932–8939, 2024, doi: 10.1109/ACCESS.2024.3352890.
- [6] S. A. Khan and D.-T. Dang-Nguyen, "Deepfake Detection: Analysing Model Generalisation Across Architectures, Datasets and Pre-Training Paradigms," *IEEE Access*, vol. 12, no. December 2023, pp. 1–1, 2023, doi: 10.1109/access.2023.3348450.
- [7] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," *IEEE Access*, vol. 11, no. December, pp. 144497–144529, 2023, doi: 10.1109/access.2023.3344653.
- [8] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022, doi: 10.1109/ACCESS.2022.3151186.
- [9] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "DeepFake Detection Based on Discrepancies Between Faces and Their Context," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6111–6121, 2022, doi: 10.1109/TPAMI.2021.3093446.
- [10] Z. Tan, Z. Yang, C. Miao, and G. Guo, "Transformer-Based Feature Compensation and Aggregation for DeepFake Detection," *IEEE Signal Process. Lett.*, vol. 29, pp. 2183–2187, 2022, doi: 10.1109/LSP.2022.3214768.
- [11] C. Rathgeb, C. I. Satnoianu, N. E. Haryanto, K. Bernardo, and C. Busch, "Differential Detection of Facial Retouching: A Multi-Biometric Approach," *IEEE Access*, vol. 8, pp. 106373–106385, 2020, doi: 10.1109/ACCESS.2020.3000254.
- [12] Y. Zhu et al., "Information-Containing Adversarial Perturbation for Combating Facial Manipulation Systems," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 2046–2059, 2023, doi: 10.1109/TIFS.2023.3262156.
- [13] L. Zhang, T. Qiao, M. Xu, N. Zheng, and S. Xie, "Unsupervised Learning-Based Framework for Deepfake Video Detection," *IEEE Trans. Multimed.*, vol. 25, pp. 4785–4799, 2023, doi: 10.1109/TMM.2022.3182509.
- [14] D. U. Leonzio, L. Cuccovillo, P. Bestagini, M. Marcon, P. Aichroth, and S. Tubaro, "Audio Splicing Detection and Localization Based on Acquisition Device Traces," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 4157–4172, 2023, doi: 10.1109/TIFS.2023.3293415.
- [15] C. C. Chang, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Cyber Vaccine for Deepfake Immunity," *IEEE Access*, vol. 11, no. August, pp. 105027–105039, 2023, doi: 10.1109/ACCESS.2023.3311461.
- [16] M. Oulad-Kaddour, H. Haddadou, C. C. Vilda, D. Palacios-Alonso, K. Benatchba, and E. Cabello, "Deep Learning-Based Gender Classification by Training With Fake Data," *IEEE Access*, vol. 11, no. October, pp. 120766–120779, 2023, doi: 10.1109/ACCESS.2023.3328210.
- [17] S. Sadiq, T. Aljrees, and S. Ullah, "Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets," *IEEE Access*, vol. 11, no. July, pp. 95008–95021, 2023, doi: 10.1109/ACCESS.2023.3308515.
- [18] K. Khan, R. U. Khan, K. Ahmad, F. Ali, and K. S. Kwak, "Face Segmentation: A Journey from Classical to Deep Learning Paradigm, Approaches, Trends, and Directions," *IEEE Access*, vol. 8, pp. 58683–58699, 2020, doi: 10.1109/ACCESS.2020.2982970.

- [19] M. F. Hashmi, B. K. K. Ashish, A. G. Keskar, N. D. Bokde, J. H. Yoon, and Z. W. Geem, "An Exploratory Analysis on Visual Counterfeits Using Conv-LSTM Hybrid Architecture," *IEEE Access*, vol. 8, pp. 101293–101308, 2020, doi: 10.1109/ACCESS.2020.2998330.
- [20] Y. X. Luo and J. L. Chen, "Dual Attention Network Approaches to Face Forgery Video Detection," *IEEE Access*, vol. 10, no. October, pp. 110754–110760, 2022, doi: 10.1109/ACCESS.2022.3215963.
- [21] S. Waseem, S. A. R. S. Abu Bakar, B. A. Ahmed, Z. Omar, T. A. E. Eisa, and M. E. E. Dalam, "DeepFake on Face and Expression Swap: A Review," *IEEE Access*, vol. 11, no. August, pp. 117865–117906, 2023, doi: 10.1109/ACCESS.2023.3324403.
- [22] Y. Patel et al., "Deepfake Generation and Detection: Case Study and Challenges," *IEEE Access*, vol. 11, no. December, pp. 143296–143323, 2023, doi: 10.1109/ACCESS.2023.3342107.
- [23] N. M. Alnaim, Z. M. Almutairi, M. S. Alsuwat, H. H. Alalawi, A. Alshobaili, and F. S. Alnezi, "DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era With Deepfake Detection Algorithms," *IEEE Access*, vol. 11, no. February, pp. 16711–16722, 2023, doi: 10.1109/ACCESS.2023.3246661.
- [24] X. Li, Z. Yang, and H. Wu, "Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks," *IEEE Access*, vol. 8, pp. 174922–174930, 2020, doi: 10.1109/ACCESS.2020.3023782.
- [25] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022, doi: 10.1109/ACCESS.2022.3154404.