Inteligencia Artificial, 28(76) (2025), 20-39
doi: 10.4114/intartif.vol28iss76pp20-39

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Deepfake Detection in Manipulated Images/ Audio/ Videos: A Three-Stage Multi-Modal Deep Learning Framework

Leema Nelson[1*], Harshita Batra[2], and Radha P.[3]

[1*]Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India., leema.nelson@gmail.com
[2]Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India., batraharshita12@gmail.com
[3]Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India., pradha@mepcoeng.ac.in

Abstract The proliferation of deepfake content presents a significant threat to digital integrity and necessitates the development of efficient detection techniques. This study aims to establish a three-stage framework utilizing advanced deep learning models for multimedia datasets encompassing audio, video, and image data. The initial stage comprises an XceptionNet-based image deepfake detection model developed by providing its capacity to capture subtle artifacts and inconsistencies through depth-wise separable convolutions. This model, developed using the CelebA dataset, achieved an accuracy of 95.56 % for the image data. The second stage, focusing on audio deepfakes, employs a novel approach combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, selected for their capacity to process both the spatial and temporal aspects of audio data. The hybrid CNN and LSTM achieved an accuracy of 98.5 % on the DEEP-VOICE dataset. The third stage, addressing video-based deepfake detection, integrates the XceptionNet and LSTM networks, harnessing the strengths of both spatial and temporal analyses. This integrated approach yields an accuracy of 97.574 % across the Forensic++, DFDC, and Celeb-DF datasets. To address class imbalances in the datasets, class weighting is employed, assigning greater weights to the minority class during training, thereby enhancing the robustness of the model. This framework is used to develop an app for detecting deepfakes across images, audio, and video data. This study underscores the significance of deep learning architectures and comprehensive datasets for accurate deepfake detection across various media forms. By advancing detection methodologies, this research contributes to combating misinformation and safeguarding the authenticity of digital content, thus supporting the preservation of online ecosystems.

Keywords: Deepfake, Convolutional Neural Networks, Long Short-Term Memory, XceptionNet, Celeb dataset.

1. Introduction

The utilization of manipulated images, videos, and audio files has increased significantly with the advent of advanced artificial intelligence (AI) technologies, particularly deep neural networks (DNNs). Although images and videos can be altered in earlier times [1], contemporary techniques facilitate the

ISSN: 1137-3601 (print), 1988-3064 (on-line)

creation of highly realistic counterfeit human facial images [2], videos [3], and human voice mimicry [4]. The application of DNN-based techniques for face replacement in deepfakes has expanded. Common methods include autoencoders (AEs), Variational Autoencoders (VAEs) [5], and Generative Adversarial Networks (GANs) [2]. These methods combine or superimpose a source face image onto a target image. Recent advancements have enabled real-time voice cloning [4, 6], which is a network-based speech synthesis technique that produces high-quality speech using target speakers [4]. Notable deepfakes created for research include those of former US Presidents Barack Obama, Donald Trump, and George W. Bush [7], with remarkably accurate lip-syncing. Deepfakes present complex technological, social and ethical challenges. Their potential negative implications have been widely discussed in social media and state news outlets [8]. Tariq et al. [9] demonstrated the significant impact of deepfake impersonation on facial recognition technology. Given the increasing utilization of deepfakes to generate false information, from fabricated news to deceptive content such as celebrity pornography, the rapid evolution of deepfakes has prompted the academic community and technology industry to emphasize the automated detection of deepfake videos. The proliferation of deepfake technology has raised significant ethical, security, and privacy concerns. Consequently, developing deepfake detection techniques has become imperative, given the potential for misuse and harm. Numerous researchers have contributed to creating publicly available deepfake detection datasets [10-15]. Most of these datasets comprise authentic and manipulated false videos, as well as edited images of individuals' faces, specifically images substituted for those of another person. Currently, deepfakes are generated using various techniques [2, 16, 17]. CelebA, a well-established large-scale collection of facial attributes, contains approximately eight million attribute labels encompassing facial images with diverse poses and complex background environments. The Deepfake Detection Challenge (DFDC) [16] incorporates the generated cloned audio, deepfake videos, or a combination of both. The deepfake detection challenge [18] and FaceForensics++ (FF++) [19] pioneered extensive datasets that contain substantial quantities of deepfake videos. FF++ contains 5,000 videos, whereas DFDC comprises 128,154 videos, both produced using multiple deepfake generation techniques (FF++: 4, DFDC: 8). FF++ generates 5,000 deepfake videos by applying four deepfake generation models to a base set of 1,000 authentic YouTube videos. Two additional deepfake datasets, Deepfake Detection (DFD) [22] and FaceShifter [20], are incorporated into FF++. The deepfake detection challenge dataset was developed through collaborative efforts involving academic researchers, Amazon Web Services, Facebook and Microsoft. Eight distinct synthesis techniques are employed to create deepfake videos using the DFDC dataset recorded under various environmental conditions. Celeb-DF [12] was introduced in 2020, utilizing 500 authentic YouTube videos featuring 59 celebrities. A publicly accessible deepfake voice dataset called DEEP-VOICE is available [21]. The AUDIO directory contains raw audio files categorized into REAL and FAKE class directories. Each filename identifies the original speaker and altered voice. For feature extraction, the authors used the "DATASET-balanced. csv"file. The characteristics are derived from one-second audio segments, and the distribution of actual and fake samples was balanced using random sampling. This study investigates deep learning approaches for the automatic classification and identification of deepfake images, audios, and videos. The researchers trained deepfake detection models for visual, audiovisual, and audio-based media using three distinct datasets. For image-based detection, XceptionNet is trained on the CelebA dataset to detect altered facial images. For video-based content, XceptionNet and long short-term memory networks are trained using the FaceForensics++, DFDC, and Celeb-DF datasets. This methodology employs temporal modeling with frame-level feature extraction to identify sequential and spatial artifacts. For audio-based detection, the CNN and LSTM networks are trained on the DEEP-VOICE dataset, focusing on voice-cloning manipulations and synthetic speech. Each model demonstrates robust detection capabilities across diverse multimedia forms. Evaluations indicate high accuracy in distinguishing between authentic and fraudulent data.

2. Related Works

This section reviews the recent advancements in the manipulation of images, videos, and audio files using artificial intelligence methodologies. This study introduces a face-NeSt detection architecture that optimally selects multiscale features for final prediction [22]. It employs an adaptively weighted multiscale attentional (AW-MSA) module to ascertain the optimal proportion of multiscale features. Face-NeSt accentuates significant feature regions across spatial and channel dimensions both locally and globally. In contrast to the prevalent contemporary computer vision models, Face-NeSt is lightweight. It demonstrates superior performance on three publicly available benchmark datasets: FaceForensics++ (FF++), CelebDF, and DFDC, achieving AUC scores of 0.9823 on CelebDF, 0.9947 on DFDC, 0.9945 on DeepFake (FF++), 0.9905 on Face2Face (FF++), 0.9978 on FaceShifter (FF++), 0.9948 on FaceSwap (FF++), and 0.9548 on neuronal textures (FF++). A two-branch structural network, called the self-attention default face discrimination network (SADFFD) [23], was developed. A branch with cascaded multiself-attention mechanism (SAM) modules was integrated in parallel with EfficientNet-B4 (EffB4). The multi-SAM branch provides additional features that focus on crucial image regions to distinguish authentic and synthetic images. EffB4 was selected for its efficiency. Experiments on FaceForensics++, Celeb-DF, and SAMGAN3 datasets showed SADFFD's superior detection accuracy, achieving 99.01 % in FaceForensics++, 98.65 % in Celeb-DF, and 99.99 % in SAMGAN3. A unified network for the detection of FaceSwap (FS) and Face-Reenactment (FR) Deepfakes, termed AUFF-Net, was presented [24]. This approach uses spatial and temporal information from video samples to detect FS and FR. An Inception-Swish-ResNet-v2 model was introduced as a feature extractor for spatial information, whereas the Bi-LSTM measured temporal information. Three dense layers were added to create a discriminative feature-vector group. Experiments on FaceForensic++ achieved average accuracies of 99.21 % and 98.32 % for FS and FR, respectively.A lightweight machine learning based framework was developed to differentiate between authentic and spoofed audio recordings [25]. This method uses handcrafted audio features, including spectral, temporal, chroma, and frequency domain features. The ASVSpoof2019, FakeAVCelebV2, and In-The-Wild databases achieved 89 % accuracy on ASVSpoof2019, 94.5 % on FakeAVCelebV2, and 94.67 % accuracy, respectively. Explainability techniques elucidate the decision-making processes, enhance transparency, and identify crucial features for audio deepfake detection.

A hybrid-optimized deep-feature fusion-based deepfake detection (HODFF-DD) framework for videos was introduced utilizing a spotted hyena optimizer [26]. HODFF-DD is robust across ethnicities and lighting conditions, and detects deepfake videos produced using various techniques. It consists of two main components: a custom model with InceptionResNetV1 and InceptionResNetV2 and bidirectional long short-term memory (BiLSTM). Faces extracted from videos underwent frame-level feature extraction using the custom model, and the resulting feature sequences were used to train a BiLSTM for the binary classification of real and fake videos. The spotted hyena optimizer optimized the network weights during training. Evaluations on datasets like Kaggle's FaceForensics++ with techniques such as DeepFakes, FaceSwap, Face2Face, FaceShifter, and NeuralTextures, and FakeAVCeleb show the method's effectiveness, achieving over 90 % accuracy on subsets like DeepFakes, FaceSwap, and Face2Face. Using a graph neural network (GNN), An enhanced method for detecting deepfakes in films has been developed [27]. This technique splits detection into two stages: a four-block CNN stream and mini-batch graph convolution network stream. Three fusion networks FuNet-A, FuNet-M, and FuNet-C were fused in two phases. After 30 epochs, the accuracy of the model for various datasets was 99.3 %.This study employed various color spaces to enhance the detection of deepfakes [28]. They used two stages: a color-space-based forgery detection network, and a representative forgery learning stage with multicolor space reasoning. The forgery learning stage employed a forgery highlighting network, color-space transformations, and manipulation cue-boosting network. The forgery highlighting network found high-level semantic forgery clues and textural anomalies, the cue boosting network enhanced feature representation, and the color spaces provided benefits over RGB. They tested the technique on FaceForensics++, DFDC, and CelebDF datasets and found it effective in identifying falsified multimedia content in various color representations.

In this study, an adaptive blind watermarking technique was used to enhance the flexibility and resilience of deepfake image detection [29]. This approach embeds coefficients to ensure good image quality while fending off attacks by using mixed modulation and a sign-altered mean value. Additionally, blind adaptive deepfake detection with a tamper detection mean value adaptively detects relative positions in marginally altered or deepfaked images. The grey wolf optimizer and denoising autoencoder further improve the performance of the method through parameter optimization and watermark detection. This technology verifies the image owner and confirms face validity by adaptively embedding watermark information while maintaining the original facial image. This research focused on current deepfake detection models for plaintext faces [30]. However, sensitive data must be computed securely for practical application. The Secure DeepFake Detection Network (SecDFDNet) is presented as a solution. An additive secret-sharing technique for safe DeepFake face detection is presented. Furthermore, protocols for multiparty secure interactions, such as SecReLU, SecSigm, SecSpatial, and SecChannel, have been presented and shown to be secure with little space and communication complexity. By combining these secure protocols with a trained plaintext DFDNet, the SecDFDNet model outperforms several other models and achieves the same accuracy as the plaintext DFDNet.

A novel deepfake detection network can distinguish between high- and low-quality facial images generated using various techniques [31]. This framework combines a regular spatial stream with a frequency stream to address low quality images. Hierarchical supervision was employed to differentiate between actual and fraudulent images. This study created an MSCR-ADD by integrating multispace channel-representation learning [32]. This system combines channel-specific, channel-differential, and channel-invariant encoders for deepfake detection. Experimental results on four benchmark datasets show that MSCR-ADD outperforms the current state-of-the-art methods. Feature representations in channel-differential and channel-invariant spaces enable effective artifact identification in false audio by highlighting the distinctions and similarities between the channels in binaural audio. In this study, AVFakeNet incorporates auditory and visual modalities to enhance deepfake detection accuracy [33]. AVFakeNet is a Dense Swin Transformer Net (DST-Net) with input, output, and feature extraction blocks. The feature extraction block used a specially designed swine transformer module in which the input and output bhads had dense layers. Comprehensive experiments on five datasets, including audio, visual, and audio-visual deepfakes, along with a cross-corpora examination, demonstrated the efficacy and generalizability of this unified architecture. The findings demonstrate that the proposed framework successfully identifies deepfake videos by analyzing both audio and visual streams.

3. Materials and Methods

This section discusses the materials used in the development of deepfake detection models, including images, audio, and videos. An XceptionNet-based image-deepfake detection model is developed using the CelebA dataset obtained from the Kaggle repository. The CNN with LSTM-based audio deepfake detection model is developed using the DEEP-VOICE dataset, while the XceptionNet with LSTM-based deepfake video model is developed using FaceForensics++, DFDC, and Celeb-DF datasets obtained from the Kaggle repository.

3.1. Dataset Description

Various deepfake datasets are used to test and identify manipulated media. They provide samples of fake and real images, videos, and audio from various sources, such as YouTube and public websites. Table 1 provides an overview of the deepfake datasets used in this study. It contains details on the dataset name, year of release, number of fake and real samples, source, and media type. This comprises datasets, such as CelebA, FaceForensic++, DFDC, Celeb-DF, and DEEP-VOICE.

Cuadro 1: Deepfake datasets details

Dataset	**Release Year**	**Fake : Real Ratio**	**Source**	**Type**
CelebA	2021	11,509 : 8,000	Celebrity images	Image
FaceForensic++	2019	4K : 1K	YouTube	Video
DFDC	2020	$> 100K :> 100K$	Celebrity videos	Video
Celeb-DF	2019	5,639 : 590	YouTube	Video
DEEP-VOICE	2020	1,000 : 2,500	Public	Audio

a) **CelebA:** The CelebA dataset, obtained from Kaggle, comprises 11,509 real images and 8,000 fake images, rendering it a valuable resource for training and testing face-detection models [10]. It is extensively utilized for facial attribute recognition, including features such as smiling, wearing glasses, and hair coloration. The dataset encompasses 202,599 facial images of various celebrities, featuring 10,177 unique individuals, five landmark points, and 40 binary attribute annotations.

In addition, the CelebA dataset exhibits variations in pose, background clutter, and expressions, presenting challenges for deep-learning-based facial analysis.

b) **FaceForensic++:**FaceForensics++ is an extensive dataset designed to facilitate the detection of manipulated facial images and videos [19]. The collection contains high-quality real and fake video footage developed using deep learning-based face manipulation techniques, including DeepFakes, Face2Face, FaceSwap, and NeuralTextures. Furthermore, the dataset incorporates raw and compressed videos, which may prove beneficial for assessing the robustness of models against varying compression levels. Researchers have extensively employed FaceForensics++ to train deepfake models and enhance real-time facial-content identification models.

c) **DFDC:** The DFDC dataset was introduced on Facebook in collaboration with industry experts [18]. This dataset comprises over 100,000 real and synthetic videos, and deep learning models are utilized to create highly realistic facial manipulations. Subjects within the dataset were captured in diverse lighting and background settings, rendering it one of the most challenging datasets for training deepfake detection models. The DFDC dataset plays a crucial role in the development of robust AI solutions to identify manipulated media.

d) **Celeb-DF:**Celeb-DF is a deepfake video dataset designed to enhance the performance of deepfake detection [13]. The dataset encompasses over 590 real videos and 5639 deepfake videos generated using advanced synthesis techniques, resulting in highly realistic facial expressions, lip movements, and eyeblinks. It addresses challenges in deepfake detection, such as the reduction of visual artifacts commonly observed in synthetic videos. High-resolution video samples from Celeb-DF provide an excellent benchmark for evaluating the efficacy of the detection algorithms.

e) **DEEP-VOICE:** The DEEP-VOICE Kaggle dataset incorporates both authentic human speech and its corresponding deepfake audio tracks [21]. This dataset facilitates the training of models to distinguish between authentic and synthetic voices, and comprises artificially altered speech examples based on AI approaches in voice synthesis. The resulting high-quality alterations in human speech through AI provide speech samples via voice synthesis techniques to develop robust mechanisms against a range of increasing cyber fraud risks associated with audio deepfakes, which may be linked to misinformation campaigns or identity spoofing attempts.

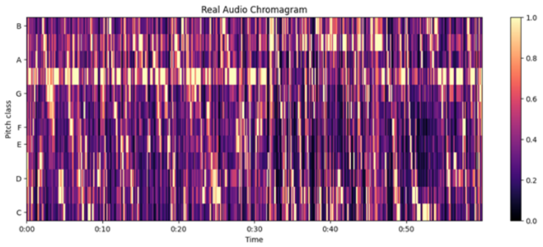


Figura 1: Pitch class of real audio chromogram over time.

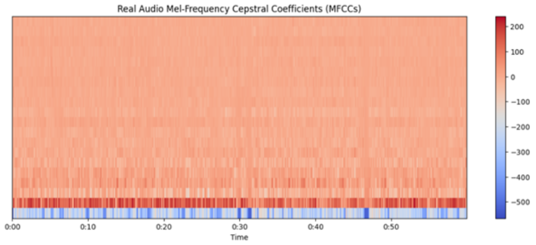


Figura 2: Frequency of real audio Mel-frequency cepstral coefficients over time.

Figure 1 illustrates the chromogram of authentic audio, demonstrating the temporal variation in pitch classes. In real audio recordings, this variation typically exhibits gradual progression and adheres to natural harmonic structures, particularly in speech or musical compositions. Authentic audio generally displays consistent harmonic relationships between the pitches over time. If artificially generated audio fails to accurately replicate natural pitch patterns, it indicates manipulation or inadequate synthesis. Subtle, unnatural alterations in pitch or abrupt variations serve as indicators of audio inauthenticity. Figure 2 shows that mel-frequency cepstral coefficients (MFCCs) represent the human auditory system's perception of frequency content in audio, compressing it into features that emphasize perceptually significant frequencies. In real audio, the MFCCs exhibit smooth characteristics and consistent transitions. In the artificially generated audio shown in Figure 4, MFCCs do not follow natural frequency patterns,

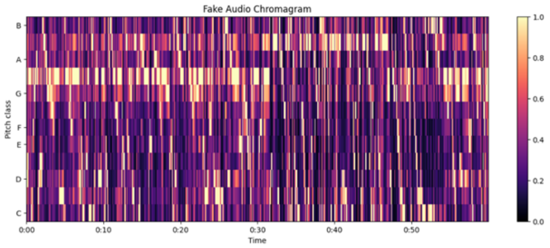


Figura 3: Pitch class of fake audio chromogram over time.

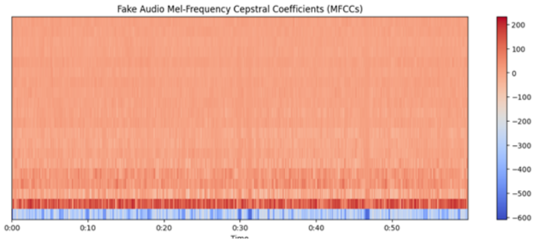


Figura 4: Frequency of fake audio Mel-frequency cepstral coefficients over time.

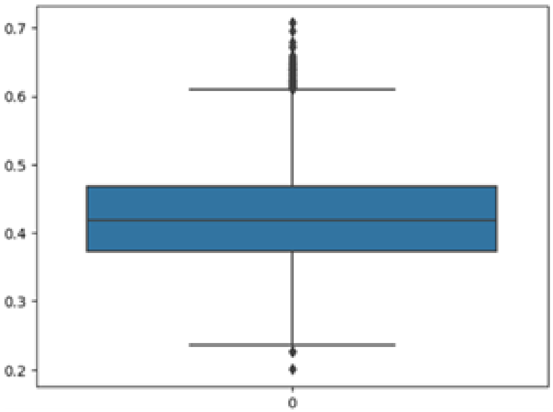


Figura 5: Feature chroma stft plotted over time and pitch class.

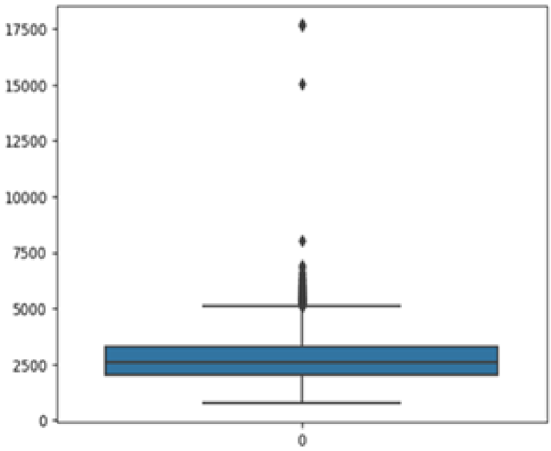


Figura 6: Spectral centroid values over time

particularly if the generative model fails to produce realistic speech or other acoustic features. A comparative analysis of authentic and artificially generated MFCCs can identify unnatural frequency variations or anomalies in synthetic audio. Figure 3 presents a chromogram for synthetic audio, providing insight into the pitch classes of the artificially generated audio clips. In numerous instances, deepfake audio exhibits difficulty in maintaining the smooth harmonic structure characteristic of authentic audio, resulting in erratic pitch shifts or unnatural-note transitions. Abrupt changes between pitch classes or irregular tonal distributions suggest the synthetic nature of audio. Chroma refers to 12 distinct pitch classes in Western music, irrespective of the octave. An octave refers to the interval between one musical pitch and another pitch, which is either double or half its frequency. The chroma_stft feature, derived from the short-time Fourier transform (STFT) of an audio signal, represents the energy distribution among the 12 pitch classes over time. It is effective in music analysis for tasks, such as chord identification, harmonic structure detection, and tonality differentiation. The time-frequency representation of the audio signal aggregates the energy into bins corresponding to the 12 pitch classes, disregarding octave differences. This provides a concise method for visualizing the temporal evolution of musical elements, making it suitable for comparing the tonal structures of authentic and synthetic audio samples. Figure 5 illustrates the chroma_stft feature over time, demonstrating the energy distribution across different pitch classes throughout the audio signal duration. This visualization facilitates the identification of harmonic and tonal structures in real audio. Figure 6 depicts the spectral centroid values over time, indicating variations in the ”brightness.ºf the audio signal. A higher centroid typically corresponds to a brighter sound, whereas a lower centroid indicates darker tonal quality. Figure 7 displays the deepfake detection challenge picture dataset from the Kaggle repository, whereas Figure 8 displays the distribution of video frames based on the width intervals. The evaluation results for several face-detection software packages with different image resolutions are presented in Table 2.

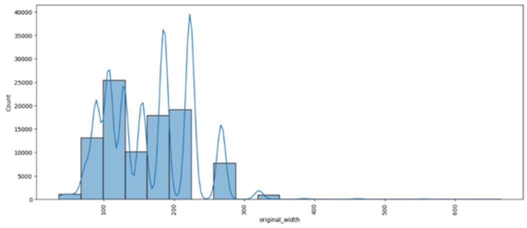


Figura 7: Distribution of video frames by width intervals

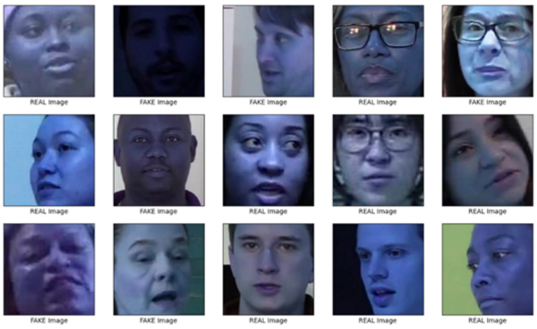


Figura 8: Deepfake detection challenge image dataset

Cuadro 2: Performance of different face detection packages under varying image resolutions.

Package	**FPS (1080 × 1920)**	**FPS (720 × 1280)**	**FPS (540 × 960)**
Facenet-pytorch	12.97	20.32	25.50
Facenet-pytorch (non-batched)	9.75	14.81	19.68
dlib	3.80	8.39	14.53
mtcnn	3.04	5.70	8.23

4. Proposed Methodology

This research aims to address the challenges introduced by deepfake media in audio, video, and image formats. To reduce the dissemination of misinformation and preserve digital integrity, this framework proposes comprehensive detection and mitigation strategies utilizing advancements in artificial intelligence, machine learning, and statistical learning. A multimodal framework for determining the authenticity of media content, including audio, videos, and images, is shown in Figure 9. The initial phase involves data collection, which encompasses the acquisition of raw data from three sources: audio, video, and images. These serve as inputs for the subsequent stages. Each data type undergoes distinct processing during the data preprocessing phase. To ensure consistency in size, scaling, and format, image data are first subjected to face detection, which isolates the facial regions, followed by normalization. Mel Frequency Cepstral Coefficients (MFCC) are extracted from audio data to capture relevant acoustic features. Video data are processed by extracting frames, identifying faces within the frames, and executing audio-video synchronization to ensure the alignment of the audio and visual streams. During the model training phase, predictive models are constructed using preprocessed data. XceptionNet, a deep learning architecture, is employed for feature extraction from the image data, with a softmax layer utilized for classification. For audio data, feature extraction is followed by classification using a Recurrent Neural Network (RNN). Video data analysis involves XceptionNet for frame-level feature extraction and Bi-LSTM for temporal feature modeling. Dynamic Time Warping (DTW) is employed to assess audio-video synchronization and ensure temporal coherence of the media. In addition, a softmax layer is utilized for video feature classification. The trained model is subsequently employed to categorize input media as either authentic or fraudulent, ensuring a comprehensive and reliable detection process across diverse data modalities.

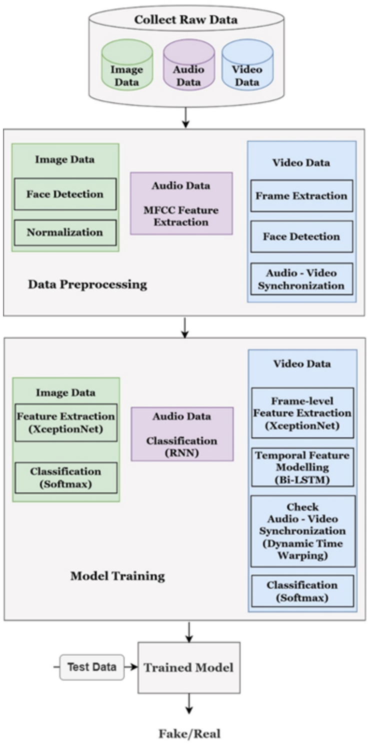


Figura 9: Proposed Methodology

4.1. Image Deepfake Detection

Image deepfake detection focuses on identifying manipulations that are often created using generative models, such as Generative Adversarial Networks. The challenge lies in detecting subtle manipulations such as face swaps, altered expressions, or synthetic creations that are visually convincing. Techniques include the analysis of inconsistencies in lighting, facial landmarks, and texture patterns. Deep learning models are widely used to identify these anomalies and are trained on large datasets of real and fake images to enhance their ability to distinguish between them. This study develops an XceptionNet-based image deepfake detection model using the CelebA dataset.

4.1.1. XceptionNet Image Deepfake Detection Model

XceptionNet is a sophisticated convolutional neural network architecture developed by Google researchers in 2016. XceptionNet is a modified version of the inception architecture that incorporates depth-wise separable convolutions to enhance the performance and reduce the model's parameter count. The XceptionNet architecture comprises three stages: entry, middle, and exit stages. With 71 layers, including 36 convolutional layers, 3 fully connected layers, and additional auxiliary layers for regularization and training purposes, XceptionNet provides a robust framework for image classification tasks. The input to the XceptionNet model is an image with dimensions of 299 × 299 × 3, where 299 represents the width and height of the image, and 3 denotes the number of color channels (RGB). The input image undergoes normalization and is subsequently processed through convolution layers for feature extraction. The architectural structure comprises three distinct stages: entry, middle, and exit, with the middle stage incorporating a series of depth-wise, separable convolutions. The entry stage focuses on reducing the spatial dimensions of the image while increasing the number of filters. This component comprises several convolution layers, followed by max-pooling for downsampling. The convolutional layer output is calculated using Equation (1).

$$Z^{(l)} = f\left(W^{(l)} * X^{(l-1)} + b^{(l)}\right) \tag{1}$$

Where $Z^{(l)}$ is the output of the l-th layer, $W^{(l)}$ represents the weights of the l-th convolutional filter, $X^{(l-1)}$ is the output from the previous layer, $b^{(l)}$ is the bias term, $*$ denotes the convolution operation, and $f(\cdot)$ is the non-linear activation function, usually ReLU. Max-pooling is applied to reduce the spatial dimensions and is calculated using Equation (2).

$$Y_{i,j} = \text{máx}(X_{i:i+k,j:j+k}) \tag{2}$$

Where k is the pooling size. The middle stage comprises a series of depth-wise separable convolutional layers designed to capture deeper and more abstract features. Multiple stacked separable convolution layers are present. Each layer consists of depthwise and pointwise convolutions. Depth-wise convolution applies convolutions over each channel independently, and pointwise convolution combines the outcomes of depthwise convolution. In the exit stage, the features undergo upsampling to restore the original resolution, followed by global average pooling, and fully connected layers to perform classification. Global average pooling reduces each feature map to a single value by averaging and is calculated using equation (3).

$$Y_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j,c} \tag{3}$$

Where H and W denote the height and width of the feature map, respectively. The fully connected layer applies a linear transformation to the pooled features and is calculated using Equation (4).

$$Z = W^{(L)} X^{(L-1)} + b^{(L)} \tag{4}$$

Where $W^{(L)}$ and $b^{(L)}$ are the weights and bias of the final fully connected layer, and $X^{(L-1)}$ is the output from the previous layer. The final fully connected layer applies a softmax function to obtain class probabilities using Equation (5).

$$P(y = k \mid X) = \frac{e^{Z_k}}{\sum_{j=1}^{K} e^{Z_j}} \tag{5}$$

Where, Z_k is the score for class k and K is the total number of classes. The model is trained using **the cross-entropy loss** for classification and is calculated using (6).

$$L = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log(P(y_i = k \mid X_i)) \tag{6}$$

Where: N is the number of training samples, $y_{i,k}$ is the true label for sample i and class k, and $P(y_i = k \mid X_i)$ is the predicted probability for class k. This XceptionNet architecture possesses approximately 22 million trainable parameters, enabling it to capture and learn complex features from data.

4.2. Audio Deepfake Detection

Audio deepfake detection aims to identify artificially altered speech typically generated using advanced text-to-speech models. These synthetic audio clips can replicate the voices, intonation, and accent of individuals, making detection challenging. Detection methodologies analyze features, including acoustic patterns, frequency inconsistencies, and unnatural speech pauses. Deep learning models focus on detecting temporal inconsistencies and abnormal voice signal fluctuations to differentiate authentic audio from deepfakes.

4.2.1. Hybrid CNN with LSTM Audio Deepfake Detection Model

A hybrid model combining a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks is highly suitable for audio deepfake detection because of its capacity to process both the spatial and temporal aspects of audio data. CNNs demonstrate high efficiency in extracting spatial features from spectrograms or other forms of audio representation, capturing significant patterns within both frequency and time domains. These features frequently reveal subtle anomalies such as unnatural frequencies, inconsistencies in pitch, or irregular energy distributions, which are typically introduced during audio synthesis. LSTM networks are structured to address sequential data, rendering them well suited for analyzing temporal dependencies within audio frames. By inputting CNN-derived high-level features into the LSTM layers, the model can comprehend the temporal evolution of these spatial features, enabling sensitivity to the detection of temporal anomalies, such as unnatural pauses, inconsistent voice modulation, or disrupted speech rhythms. This effective collaboration is achieved through a combination of feature extraction via a CNN and sequential analysis via LSTM, thereby providing a robust framework that distinguishes real audio from fake audio.

Raw audio data are converted into a melspectrogram, which represents the frequency content of the signal as a function of time. The mel-spectrogram represents the audio signal in terms of the time, frequency, and amplitude. The Mel-spectrogram is calculated using Equation (7).

$$S_{\text{mel}}(t, f) = \text{mel}(\text{STFT}(x(t))) \tag{7}$$

Where, $x(t)$ is the raw audio signal over time, STFT is the Short-Time Fourier Transform, S_{mel} is the mel-scaled spectrogram, t is time, and f is frequency in the mel scale. Figure 11 and 13 show the melspectrograms of real and fake audio signals over time. The mel-spectrogram image is given as the input to the CNN. The CNN extracts spatial features from the mel-spectrogram, including edges and frequency patterns. The convolution operation applies filters to detect these patterns using equation (8).

$$\text{Conv}(x) = x * w + b \tag{8}$$

Where, x is the input image mel-spectrogram, w is the filter (weights), b is the bias term and $*$ is the convolution operation. Max-pooling is used to down-sample the feature maps, thereby preserving the most salient features while reducing dimensionality utilizing equation (9).

$$\text{MaxPool}(x) = \text{máx}(\text{pool region}(x)) \tag{9}$$

The CNN output is flattened into a one-dimensional vector before being input into the LSTM. The LSTM receives this flattened output as the input sequence, with each time step representing an audio frame. CNN extracts high-level spatial features, whereas LSTM processes temporal dependencies by capturing information across time frames. After the LSTM layer, a dense layer is used for the binary classification of the real and fake audio. During the model training, the loss function is used as the binary cross-entropy for classification using equation (10).

$$L(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{10}$$

Where y is the actual output (0 for real, 1 for fake), and $\hat{y}$ is the predicted output.

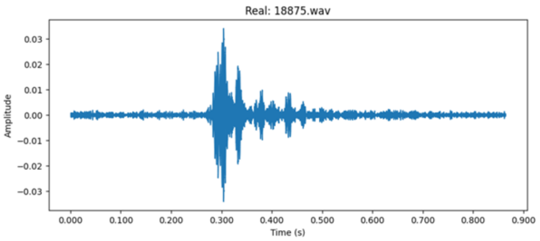


Figura 10: Amplitude of real audio signal over time.

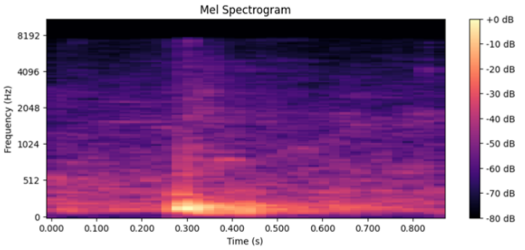


Figura 11: Frequency content of real audio signal over time.

Figure 10 illustrates the variation in amplitude of real audio. In real audio, amplitude fluctuations typically exhibit smooth transitions, reflecting natural speech patterns, in which different sounds have varying loudness levels. Conversely, Figure 13 shows that fake audio amplitude patterns appear irregular or unnaturally consistent. Fake audio lacks the natural dynamic characteristics of real speech or exhibits

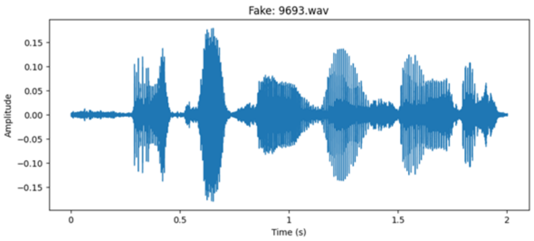


Figura 12: Amplitude of fake audio signal over time.

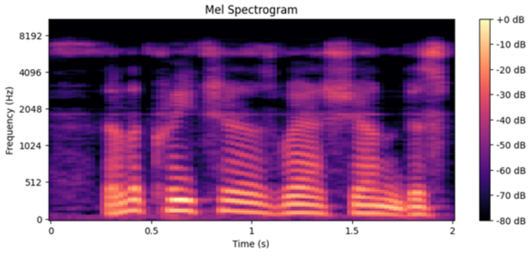


Figura 13: Frequency content of fake audio signal over time.

artifacts such as abrupt volume changes, indicating its fake nature. Figure 11 shows the spectrogram for authentic audio, displaying the energy distribution across frequencies over time. The real audio typically exhibits smooth frequency transitions and distinct harmonic patterns, particularly in speech and music. In contrast, Figure 12 shows that deepfake audio displays irregular frequency content in its spectrogram, such as absent harmonics, unnatural frequency band shifts, and unexpected noise. These frequency content deviations provide compelling evidence of the fake nature of audio. The hybrid approach is particularly synergistic for addressing the complex nature of audio deepfakes, capturing both local patterns and long-term dependencies to enhance the detection accuracy and reliability in real-world applications.

4.3. Video Deepfake Detection

Video deepfake detection involves identifying manipulated videos, in which faces and gestures are synthetically generated or modified. These manipulations are often created through face swapping, which makes it difficult to discern authenticity. Detection methods analyze inconsistencies in facial expressions, blinking patterns, and lip-sync mismatches. Advanced deep learning approaches capture spatial and temporal features, helping detect irregularities in visual information across video frames.

4.3.1. XceptionNet with LSTM Video Deepfake Detection Model

The hybrid model combining XceptionNet with LSTM networks presents a robust approach for video deepfake detection, providing the strengths of both spatial and temporal analyses. XceptionNet, a deep convolutional neural network, is proficient in extracting fine-grained spatial features from individual video frames. Its depth-wise separable convolutions enable the capture of subtle visual artifacts, including unnatural textures, facial irregularities, or blending errors that may be introduced by deepfake algorithms. These spatial features provide critical information regarding frame-level inconsistencies. Conversely, LSTM networks are suitable for modeling temporal dependencies across sequential data. While processing the spatial features extracted by XceptionNet, the LSTM layer learns the temporal patterns in video frames, such as anomalies in facial movements, lip synchronization, or unnatural transitions. This allows the hybrid model to detect both spatial inconsistencies in single frames and temporal anomalies across sequences, thereby enhancing its robustness against advanced deepfake techniques. Each video is divided into individual frames at a fixed rate of 30 fps. Each frame is input to XceptionNet for spatial feature extraction using Equation (11).

$$F_n = \text{Frame}(t_n), \text{ where } t_n = \frac{fps}{n} \tag{11}$$

where, F_n is the n-th frame, t_n is the time stamp of the n-th frame, fps is the frames per second. Each frame is resized to the input size required by the XceptionNet. XceptionNet utilizes depth-wise separable convolutions to reduce the computational complexity while preserving spatial information. This approach enables the capture of fine-grained spatial features from each frame, such as facial irregularities or artifacts introduced by deepfake generation. The depth-wise separable convolution is represented by Equation (12).

$$\text{Conv}(X) = \text{Depthwise}(X, W_d) + \text{Pointwise}(X, W_p) \tag{12}$$

Where, X is the input, W_d and W_p are the depthwise and pointwise filters, respectively. Each frame was analyzed using the XceptionNet pretrained model on ImageNet to extract spatial features. These features constitute a lower-dimensional representation of the frame and serve as an input to the LSTM. Following the extraction of spatial features from all frames in a video, these features are sequentially arranged and input into the LSTM layer. LSTM networks analyze the temporal dependencies in sequential data and examine the evolution of spatial features extracted from video frames over time. LSTM detects inconsistencies in facial movements, unnatural transitions, and synchronization errors across frames. The model architecture involves processing each frame through XceptionNet to extract spatial features and subsequently transmitting these features through LSTM to detect temporal dependencies. This hybrid architecture is particularly well suited for addressing the challenges posed by deepfakes, which are characterized by subtle spatial artifacts and compromised temporal coherence. The integration of the XceptionNet and LSTM capabilities in this model facilitates improved detection accuracy and demonstrates enhanced suitability for real-world video deepfake detection tasks, where both spatial and temporal cues are of significant importance.

5. Results and Discussion

The use of advanced deep learning models, such as XceptionNet, and their integration with Long Short-Term Memory networks have significantly improved deepfake detection capabilities. These models employ state-of-the-art methods to identify audio deepfakes and detect manipulated media content including video and image alterations with high accuracy. XceptionNet, a convolutional neural network, is considered one of the most effective approaches for identifying image-based modifications, particularly in deepfake images. It uses depth-wise separable convolutions, enhancing its ability to focus on fine-grained features and subtle artifacts introduced during manipulation. By isolating pixel-level discrepancies, XceptionNet effectively detects altered content. It has shown reliability in image-based deepfake detection, with accuracy rates often ranging between 95 % and 96 %. XceptionNet, trained on datasets like

CelebA, has exhibited remarkable performance in recognizing face-swapped images and detecting manipulation inconsistencies. Its capacity to extract fine-grained information is crucial for combating deepfake content. Combining XceptionNet and LSTM networks extends the deepfake detection capabilities to video-based media. While XceptionNet performs spatial analysis by detecting frame-specific pixel-level anomalies, LSTM networks capture sequential temporal changes across video frames. This hybrid approach ensures a comprehensive examination of video alterations by detecting both temporal and spatial artifacts. Datasets such as Celeb-DF and DeepFake Detection Challenge have shown that this combined method yields accuracy rates exceeding 97 %. The integration of LSTM temporal modeling with XceptionNet spatial precision provides a robust approach for identifying deepfakes in dynamic video content. Convolutional Neural Networks (CNNs) and LSTM networks have demonstrated effective synergy for audio deepfake detection. This architecture combines the spectral feature detection capabilities of CNNs for audio data with the sequential pattern recognition abilities of LSTMs. This enables the detection of temporal inconsistencies and audio artifacts, which are crucial for identifying deepfake audio. CNN+LSTM architectures play a vital role in detecting manipulated audio content, such as voice cloning and synthetic speech, with an accuracy rate of 98 %.

Figure 14 shows the time-domain waveform, with the x-axis showing the time in seconds and the y-axis representing the audio signal amplitude. The waveform illustrates temporal variation in sound amplitude, with higher peaks indicating increased intensity and lower troughs showing diminished acoustic energy. This representation is used to analyze the temporal dynamics of the audio signals. Figure 15 shows a mel-frequency cepstral coefficient (MFCC) visualization, with the x-axis showing time and the y-axis depicting MFCC coefficients. Each value corresponds to the magnitude of a particular MFCC coefficient at a specific time, as indicated by color intensity variations. This visualization captures the short-term power spectrum of a sound signal, useful in speech and audio processing tasks for analyzing temporal variations in frequency content. Figure 16 shows a spectrogram, with the x-axis denoting time and y-axis showing frequency on a mel scale. The color intensity indicates the power magnitude of the signal at each frequency and time point. This visualization captures the energy distribution across frequencies over time and is valuable for analyzing audio signals in tasks, such as speech recognition, music analysis, and environmental sound classification. Figure 17 shows the chroma feature visualization, with the x-axis

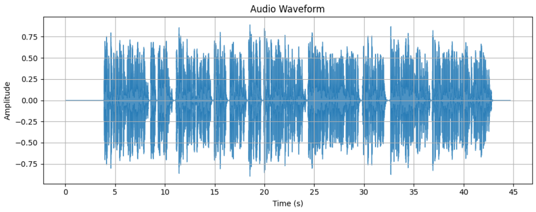


Figura 14: Time-domain waveform for amplitude variation over time

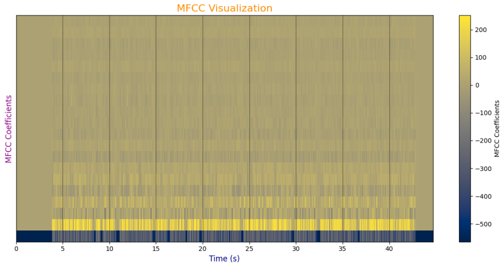


Figura 15: MFCC visualization for mel-frequency cepstral coefficients (MFCC) over time

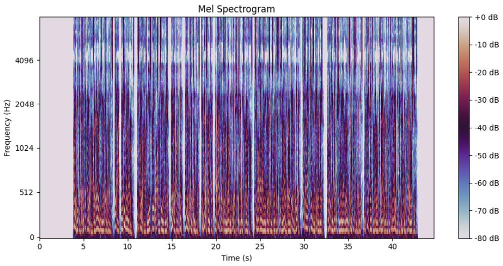


Figura 16: Mel-spectrogram for frequency distribution mel scale over time

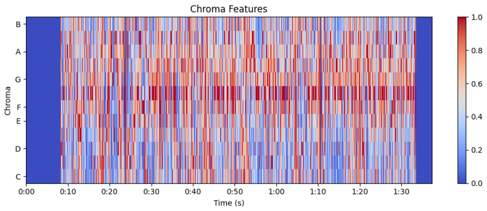


Figura 17: Chroma feature visualization for pitch class intensity over time

showing time and y-axis depicting 12 distinct chroma values corresponding to 12 pitch classes in western music. Each color indicates the intensity of a particular pitch class at a given time. This visualization is useful in music processing for analyzing harmonic and melodic content, highlighting the prominence of different notes or chords over time. Figure 18 shows the zero-crossing rate (ZCR), with the x-axis denoting time and the y-axis showing the rate of zero crossings. ZCR indicates how often the audio signal changes signs within a given timeframe. This feature is used in audio analysis to differentiate between sound types, with higher ZCR values indicating more rapid signal fluctuations. Figure 19 shows the Spectral Centroid over time, with the x-axis showing time and the y-axis depicting the spectral centroid in terms of frequency. The spectral centroid indicates the çenter of mass.ºf the spectrum, and is often perceived as a measure of sound brightness. Higher values correspond to brighter sounds and lower values correspond to darker sounds. This feature is used to characterize the timbral quality of the sounds. Figure 20 shows the Spectral Flatness over time, with the x-axis depicting time and the y-axis showing the spectral flatness value. Spectral flatness quantifies how noise-like or tonal a sound is, with values near 1 indicating a flatter, more noise-like spectrum and values near 0 indicating a more tonal signal. This feature helps distinguish

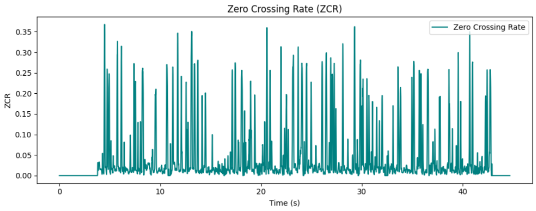


Figura 18: Zero Crossing Rate (ZCR) over time for rate of sign change

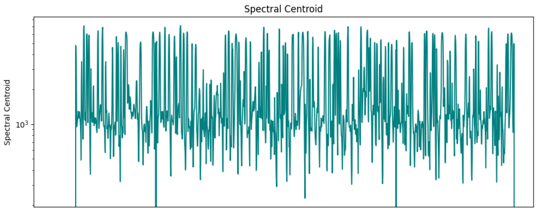


Figura 19: Spectral Centroid over time for audio spectrum brightness

between harmonic and noise-like sounds, and is used in various audio signal processing tasks.

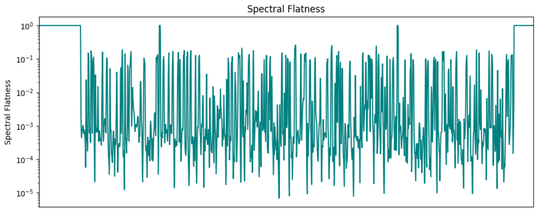


Figura 20: Spectral Flatness over time for the tonal nature of audio signals

In conclusion, XceptionNet and its integration with LSTM networks represent state-of-the-art deepfake detection methodologies for image and video media. The model's ability to focus on minute temporal and spatial anomalies ensures accurate detection of manipulated content. These capabilities extend to audio-based deepfakes through CNN + LSTM architectures, offering a comprehensive approach to address the challenges posed by deepfake media. As datasets and models evolve, these methodologies remain at the forefront of the accurate and consistent mitigation of media manipulation.

Cuadro 3: Performance of Different Models on Deepfake Detection across Media Types

Model	**Accuracy (%)**	**Media Type**	**Dataset**
XceptionNet	95.56	Image-based	CelebA
XceptionNet + LSTM	97.00	Video-based	FaceForensics++, DFDC, Celeb-DF
CNN + LSTM	98.00	Audio-based	DEEP-VOICE

Table 3 presents the accuracies of various deepfake detection models for different media types. The image-based XceptionNet model achieved an accuracy of 95.56 % on the CelebA dataset. In video-based deepfake detection, the performance improved to 97 % through the utilization of a combination of XceptionNet and LSTM, tested on FaceForensics++, DFDC, and Celeb-DF datasets. CNN combined with LSTM demonstrated the highest accuracy of 98 % in detecting deepfake audio using the DEEP-VOICE dataset. These findings suggest that deep learning architectures are effective in identifying deepfake content; however, their accuracy varies depending on the media type and dataset employed. The accuracy

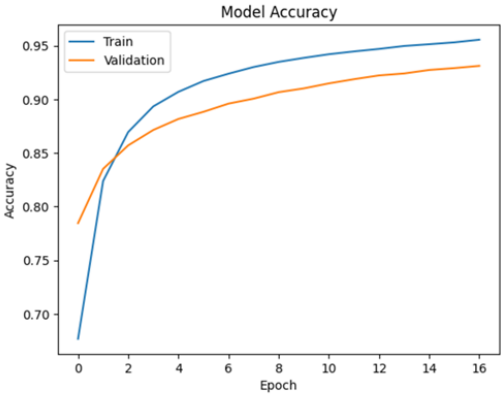


Figura 21: Training vs validation accuracy of XceptionNet Model

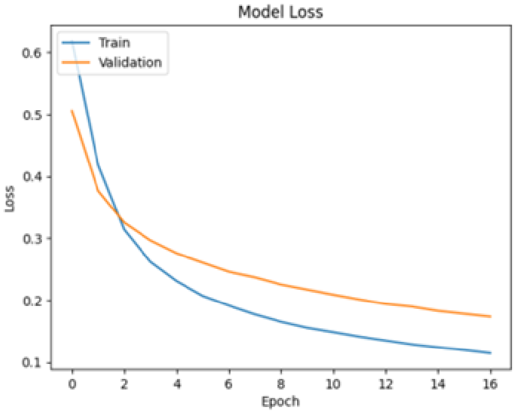


Figura 22: Training vs validation loss of XceptionNet Model

of the model during training and validation is shown in figure 21. The accuracy values are shown on

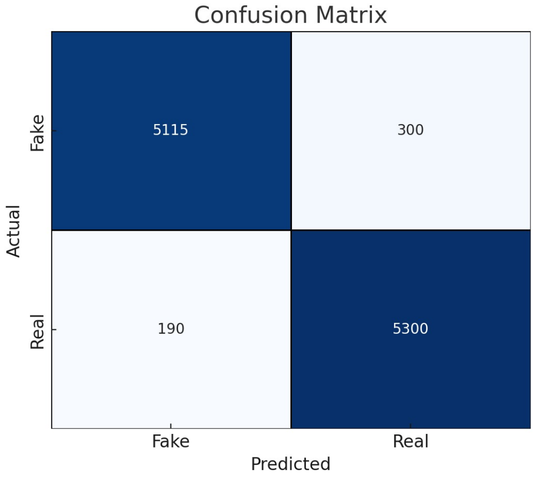


Figura 23: Confusion Matrix for Deepfake Image Detection Model

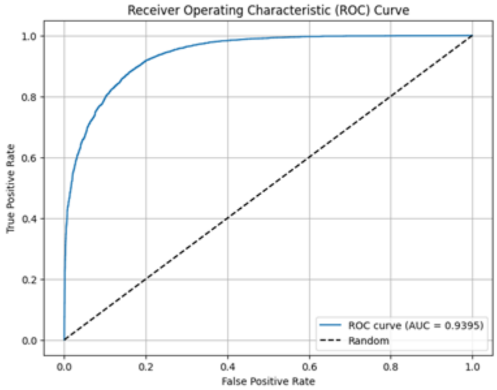


Figura 24: XceptionNet Model's discriminative capability

the y-axis, and the epochs are shown on the x-axis. As the model discovered patterns, the training accuracy initially decreased and then quickly increased. However, the validation accuracy increased. Both accuracies stabilize as epochs progress, suggesting that the model is convergent and has absorbed the most significant patterns. The discrepancy between the higher training accuracy and lower validation accuracy indicates a potential slight overfitting. This suggests that the model performs well on training data, but not on unseen data. The model achieved stable performance, as indicated by the plateau at high-accuracy levels. The model loss during training and validation across epochs is shown in figure 22. The y-axis represents the loss values and the x-axis represents the epochs. As the model learns, the training loss decreases from the initial high level. The validation loss also initially decreased before stabilizing over a few epochs. The validation loss remains higher than the training loss at the end, indicating possible overfitting and a generalization gap. A small difference between the two loss curves is acceptable, but a significant difference suggests the need for more training data or regularization methods to improve generalization.

The confusion matrix of the XceptionNet-based deep fake image detection model for fake and real classes is shown in figure 23. Fake labels are shown in columns, and real labels in rows. The model correctly identified 4882 fake instances (top-left value: 5115). False positives are indicated by the top-right values (300). The bottom-left value (190) represents false negatives, where real cases are mislabeled as fakes. True positives are bottom-right value (5300). This matrix shows high accuracy, with most of the predictions being correct. The model has a moderate proportion of false positives and negatives, indicating room for improvement. Overall, the model performed well, but could reduce misclassifications for a more balanced performance. The capacity of the model to differentiate between classes at various thresholds is assessed using the Receiver Operating Characteristic (ROC) curve, as shown in figure 24. The true positive rate (TPR) and false positive rate (FPR) are displayed on the y- and x-axis, respectively. The curve illustrates the effectiveness of the model in distinguishing between classes, as the decision threshold varies. A curve hugging in the upper-left corner indicates a perfectly discriminative model. With an Area Under the Curve (AUC) score of 0.9395, the model demonstrates significant discriminative power, indicating high success in differentiating authentic and fraudulent cases. The performance of the model improves as the AUC approaches 1. The dotted diagonal line denotes random guessing with an AUC of 0.5, whereas the curve of the model is considerably higher, indicating high predictive power. The slight dips in the curve suggest that the model requires further refinement. Overall, the ROC curve and AUC values showed that the model could reliably categorize the occurrences.

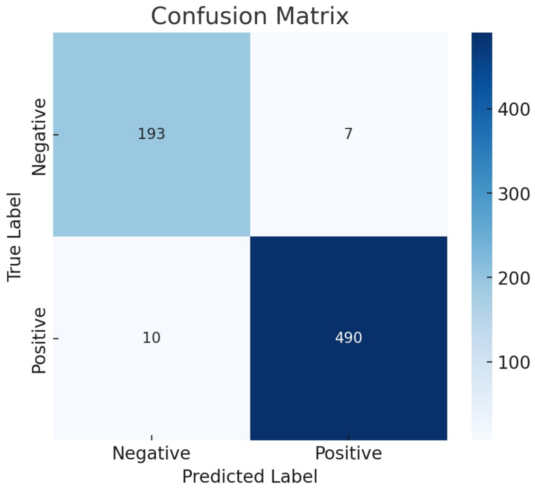


Figura 25: Confusion Matrix for Deepfake Video Detection Model

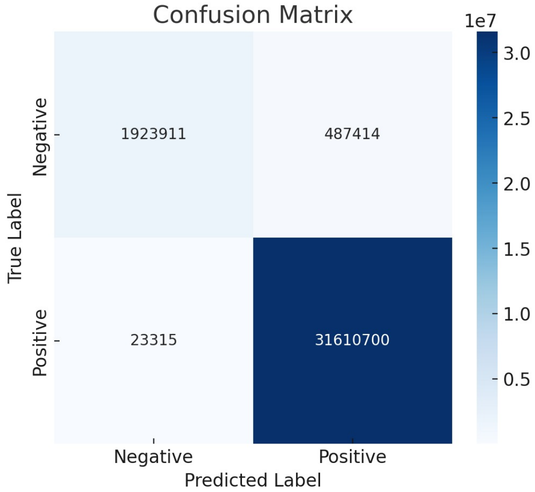


Figura 26: Confusion Matrix for Deepfake Audio Detection Model

This effectively demonstrates the performance of advanced deep learning models, such as XceptionNet, when used with LSTM networks for detecting deepfake content in any form of media. The extraction of spatial and temporal features is the reason for their accuracy in recognizing fake images, videos, and audio streams. The experimental results showed a very high accuracy for the developed models. Figure 25 and 26 show the confusion matrices for the deepfake video and audio detection models. The detection accuracies of the video- and audio-based systems are97 %, and 98 %, respectively, using XceptionNet + LSTM and CNN + LSTM.

5.1. Application Overview

The developed application processes an input video by segmenting it into individual frames, which are subsequently analyzed using a deepfake detection model to identify facial features. The detected faces are then converted into a NumPy array suitable for input into a classification model capable of determining the authenticity of the content. The model ultimately generates an output that indicates whether the provided video contains manipulated content. This pipeline facilitates efficient deepfake detection through the application of deep learning techniques for facial data extraction, as illustrated in Figure 27.

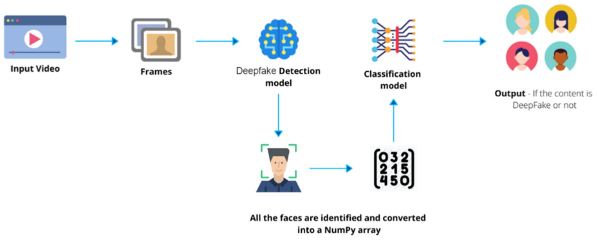
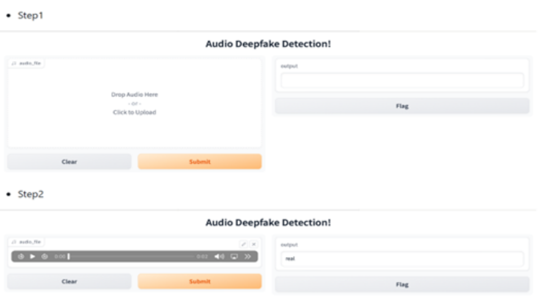

Input Video
Frames
Deepfake Detection model
Classification model
Output - If the content is DeepFake or not
032
215
450
All the faces are identified and converted into a NumPy array

Figura 27: Working of app for predicting video as deepfake or real



- Step1

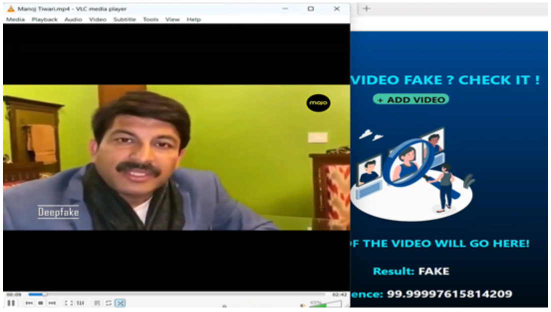
Audio Deepfake Detection!

- Step2

Audio Deepfake Detection!

real

Figura 28: App predicting real/ fake audio



Manoj Tiwari.mp4 - VLC media player
Media Playback Audio Video Subtitle Tools View Help
mojo
Deepfake
00:09
02:42
+ ADD VIDEO

VIDEO FAKE ? CHECK IT !

F THE VIDEO WILL GO HERE!

Result: FAKE

ence: 99.99997615814209

Figura 29: App predicting real or fake Manoj Tiwariâs Video

The developed applications utilized the capabilities of the XceptionNet, XceptionNet with LSTM, and CNN with LSTM models to demonstrate the efficacy of the proposed deepfake detection model in real-world scenarios. These applications illustrate how the models can be applied to routine media verification tasks by enabling individuals to identify modified photographs, videos, and audio in real time. Figures 28 and 29 depict how the application can determine whether an audio or video file is authentic or deepfake. To assess the validity of the video, it considers temporal anomalies between frames and extracts spatial features from individual frames. These applications provide a practical demonstration of how the proposed models can be employed for real-world deepfake detection challenges. This study demonstrates the effectiveness of deep learning models in detecting deepfake content across various media types, including images, videos, and audio. The results indicate the precision of models such as XceptionNet, CNN, and LSTM in distinguishing between authentic and manipulated content. The application's capability to classify real and fake instances, such as audio deepfakes and Manoj Tiwari videos, is noteworthy. The findings of this investigation reflect a growing necessity: the requirement for robust deepfake detection techniques coupled with an increased need to combat digital misinformation, thereby ensuring the authenticity of multimedia content in the era of AI-generated fabrications.

6. Conclusion

A novel three-stage deepfake detection framework is developed using advanced deep learning techniques to address the challenge of detecting manipulated media across images, videos, and audio. XceptionNet demonstrates promising performance for image-based media, achieving 95.56 % accuracy on the Celeb dataset. In the field of audio deepfakes, a new method combining CNN and LSTM networks attains an accuracy of 98.5 % on the DEEP-VOICE dataset. The combination of XceptionNet and LSTM networks proves effective for video-based deepfake detection, achieving an accuracy of 97.574 % on the Forensic++, DFDC, and Celeb-DF datasets. This represents highly accurate detection of various classes of deepfakes, indicating significant progress toward a comprehensive solution for deepfake detection. The implications of these findings extend to practical applications in media verification platforms, social media companies, and government organizations to combat misinformation. The implementation of these models in real-time systems will enable stakeholders to substantially enhance their ability to track and contain deepfakes, thereby ensuring the integrity of digital content in contemporary information ecosystems. However, the performance varies when the model is exposed to novel or unseen manipulations, necessitating additional experimentation to enhance generalizability. Furthermore, these models face challenges in terms of scalability and real-time applicability, particularly in resource-constrained environments.

Acknowledgements

The authors would like to thank Chitkara University Centre for Cybersecurity Research and Education, Chitkara University, Punjab, India for their support.

Referencias

[1] Sridevi, M., Mala, C. and Sanyam, S., 2012. Comparative study of image forgery and copy-move techniques. In *Advances in Computer Science, Engineering & Applications: Proceedings of the Second International Conference on Computer Science, Engineering and Applications (ICCSEA 2012)*, Volume 1, pp. 715-723. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-30157-5$_7$3

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2020. Generative adversarial networks. *Communications of the ACM*, 63(11), pp. 139-144. doi: 10.1145/3422622

[3] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. and Ortega-Garcia, J., 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, pp. 131-148. doi: 10.1016/j.inffus.2020.07.008

[4] Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Lopez Moreno, I. and Wu, Y., 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31. `https://proceedings.neurips.cc/paper/2018/file/a60bbce6720b1eb6fac88782dbb2126e-Paper.pdf`

[5] Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986. Learning representations by back-propagating errors. *Nature*, 323(6088), pp. 533-536. doi: 10.1038/323533a0

[6] Arik, S., Chen, J., Peng, K., Ping, W. and Zhou, Y., 2018. Neural voice cloning with a few samples. *Advances in Neural Information Processing Systems*, 31. `https://proceedings.neurips.cc/paper/2018/file/919cc2782b36a18ef440bb11f1f7e60e-Paper.pdf`

[7] Thies, J., Elgharib, M., Tewari, A., Theobalt, C. and NieÃner, M., 2020. Neural voice puppetry: Audio-driven facial reenactment. In *Computer VisionâECCV 2020: 16th European Conference*, Glasgow, UK, August 23â28, 2020, Proceedings, Part XVI 16, pp. 716-731. Springer International Publishing. doi: 10.1007/978-3-030-58526-6$_4$2

[8] Smith, A., 2020. Deepfakes are the most dangerous crime of the future, researchers say. [Online; accessed 03-August-2024]. URL: `https://www.independent.co.uk/tech/deepfakes-dangerous-crime-artificial-intelligence-a9655821.html`

[9] Tariq, S., Jeon, S. and Woo, S.S., 2021. Am I a real or fake celebrity? Measuring commercial face recognition web APIS under deepfake impersonation attack. *arXiv preprint arXiv:2103.00847*. DOI: N/A (arXiv preprint)

[10] Korshunov, P. and Marcel, S., 2018. Deepfakes: A new threat to face recognition? Assessment and detection. *arXiv preprint arXiv:1812.08685*. DOI: N/A (arXiv preprint)

[11] Yang, X., Li, Y. and Lyu, S., 2019, May. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8261-8265). IEEE. DOI: 10.1109/ICASSP.2019.8683164

[12] Jiang, L., Li, R., Wu, W., Qian, C. and Loy, C.C., 2020. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2889-2898). DOI: 10.1109/CVPR42600.2020.00294

[13] Li, Y., Yang, X., Sun, P., Qi, H. and Lyu, S., 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207-3216). DOI: 10.1109/CVPR42600.2020.00327

[14] Kwon, P., You, J., Nam, G., Park, S. and Chae, G., 2021. Kodf: A large-scale Korean deepfake detection dataset. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10744-10753). DOI: 10.1109/ICCV48922.2021.01057

[15] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M. and Ferrer, C.C., 2020. The deepfake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397.* DOI: N/A (arXiv preprint)

[16] GÃ¼era, D. and Delp, E.J., 2018, November. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1-6). IEEE. DOI: 10.1109/AVSS.2018.8639163

[17] Liu, Z., Luo, P., Wang, X. and Tang, X., 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (pp. 3730-3738). DOI: 10.1109/ICCV.2015.425

[18] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and NieÃner, M., 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1-11). DOI: 10.1109/ICCV.2019.00009

[19] Pham, L., Lam, P., Tran, D., Tang, H., Nguyen, T., Schindler, A., Skopik, F., Polonsky, A. and Vu, H.C., 2025. A comprehensive survey with critical analysis for deepfake speech detection. *Computer Science Review*, 57, p.100757. https://doi.org/10.1016/j.cosrev.2024.100757.

[20] Li, L., Bao, J., Yang, H., Chen, D. and Wen, F., 2019. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457.* DOI: N/A (arXiv preprint)

[21] Bird, J.J. and Lotfi, A., 2023. Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion. *arXiv preprint arXiv:2308.12734.* DOI: N/A (arXiv preprint)

[22] Yadav, A. and Vishwakarma, D.K., 2024. AW-MSA: Adaptively weighted multi-scale attentional features for DeepFake detection. *Engineering Applications of Artificial Intelligence*, 127, p.107443. DOI: 10.1016/j.engappai.2023.107443

[23] Wang, S., Zhu, D., Chen, J., Bi, J. and Wang, W., 2024. Deepfake face detection with multi-layer fusion of diverse features. *Pattern Recognition Letters*, 151, pp. 20-30. DOI: 10.1016/j.patrec.2024.01.003

[24] Nawaz, M., Javed, A. and Irtaza, A., 2024. A deep learning model for FaceSwap and Face-Reenactment Deepfakes Detection. *Applied Soft Computing*, p.111854. DOI: 10.1016/j.asoc.2024.111854

[25] Bisogni, C., Loia, V., Nappi, M. and Pero, C., 2024. Acoustic features analysis for explainable machine learning-based audio spoofing detection. *Computer Vision and Image Understanding*, 249, p.104145. https://doi.org/10.1016/j.cviu.2024.104145.

[26] Jayashre, K. and Amsaprabhaa, M., 2024. Safeguarding media integrity: A hybrid optimized deep feature fusion based deepfake detection in videos. *Computers & Security*, 142, p.103860. https://doi.org/10.1016/j.cose.2024.103860.

[27] El-Gayar, M.M., Abouhawwash, M., Askar, S.S. and Sweidan, S., 2024. A novel approach for detecting deep fake videos using graph neural network. *Journal of Big Data*, 11(1), p.22. https://doi.org/10.1186/s40537-024-00744-z.

[28] Amin, M.A., Hu, Y., Guan, Y. and Amin, M.Z., 2024. Exploring varying color spaces through representative forgery learning to improve deepfake detection. *Digital Signal Processing*, p.104426. https://doi.org/10.1016/j.dsp.2024.104426.

[29] Hsu, L.Y., 2024. AI-assisted deepfake detection using adaptive blind image watermarking. *Journal of Visual Communication and Image Representation*, p.104094. https://doi.org/10.1016/j.jvcir.2024.104094.

[30] Chen, B., Liu, X., Xia, Z. and Zhao, G., 2023. Privacy-preserving DeepFake face image detection. *Digital Signal Processing*, 143, p.104233. https://doi.org/10.1016/j.dsp.2023.104233.

[31] Liang, Y., Wang, M., Jin, Y., Pan, S. and Liu, Y., 2023. Hierarchical supervisions with two-stream network for Deepfake detection. *Pattern Recognition Letters*, 172, pp.121-127. https://doi.org/10.1016/j.patrec.2023.06.009.

[32] Liu, R., Zhang, J. and Gao, G., 2024. Multi-space channel representation learning for mono-to-binaural conversion-based audio deepfake detection. *Information Fusion*, 105, p.102257. https://doi.org/10.1016/j.inffus.2024.102257.

[33] Ilyas, H., Javed, A. and Malik, K.M., 2023. AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audioâvisual deepfakes detection. *Applied Soft Computing*, 136, p.110124. https://doi.org/10.1016/j.asoc.2023.110124.