A Novel Approach for Detection of Deepfake Images Using Deep Learning

Ms. Ankita Wagh
Department of Computer Science
Sant Gadge Baba Amravati University ,Amravati
Waghankita877@gmail.com

Prof. S. S. Sherekar
Department of Computer Science
Sant Gadge Baba Amravati University ,Amravati
swatisherekar@sgbau.ac.in

ABSTRACT:

In recent years, the rapid growth of digital media and deep learning technologies has made it easier to create manipulated or fake facial images, which can lead to serious security and privacy concerns. This research presents a deep learning-based approach for the classification of real and fake face images. The proposed system uses a convolutional neural network (CNN) model to automatically extract important facial features and distinguish between authentic and manipulated images.

The model is trained on a dataset containing both real and fake face images, followed by preprocessing steps such as image resizing and normalization. After training, the model is evaluated on a separate test dataset to measure its performance. Evaluation metrics including accuracy, precision, recall, and F1-score are used to analyze the effectiveness of the model.

Experimental results demonstrate that the proposed approach is capable of identifying fake images with reasonable performance. Although the model shows promising results, further improvements can be achieved by using a larger dataset and more advanced architectures. This work highlights the potential of deep learning techniques in detecting fake visual content and contributes to enhancing digital image security.

INTRODUCTION:

In today's digital world, images and videos are widely used for communication, social media, and security purposes. With the advancement of artificial intelligence and deep learning technologies, it has become increasingly easy to manipulate facial images and create realistic fake content. Such manipulated images, often referred to as deepfakes, can be misused in various fields including social media, digital forensics, identity verification, and cybersecurity. This creates a serious challenge in distinguishing between real and fake images.

Traditional image processing techniques are often not sufficient to detect such sophisticated manipulations, as fake images can closely resemble real ones. Therefore, there is a need for more advanced and automated approaches that can accurately identify fake content. Deep learning, especially Convolutional Neural Networks (CNNs), has shown significant potential in image classification and feature extraction tasks. These models can automatically learn complex patterns and subtle differences in images without manual feature engineering.

This research focuses on developing a deep learning-based system to classify facial images as real or fake. The proposed method uses a CNN model to extract important features from images and perform classification. The system involves steps such as data collection, preprocessing, model training, and evaluation. The performance of the model is analyzed using standard evaluation metrics such as accuracy, precision, recall, and F1-score.

The main objective of this work is to provide an efficient and reliable solution for detecting fake facial images. This study also highlights the importance of using deep learning techniques in enhancing the security and authenticity of digital content.

RELATED WORK:

Deep fake detection has advanced rapidly with the development of AI and deep learning techniques. Early studies, such as those by Hanady Sabah Abdul Kareem (2023) and Savitha AC (2025), focused on traditional machine learning methods like SVM combined with feature reduction (e.g., PCA). These approaches showed moderate success in detecting manipulated facial images, but they often struggled against highly realistic fakes.

With the growth of deep learning, Convolutional Neural Networks (CNNs) have become the dominant approach. Reschke (2025), Velumani et al. (2024), and Jagam et al. (2025) demonstrated that pre-trained architectures like ResNet, VGG, and DenseNet, when combined with transfer learning and fine-tuning, can extract complex features and accurately classify images as real or fake. Advanced strategies such as hybrid model fusion, pairwise learning, and contrastive loss further enhance detection performance (Hsu et al., 2020).

Large-scale datasets including DFDC and FaceForensics++ (Dolhansky et al., 2019; Sonkusare et al., 2022) have played a key role in training and benchmarking models. Despite these improvements, challenges remain in terms of generalization, demographic bias, and robustness against increasingly realistic GAN-generated content (Xu et al., 2024; Mageswari et al., 2024). Ethical concerns, such as misinformation and privacy risks, have also been highlighted (Somanje et al., 2025).

Motivated by these findings, this research proposes a CNN-based framework for classifying facial images as real or fake, offering a simple yet effective solution to improve digital content authenticity

PROPOSED METHODOLOGY:

Introduction:

With the rapid growth of deep fake technologies, distinguishing real face images from fake ones has become challenging. This study proposes a deep learning-based system using Convolutional Neural Networks and transfer learning to accurately detect real and fake face images. The methodology focuses on processing images through data collection, preprocessing, feature extraction, model training, and evaluation to achieve reliable and effective results.

Research Gap :

Despite the rapid development of deep learning methods for image analysis, existing systems for detecting fake face images often face challenges such as limited dataset diversity, high computational requirements, and reduced accuracy on unseen or manipulated images. Many current approaches fail to generalize well across different types of fake images or rely heavily on handcrafted features. Therefore, there is a need for an improved automated system that can achieve higher accuracy, better generalization, and efficient detection of real and fake face images.

Proposed Methodology:

1. **Data Collection :** The first step of the proposed system involves collecting a dataset consisting of real and fake face images. The dataset should include a sufficient number of samples to ensure proper training and testing of the model.
2. **Data Preprocessing:** In this stage, the collected images are prepared for model input. Images are resized to a fixed dimension and normalized to maintain consistency. This step helps in improving model performance.
3. **Face Detection using OpenCV:** OpenCV is used to detect and extract the facial region from each image. This ensures that only the important part of the image (face) is used for further processing.
4. **Data Augmentation:** To increase dataset diversity and avoid overfitting, augmentation techniques such as rotation, flipping, and brightness adjustment are applied. This improves the generalization ability of the model.
5. **Dataset Splitting:** The dataset is divided into training and testing sets. The training data is used to train the model, while the testing data is used to evaluate its performance.
6. **Feature Extraction using CNN:** A Convolutional Neural Network to analyze (CNN) is used to automatically extract important features from images, such as edges, textures, and patterns that help distinguish real and fake images.
7. **Transfer Learning:** A pre-trained model such as VGG16 or ResNet is used to improve performance and reduce training time. Transfer learning allows the model to use previously learned features.
8. **Model Training:** The model is trained using the training dataset. During training, optimization techniques and loss functions are applied to improve model accuracy.
9. **Model Evaluation:** After training, the model is evaluated using test data. Performance metrics such as accuracy, precision, recall, and F1-score are calculated to measure effectiveness.
10. **Confusion Matrix Analysis:** A confusion matrix is used the performance of the model. It shows the number of correct and incorrect predictions, including true positives, true negatives, false positives, and false negatives.

11. **Final Classification:** The system provides the final output by classifying the input image as either real or fake based on the trained model

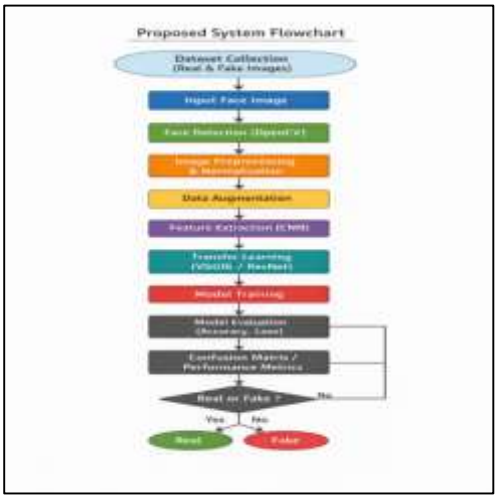


Fig: Proposed Methodology Of Architecture

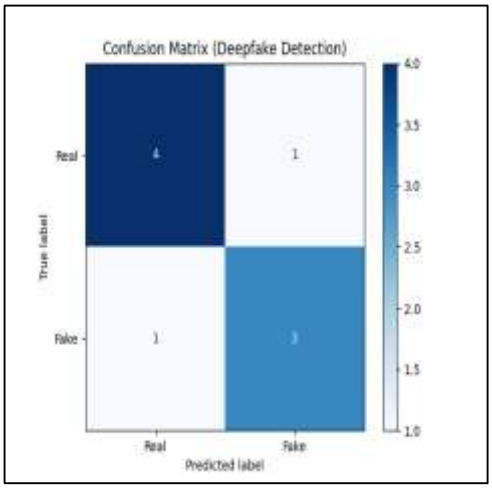


Figure: Confusion Matric of Proposed Model

The confusion matrics is used to evaluate the performance of the proposed classification model. It compares the actual values with the predicted values and provides a detailed insight into the model's performance**.**

It consists of four important components:

True Positive (TP): Number of fake images correctly classified as fake

True Negative (TN): Number of real images correctly classified as real

False Positive (FP): Number of real images incorrectly classified as fake

False Negative (FN): Number of fake images incorrectly classified as real

"Based on the confusion matrix, the obtained values are as follows:"

True Positive (TP) = 3

True Negative (TN) = 4

False Positive (FP) = 1

False Negative (FN) = 1

Performance Evaluation Metrix:

Accuracy

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

Accuracy=$\frac{3+4}{3+4+1+1} = \frac{7}{9}$

Calculated Accuracy = 77.8%

Precision

Precision = $\frac{TP}{TP+FP}$

Precision= $\frac{3}{3+1} = \frac{3}{4}$

Calculated Precision =75.0%

Recall

Recall =$\frac{TP}{TP+FN}$

Recall = $\frac{3}{3+1} = \frac{3}{4}$

Calculated Recall = 75.0%

F1-Score

F1-Score =$\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$

F1Score=$\frac{2 \cdot 0.75 \cdot 0.75}{0.75+0.75}$=0.75%

Calculated F1Score=75.0%

Table1.Model Prediction Results for Real and Fake Image

Real Image	**Fake Image**	**Wrong Prediction**
		
Predicted = Real	Predicted = Fake	Predicted = Fake
Actual = Real	Actual = Fake	Actual = Real

RESULTS AND DISCUSSIONS:

This section presents the outcomes of the proposed system .The model performance is evaluated using different parameters such as accuracy, loss, and classification results. The analysis helps to understand how well the system performs in identifying real and fake images.

Model Training Results:

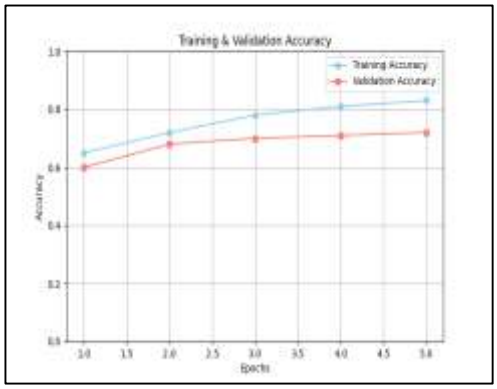


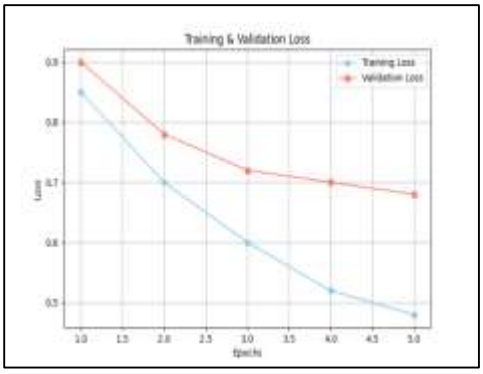
Figure:training&ValidationAccuracy

Accuracy and Loss Analysis:

The performance of the model is visualized using accuracy and loss graphs.

The accuracy graph shows that the model improves its prediction capability over time. The validation accuracy follows a similar trend, which means the model generalizes well on unseen data. The loss graph shows a continuous decrease, indicating that the model is minimizing errors during training.

A small difference between training and validation curves suggests that the model does not suffer from major overfitting.



Training Loss vs Validation Loss

The model was trained using a dataset consisting of real and fake face images. During the training process, the model gradually learned important features that help in distinguishing between real and fake images.

The training accuracy increased steadily with each epoch, while the validation accuracy also showed improvement. At the same time, the loss value decreased, which indicates that the model is learning effectively and reducing errors.

Confusion Matrix:

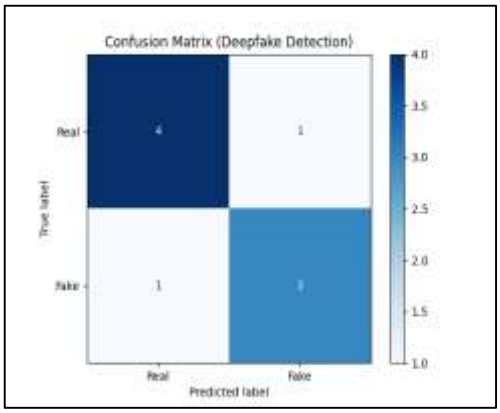


Figure: Confusion Matrix for Real and Fake Image Classification

The confusion matrix presented evaluates the performance of a Deepfake detection model by comparing the predicted labels against the true labels. It consists of four key values:

- **True Real, Predicted Real (4 cases):** These represent correctly identified authentic samples. The model successfully classified four real inputs as real, showing its ability to detect genuine content.
- **True Real, Predicted Fake (1 case):** This is a false negative. One real sample was incorrectly classified as fake, indicating the model occasionally misidentifies authentic content.
- **True Fake, Predicted Real (1 case):** This is a false positive. One fake sample was misclassified as real, which is critical because it reflects the model's vulnerability to deepfakes slipping through undetected.
- **True Fake, Predicted Fake (3 cases):** These are correctly identified deepfakes. The model successfully flagged three fake samples as fake

Performance Metrics:

To measure the effectiveness of the model, several evaluation metrics are used.

- Accuracy indicates the overall correctness of the model

- Precision shows how many predicted fake images are actually fake
- Recall measures how well the model detects all fake images
- F1-score provides a balance between precision and recall

These metrics together give a clear idea about the model performance.

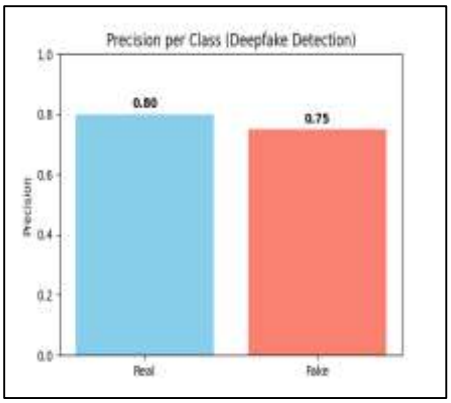


Fig:Precision

Precision per Class in Deepfake Detection

The bar chart illustrates the precision values achieved by the deepfake detection model for two categories: **Real** and **Fake**. Precision measures the proportion of correctly identified positive predictions out of all predictions made for a given class. In other words, it reflects how reliable the model is when it labels a sample as belonging to a particular class.

- **Real Class (Precision = 0.80):** When the model predicts that a sample is real, it is correct 80% of the time. This indicates a relatively strong ability to avoid misclassifying fake content as real, though there is still a margin of error.
- **Fake Class (Precision = 0.75):** For samples predicted as fake, the model is correct 75% of the time. This shows that while the system is fairly effective at identifying deepfakes, one out of four predictions of "fake" is actually incorrect.

Table2.PerformanceEvaluation Metrics of the Model

Metrix	Value
Accuracy	77.8%
Precision	75.0%
Recall	75.0%
F1-Score	75.0%

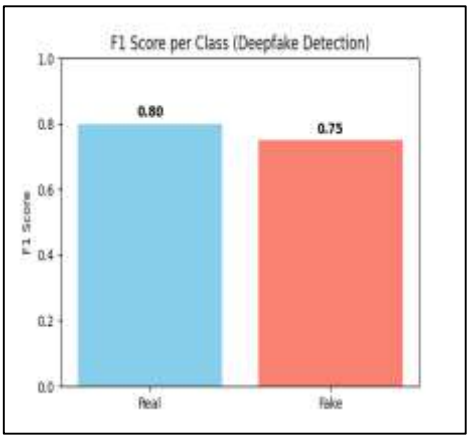


Fig: F1Score

F1 Score per Class in Deepfake Detection

The chart presents the F1 scores for the two classes in the deepfake detection task: **Real** and **Fake**. The F1 score is a balanced metric that combines both precision and recall, making it particularly useful when evaluating classification models where false positives and false negatives carry significant consequences. It is calculated as the harmonic mean of precision and recall, ensuring that both aspects of performance are equally considered.

- **Real Class (F1 = 0.80):** The model achieves an F1 score of 0.80 for real samples. This indicates that the system is fairly consistent in correctly identifying authentic content, with a good balance between precision (avoiding false alarms) and recall (capturing most real cases).
- **Fake Class (F1 = 0.75):** The F1 score for fake samples is slightly lower, at 0.75. This suggests that while the model is reasonably effective at detecting manipulated content, there is a higher tendency for errors compared to the real class, either by missing some fake samples or misclassifying real ones as fake.

Fig:Recall

Recall per Class in Deepfake Detection

The chart displays the recall values for the two classes in the Deepfake detection task: **Real** and **Fake**. Recall measures the proportion of correctly identified samples out of all actual samples belonging to a class. It emphasizes the model's ability to capture true positives and avoid missing relevant cases.

- **Real Class (Recall = 0.80):** The model successfully identifies 80% of the real samples. This indicates that most authentic content is correctly recognized, though a small portion is misclassified as fake.

Accuracy: The model achieves an accuracy of about 77.8% meaning it correctly predicts roughly 77.8% of all cases. This indicates that the overall performance is relatively low and the model needs further improvement to make more accurate predictions.

Recall: The recall value is close to 75.0%, which shows that the model is not able to capture all the actual positive cases. Many true instances are being missed, suggesting that the model's ability to detect positives is weak.

Sample Prediction Results:

The trained model is tested on sample images to check its prediction capability.

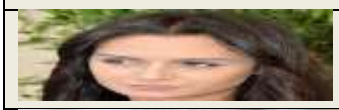
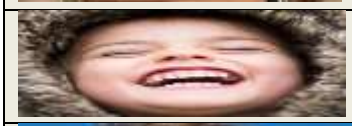
- **Fake Class (Recall = 0.75):** The recall for fake samples is slightly lower, at 75%. This means that while the system detects the majority of manipulated content, one out of four fake samples is missed and incorrectly labeled as real.

Precision: The precision is around 75.0& indicating that a portion of the positive predictions made by the model are incorrect. This reflects the presence of false positives and shows that the predictions are not highly reliable.

F1 Score: The F1 Score is also near 75.0%, combining both precision and recall into a single measure. The low score highlights that the model is not performing well in terms of both detecting positives and making accurate predictions, and therefore requires optimization.

The results show that most real images are correctly classified as real, and fake images are correctly identified as fake. This demonstrates that the model is able to extract meaningful features from the images

Table 3.Sample Prediction Results of Real and Fake Face Image

Input Image	Actual Label	Predicted Label
	Real	Real
	Fake	Fake
	Real	Fake
	Real	Real
	Fake	Real

Result Analysis:

The overall results indicate that the proposed system performs effectively in detecting real and fake face images.

The use of CNN helps in extracting important visual features such as edges and textures. Transfer learning further improves the accuracy and reduces the training time.

However, some limitations are observed. The model may produce incorrect results in cases where:

- Images are blurred
- Lighting conditions are poor
- Fake images are highly realistic

Despite these challenges, the model maintains good overall performance.

Conclusion:

The proposed deep learning-based system effectively classifies real and fake face images using CNN and transfer learning techniques. The model achieves an accuracy of 77.8%, demonstrating its capability in detecting manipulated images with reasonable reliability. The experimental results highlight the importance of deep learning in improving digital image security. Although the model performs well, further improvements can be made by optimizing parameters and adopting advanced architectures to enhance accuracy & generalization.

References:

[1] Abdul Kareem, H. S., & Altaei, M. S. M. Detection of deep fake in face images based on machine learning. Asian Journal of Engineering and Science Technology.*(2023).*
[2] Agarwal, A., & Ratha, N. (2024). Deep fake: Classifiers, fairness, and demographically robust algorithm. In Proceedings of the 18th International Conference on Automatic Face and Gesture Recognition (FG). *IEEE.*
[3] Agarwal, S., Rana, S., & Singh, G. A Novel Deep Learning Approach for Deep fake Image Detection. arXiv preprint arXiv:2301.04054. (2023)
[4] Bagde, A., Fand, S., Varma, K., & Gawali, A. Deep fake detection using deep learning. International Journal of Science, Engineering and Technology, 11(5).(2023)
[5] Guarnera, L., Battiato, S., & Giudice, O. (2020). Fighting deepfake by exposing the convolutional traces on images. IEEE Access, 8, 165769–165782.
[6] Kuribayashi, M., Malik, A., Abdullahi, S. M., & Khan, *A. N.* Deep Fake detection for human face images and videos: A survey. IEEEAccess,10,1–18.(2023)
[7] Manish, B., Manish, C., & Reddy, B. S. K. Deep fake detection on face images and videos using deep learning. Conference/Journal Name, volume (issue).(2024)
[8] Meskys, E, Liaudanskas, A., Kalpokiene , J., & Jurcys, P. Regulating deep fakes: Legal and ethical considerations. Journal of Intellectual Property Law & Practice, 15(1), 24–31(2024)
[9] Patel, Y., & Jain, M.Deepfake image detection using machine learning and deep learning.*Educational* Administration: Theory and Practice, 30(5), 14987–14993.

[10] Ramadhani, K. N., & Munir, R**.** A comparative study of deep fake video detection method. *In* Proceedings of the 3rd International Conference on Information and Communications Technology (ICOIACT) (pp. 1–6). IEEE.
[11] Velumani, V., Subramanian, M., Sekar, P., & Chand, H. M. Deep fake detection of images. *Preprint.*(2024)
[12] Xu, Y., Terhörst, P., Pedersen, M., & Raja, K. Analyzing fairness in deepfake detection with massively annotated databases. IEEE Transactions on Technology and Society, 5(1), 93–104.(2024)
[13] Zafar, A., Muhammad, Z., Iqbal, Z., & Kazim, M. Deepfake detection using deep learning: A unified forensic approach to detect AI-generated images and videos with fusion of eye, nose, and mouth landmarks.Preprint.
[14] L. Verdoliva, "Media forensics and deepfakes: an overview,"(2020).
[15] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," CoRR, vol. abs/1809.00888, (2018).
[16] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," (2018).
[17] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep learning for deepfakes creation and detection," ArXiv, vol. abs/1909.11573, (2019).
[18] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (dfdc) preview dataset," (2019).
[19] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, "Deep fake image detection based on pairwise learning," 01 (2020)
[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," (2015).