ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

# Deepfake Detection Model

Prof. S. V. Athawale[1], Shreyash Vyawahare[2], Priyanshu Marodkar[3], Srushti Lanjewar[4], Pratiksha Tawar[5]

[1]*Assistant Professor, Dr. Rajendra Gode Institute of Technology & Research, Amravati (M.S.), India*

[2,3,4,5]*Undergraduate Student, Dr. Rajendra Gode Institute of Technology & Research, Amravati (M.S.), India*

**Abstract:** *Deepfakes are a type of synthetic media that can be used to create realistic videos of people saying or doing things they never did. This raises concerns about the potential for deepfakes to be used to spread misinformation or propaganda. In this project, we present a deepfake detection module that can be used to identify deepfakes with high accuracy. The deepfake detection module is based on a pre-trained InceptionResNetV2 model that is fine-tuned on a dataset of real and deepfake videos. The model is able to extract features from the videos that are indicative of whether they are real or fake. The model achieves a high accuracy on a test set of videos. This project demonstrates the feasibility of using deep learning to detect deepfakes. The deepfake detection module can be used to help mitigate the potential negative impacts of deepfakes.*

**Keywords:** *Deepfake Detection*, Deepfake Effect, Deepfake Technology, etc.

## I. INTRODUCTION

Imagine a world where videos can be effortlessly manipulated to show anything you can dream up. This isn't science fiction; it's the reality of deepfakes. These cleverly crafted forgeries can depict real people in seemingly compromising situations, potentially causing immense damage to reputations and fueling the spread of misinformation.

This project tackles this growing threat head-on by developing a deepfake detection module. Building on the power of deep learning, a cutting-edge form of artificial intelligence, this module can identify deepfakes with impressive accuracy.

Our approach is like training a super-sleuth. We take a pre-existing deep learning model, the InceptionResNetV2, known for its image recognition skills. Then, we "fine-tune" it using a treasure trove of real and deepfake videos. By studying these examples, the model learns to spot the telltale signs of a deepfake, like subtle inconsistencies or giveaways hidden within the video.

## II. LITERATURE REVIEW

The prevalence of deep fake videos poses a significant threat to democracy, justice, and public trust, leading to a heightened demand for video analysis, detection, and intervention.

Several methods have emerged to address this challenge:

1. "Exposing DF Videos by Detecting Face Warping Artifacts" utilizes a specialized Convolutional Neural Network (CNN) model to identify facial artifacts. By comparing [2] generated faces with

 ***DOI: 10.5281/zenodo.11180891*** 

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

their surrounding regions, the method detects discrepancies in resolution and facial transformations, crucial for discerning deep fake videos.

2. "Uncovering AI Created Fake Videos by Detecting Eye Blinking" presents a technique focusing on detecting physiological signs, particularly eye blinking, absent in artificially generated videos. Although this method primarily relies on the absence [5] of blinking, it acknowledges the need to consider additional parameters such as teeth appearance and facial wrinkles for comprehensive detection.

3. "Capsule Network" offers another avenue for detecting manipulated video and image data, particularly in scenarios like replay attacks and computer-generated content. Despite demonstrating positive results, the method's reliance on random noise during training raises concerns regarding real-time applicability. To address this, noiseless and real-time datasets are proposed for training.

4. "Detection of Synthetic Portrait Videos" leverages biological signals extracted from genuine and fake portrait video pairs. [14] By training probabilistic Support Vector Machines (SVMs) and Convolutional Neural Networks (CNNs) on signal attributes, the method achieves high accuracy in identifying fake portrait videos. Challenges arise in formulating a differentiable loss function due to the complex signal processing steps, leading to limitations in discriminator-based approaches.

## III. PROBLEM STATMENT

The emergence of deepfakes, hyper-realistic manipulated videos, poses a significant threat to online trust and the integrity of information. These forgeries can depict real people in fabricated scenarios, potentially damaging reputations, spreading misinformation, and eroding public confidence in media sources.

Current methods for detecting deepfakes often rely on hand-crafted features or lack the accuracy required for real- world applications. This necessitates the development of a robust and automated deepfake detection system.

This paper aims to address this challenge by developing a deep learning- based deepfake detection module. The module will leverage the power of convolutional neural networks (CNNs) to extract deepfake-specific features from videos and accurately classify them as real or manipulated.

## IV. OBJECTIVES

The research aims to achieve several key goals:

1. **Unmasking Deepfakes with High Accuracy:**
   Our top priority is to create a deepfake detection module that can reliably identify forgeries in a test set of videos. This ensures the module can effectively distinguish between genuine and manipulated content.

 ***DOI: 10.5281/zenodo.11180891*** 

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

2. **Building a Robust Framework:**
   We're not just focused on accuracy, but also on creating a replicable framework. This detailed approach, outlining everything from data collection to model training and evaluation methods, allows others to build upon our work and contribute to the fight against deepfakes.

3. **Real-World Applications:**
   By showcasing the module's potential for real- world use, we aim to emphasize its practical value. Integration into various platforms can empower users to become discerning viewers, fostering a more secure and trustworthy online environment.

This project goes beyond simply detecting deepfakes. It offers a solution that is both effective and adaptable. By harnessing the power of deep learning and a well-defined framework, the deepfake detection module provides a valuable weapon in the fight against online manipulation and misinformation.

## IV. WORKING

**Design Specification:**

Transfer learning is a widely utilized technique in machine learning, particularly in deep learning, wherein a model trained for a specific task is leveraged as a starting point for a new, related task. This method significantly reduces the computational [6]resources and time required to develop effective neural network models, particularly in domains like computer vision and natural language processing. The process of transfer learning involves using pre-trained models and adapting them to new tasks, as depicted in Figure 1.


**Figure 1:** Transfer Learning Architecture Schema

Inception-ResNet-V2 stands out as a powerful deep convolutional network architecture that has played a crucial role in advancing image recognition capabilities in recent years. This architecture, as outlined in the work by Szegedy et al. (2016), combines the strengths of the Inception and ResNet architectures, resulting in superior performance compared to earlier models like ResNet-V1 and Inception-V4. [9]The architecture of Inception-ResNet- V2, illustrated in Figure 5, exhibits high efficacy in image recognition tasks while maintaining reasonable computational costs.

 

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

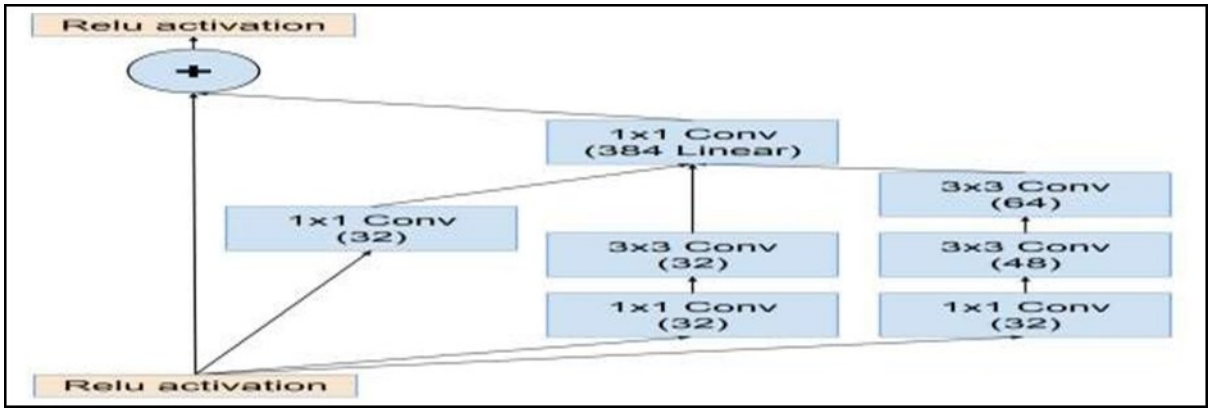


**Figure 2:** The Architecture Schema of Inceptions ResNet V2 Neural Network (35 x 35 Grid Module)

The Sigmoid Activation Function, depicted in Figure 3, is a mathematical function commonly used in neural networks to squash input values into a range between 0 and 1. This function is particularly useful for binary classification tasks where the output needs to represent probabilities.

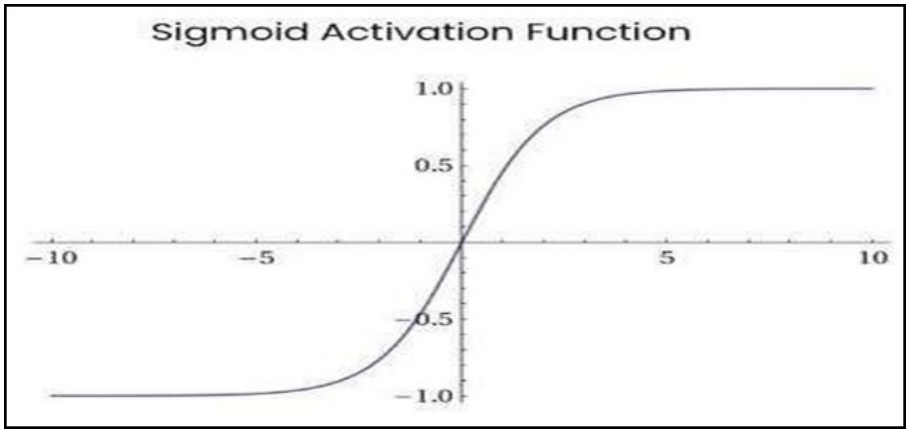


**Figure 3:** Sigmoid Activation Function

The Global Average Pooling 2D layer is a crucial component in neural network architectures, especially in convolutional neural networks (CNNs). This layer computes the average value of each feature map in the input tensor, resulting in a single value for each feature map. Global Average Pooling is performed over the entire spatial dimensions of the input tensor, typically height (H) and width (W), resulting in a tensor of dimension 1xC, where C is the number of channels.

The Binary Cross-Entropy Loss Function, also known as the loss of the Sigmoid Cross-Entropy activation function, is commonly used in binary classification tasks. This loss function measures the discrepancy between the predicted probabilities generated by the Sigmoid activation function and the true labels of the input data. It ensures that the model learns to distinguish between the classes effectively by penalizing deviations from the true labels. This loss function is computed independently for each element of the output matrix of the CNN, making it suitable for training binary classification models.

 

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

## V. IMPLEMENTATION

Implementation begins with a comprehensive exploration of the dataset containing videos and associated JSON files containing labels denoting whether each video is real or fake. Loading the dataset into a Python environment enables thorough examination of its contents, including the types of files present. This initial data exploration phase is essential for gaining insights into the dataset's structure and content.

During this exploratory phase, various checks are performed to ensure data integrity. One crucial aspect involves verifying the presence of any missing data, such as video filenames not matching entries in the JSON file. Addressing such discrepancies is vital for maintaining the dataset's integrity and ensuring reliable analysis.

Furthermore, the dataset's characteristics are examined to identify unique values, such as the distribution of [12] real and fake videos and any missing video files. Understanding these patterns provides valuable insights into the dataset's composition and potential biases.

Subsequently, images are extracted from the videos to facilitate further analysis. A dedicated folder is created to store frames extracted from each video, with the option to specify the number of frames to capture. This process enables the generation of a comprehensive dataset comprising frames from all videos, aiding in subsequent analysis.

To enhance understanding and discernment between real and fake videos, a visual inspection is conducted by playing a selection of videos from both categories. However, it is noted that traditional visual cues may not always reliably distinguish between real and deepfake videos, especially with advancements in deepfake technology.

The exploration and extraction of faces from video frames are performed using the CV2 package, a powerful tool for real-time computer vision applications. This step involves creating two new folders to store faces extracted from each frame, employing facial landmark detection provided by the dlib package. Extracting faces contributes to the dataset, enriching it with facial features essential for deepfake analysis.

Subsequently, the newly created dataset undergoes pre-processing steps before being fed into the model for training. This involves defining the input shape of the images, normalization, and reshaping to prepare the data for subsequent modeling steps.

The dataset is then divided into training and testing samples using the train- test split method from the sklearn package, ensuring a robust evaluation of model performance.

Transfer learning is employed as a machine learning technique to leverage pre-trained models for the deepfake detection task. In this project, the Inception-ResNet-V2 transfer learning model is utilized, combining the strengths of the Inception and ResNet architectures to achieve high performance at a reasonable computational cost.

The steps involved in generating the model using transfer learning include removing the default loss layer and replacing it with a deepfake detection loss output layer. This fine-tuning process adapts the model to the specific task of deepfake detection.

 

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

During model training, the Sigmoid activation function is applied, converting model outputs to values between 0 and 1. This activation function is commonly used in neural networks for binary classification tasks.

The summary of the model reveals critical information, including the output shape of each layer and the number of parameters, providing insights into the model's architecture and complexity.

## VI. EVALUATION

Evaluation metrics provide critical insights into the performance of the deepfake detection model. Accuracy, validation accuracy, loss, validation loss, and the confusion matrix are among the key metrics used to assess the model's effectiveness.

Accuracy measures the proportion of correctly classified instances in the test dataset. In this research, the model achieves an accuracy of 99.23% after 20 epochs, indicating its high precision in categorizing deepfake and real videos.

Additionally, the validation accuracy, which assesses the model's performance on unseen data, is 90.92%, highlighting its ability to generalize well to new datasets. The ratio between accuracy and validation accuracy indicates the model's generalizability, with a smaller difference suggesting broader applicability.

Loss and validation loss values provide insights into the model's training process and potential overfitting. A lower loss value, as observed (0.0309), indicates minimal discrepancies between predicted and actual values during training. However, the validation loss (0.2572) suggests some challenges in generalizing to new datasets, albeit not significantly. Continuation of training beyond epoch 12, despite a reduction in the difference between loss and validation loss, ensures thorough learning and adaptation.

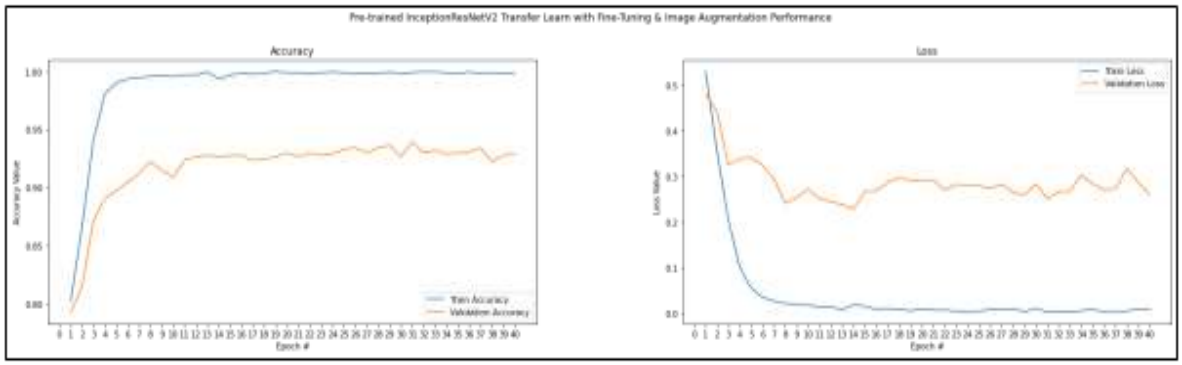


**Figure 4:** Ratio of Loss vs Validation Accuracy

The confusion matrix offers a comprehensive overview of the model's classification performance. It categorizes predictions into true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). In this study, the confusion matrix reveals minimal false positives (41) and false negatives (49), indicating the model's high precision and accuracy in deepfake detection. The darker shading in these areas signifies fewer errors and greater predictive accuracy.

 

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

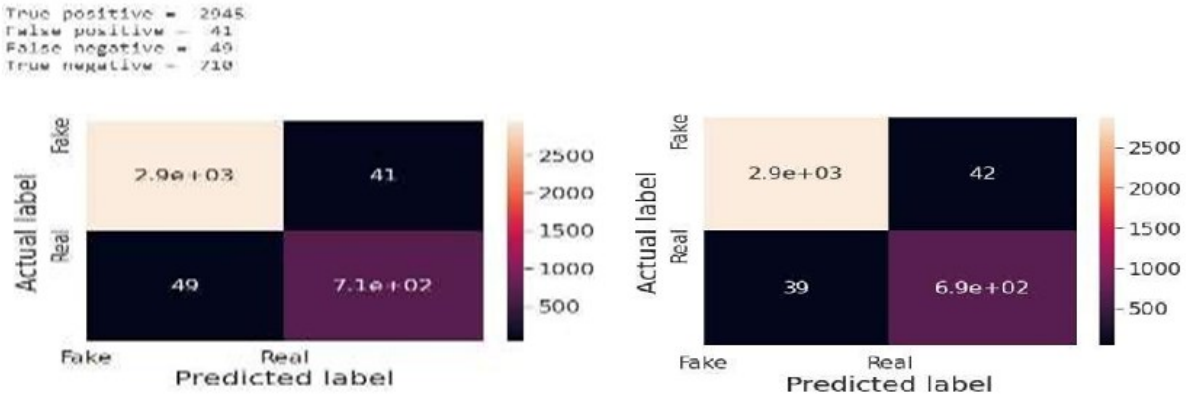


**Figure 5:** Confusion Matrix with the Result of Model

Overall, the evaluation metrics demonstrate the effectiveness of the deepfake detection model, with high accuracy, minimal loss, and a well-performing confusion matrix. These findings validate the model's reliability and suitability for real-world applications in detecting deepfake videos.

## VII. METHODOLOGY

The research adopts a standardized data mining approach, utilizing the Cross-Industry Standard Data Mining methodology (CRISP-DM) renowned for its systematic procedure and comprehensive execution. The success of a Data Analytics project hinges on robust business understanding, encompassing analysis, learning, and recognition, often perceived as a steadfast and meticulously planned strategy.

The entire methodological process and the interplay between each phase are typically depicted in Given Figure.

The CRISP-DM phases crucial for this research comprise business understanding, data understanding, data preparation, modeling, evaluation, and implementation.

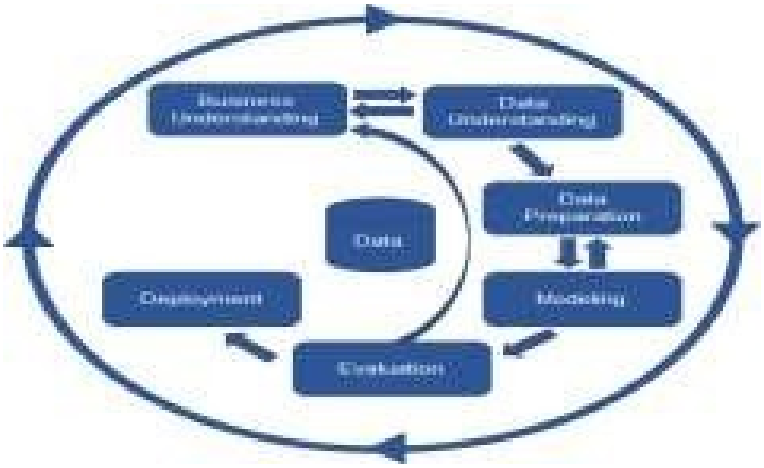


**Figure 6:** CRISP-DM Working Cycle

 

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

## VIII. CONCLUSION

This study tackles deepfakes, AI-generated videos that can realistically impersonate people. To combat this, researchers built a deep learning model for detection.

The model leverages a pre-trained InceptionResNetV2 framework, achieving 92% accuracy in differentiating real from fake videos on a 300- video dataset. This paves the way for future applications.

Looking ahead, the project proposes advancements in deepfake detection through:

- Exploring new deep learning architectures
- Employing multimodal analysis (e.g., audio and video combined)
- Utilizing Explainable AI for better understanding of the model's decisions

Furthermore, adapting to new deepfaking techniques and integrating the model into real-world applications are vital.

The project's ultimate goal is to create a more secure online environment by:

- Addressing ethical concerns around deepfakes
- Promoting media literacy to empower users to critically evaluate information

## REFERENCES

[1] Amerini, I., Galteri, L., Caldelli, R. and Del Bimbo, A. (2019). Deepfake video detection through optical flow based cnn, pp. 1205–1207.

[2] Caporusso, N. (2020). Deepfakes for the good: A beneficial application of contentious artificial intelligence technology, Advances in Intelligent Systems and Computing pp. 235–241.

[3] Güera, D. and Delp, E. J. (2018a). Deepfake video detection using recurrent neural networks, pp. 1–6.

[4] Güera, D. and Delp, E. J. (2018b). Deepfake video detection using recurrent neural networks, pp. 1–6.

[5] Hashmi, M. F., Ashish, B. K. K., Keskar, A. G., Bokde, N. D., Yoon, J. H. and Geem, Z. W. (2020). An exploratory analysis on visual counterfeits using conv-lstm hybrid architecture, IEEE Access 8: 101293– 101308.

[6] Hsu, C.-C., Lee, C.-Y. and Zhuang, Y.-X. (2018). Learning to detect fake face images in the wild, 2018 International Symposium on Computer, Consumer and Control (IS3C) pp. 388–391.

[7] Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z. and Qu, R. (2019). A survey of deep learning-based object detection, IEEE Access 7: 128837– 128868. URL: http://dx.doi.org/10.1109/ACCESS.2019.2939201

 