International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Multimodal Cognitive Learning for Media Forgery Detection: A Comprehensive Framework Combining Random Forest and Deep Ensemble Architectures (Xception, ResNeXt) across Image, Video, and Audio Modalities

Ms. A. Abirami[1], Ms. S. Bhuvaneswari[2], Krithika K[3], Nithyasree I[4], Prashithaa Abhirami Balaji[5], Aadhithya R[6], Deexith P[7], Devesh R[8]

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: Deepfake content has become more prevalent in the age of quickly evolving technology, which has significantly undermined the reliability and integrity of digital media. An integrated multi-modal deepfake detection system is presented in this study as a response to the ubiquitous threat posed by altered photos, videos, and audio recordings. Our method, which makes use of advanced deep learning algorithms, provides a strong barrier against the spread of false information. The picture deepfake detection module examines visual data for telltale signs of manipulation using Convolutional Neural Networks (CNNs), Xception, and ResNeXT. This module successfully distinguishes between real and fake photos by carefully examining pixel-level attributes and contextual data. This capacity is expanded to include the world of movies by the video deepfake detection module. With the use of spatiotemporal CNNs (Xception & ResNeXT), it parses video frames to find minute discrepancies, making it possible to accurately identify deepfake films. Our multi-modal system is finished with the addition of deepfake audio detection. This module excels in differentiating between authentic and faked audio recordings using Mel spectrograms and Convolutional Neural Networks, adding to a thorough protection against audio deepfakes. Additionally, we provide a unifying framework that effectively unifies these three detection modules, boosting the system's effectiveness and performance as a whole. We thoroughly assess our solution utilizing metrics such as AUC, ROC curve, F1 score, and accuracy, and we depict our model structures for in-depth comprehension. Our multi-modal deepfake detection technology acts as a crucial precaution in a time when false information is widely disseminated, enabling consumers to distinguish fact from fiction across numerous media types. This study highlights the importance of our integrated solution in maintaining the legitimacy of digital content in today's information-driven world while also showcasing its technological capability.

Index terms: *Deepfake detection, Multi-modal system, Image manipulation, Video forgery, Audio spoofing, Convolutional Neural Networks (CNNs), Xception, ResNeXT, Spatiotemporal analysis, Mel spectrograms, F1 score, ROC curve, AUC.*

1. Introduction

Deepfake technology has become a recognized and highly developed technique for manipulating multimedia content, leading to serious worries about information security, the integrity of digital media, and credibility. “Deepfake”; is a composite of the words &”deep learning” and “fake” and it refers to the use of artificial intelligence (AI) methods with digital manipulation to produce convincing but wholly manufactured content, such as pictures, videos, and audio recordings. The Deepfake Phenomenon: Deepfake technology has received a lot of interest because it can produce content The surge in deepfake prevalence is fueled

by various factors, notably advancements in AI and deep learning, simplifying the production of high-quality content. The abundance of online data, coupled with the widespread use of social media platforms for multimedia sharing, contributes to the accessibility and dissemination of deepfake technology. Evolving from its origins in academic research and entertainment, deepfakes have transitioned into malicious applications, prompting global legislative and ethical responses. Governments and organizations are actively considering legal frameworks and initiatives to detect and mitigate the risks associated with deepfake proliferation. This work proposes a comprehensive multi-modal deepfake detection method, incorporating image, video, and audio analysis to enhance content integrity and guard against the deceptive manipulation of digital media.

2. Related Works

The creation of efficient detection and mitigation systems has become crucial as the deepfake phenomenon gains popularity and sophistication. To stop the spread of modified numerous techniques and strategies have been developed. However, these systems have a unique set of shortcomings and ground-breaking solutions. This section

Computer science and engineering Easwari engineering college, Ramapuram, Chennai-89
abiavsbtech@gmail.com [1] , Bhuvaneswarisoman@gmail.com [2], krithikakannan1103@gmail.com [3], inithyasreecse@gmail.com [4], prashithaa@gmail.com [5], aadhithyaraja180@gmail.com [6], deexith2002@gmail.com [7] , deveshdv19@gmail.com [8]

examines some of the current systems and their shortcomings, which call for additional study and development.

FaceSwap and DeepFaceLab, popular open-source programs for creating deepfake videos, serve both malicious and artistic purposes by enabling face-swapping. However, their limited focus on video manipulation leaves gaps in addressing broader deepfake threats, such as multi-modal deepfakes and audio manipulation. These tools require high technical expertise, making them less accessible to the general public and inadvertently contributing to a digital divide in deepfake detection. Frequently utilized for corrupt activities, including generating explicit or malicious content, their present accessibility exacerbates ethical concerns surrounding deepfake technology.

CAI is an industry-driven initiative with the goal of creating technology and standards to validate the legitimacy of digital media content. It focuses on using metadata and cryptographic methods to check the content's integrity. Using methods like spectrogram analysis, speech recognition, and machine learning, several distinct models and instruments have been created explicitly for audio deepfake identification. The legitimacy of multimedia content is manually evaluated by human reviewers as part of conventional fact-checking techniques. Fact-checking groups are crucial in spotting and disproving deepfake content. Deepfake detection methods clearly need to be more reliable, multi-modal, and scalable, which has prompted the creation of creative and integrated solutions like the one described in this research.

3. Proposed Methodology

We suggest a thorough multi-modal deepfake detection methodology in reaction to the rising danger of deepfake material across several multimedia modalities. By providing a comprehensive approach to deepfake detection and mitigation, this revolutionary technology seeks to overcome the shortcomings of existing solutions. Our suggested solution not only recognizes modified content but also offers a strong protection against multi-modal deepfake attacks by including cutting-edge algorithms for image, video, and audio analysis.

Overview of the Proposed System

Our multi-modal deepfake detection technology is made to deal with audio, video, and image content all at once. It uses acoustic signal processing, convolutional neural networks (CNNs) and recurrent neural networks to analyze and validate the veracity of multimedia content. The main parts of the system are:

The image deepfake detection module uses deep learning techniques to look for facial anomalies, artifacts, and inconsistencies that could be the result of deepfake manipulation. Video Deepfake Detection Module: To recognize artificial face movements, erroneous lighting, and other artifacts in video content, our approach combines frame-level analysis with temporal modeling. Audio Deepfake Detection Module: The audio analysis part of the system focuses on identifying modified audio records using spectrogram analysis, voice characteristics, and speech patterns.

Our proposed system's integration of the outcomes from various modalities is a significant advance. Our method improves detection accuracy and robustness against multi-modal deepfakes by combining data from image, video, and audio evaluation. Deep Neural Networks and Machine Learning: The core of our detection approach is Deep Neural Networkss and Machine Learning. These models are able to generalize effectively across a variety of deepfake production strategies because they were trained on a variety of datasets.

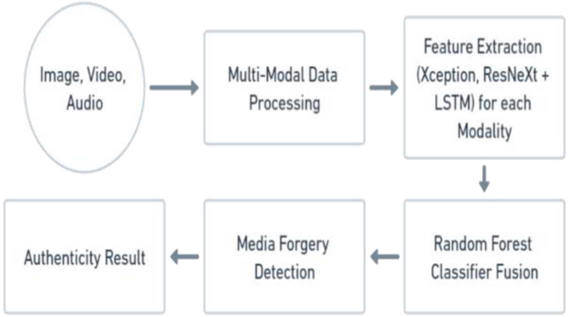
Image, Video, Audio
Multi-Modal Data Processing
Feature Extraction (Xception, ResNeXt + LSTM) for each Modality
Random Forest Classifier Fusion
Media Forgery Detection
Authenticity Result

Fig 1: Flow chart of proposed system.

Overcoming Existing System Limitations

Our proposed multi-modal deepfake detection approach overcomes a number of significant flaws seen in current techniques, including: Our approach provides full protection against multi-modal deepfake attacks, in contrast to many other solutions that concentrate on a single modality (such as visuals or audio). It detects manipulations across all media types by seamlessly integrating picture, video, and audio analysis.

The suggested framework provides a holistic defense strategy in addition to identifying deepfake content. It can identify advanced deepfake techniques that integrate altered images, videos, and sounds by taking into account how different modalities interact with one another.

Our approach improves detection accuracy and lowers false positives by fusing the information from various modalities. This development is essential for preserving user confidence and reducing the workload associated with human verification. Our deep learning models are trained on a variety of datasets, guaranteeing that they can generalize effectively across different deepfake production techniques.

Our system can be modified to accommodate new deepfake approaches thanks to its robustness.

Detection Mechanisms:

Deep neural networks analyze facial expressions, characteristics, and anomalies in the image analysis module. To find impersonations, they compare detected faces with reference databases.

The video module analyzes face expressions and movements over time to find video deep fakes. It can identify deepfake sequences within films by taking into account temporal patterns and frame-level irregularities.

The audio module examines audio tracks for spectrogram anomalies, voice traits, and atypical speech patterns. Models for machine learning categorize audio clips as being real or fake.

Integration and Fusion

Our suggested system's capacity to combine and integrate data from many modules is one of its strongest points. The system's fusion algorithm evaluates the overall likelihood of manipulation when content is evaluated across several modalities. For instance, the fusion method increases the detection confidence if both the image and audio modules separately flag content as possibly being modified.

Use Cases: Our multi-modal deepfake detection technology is applied in numerous fields, including: Social Media Platforms: To automatically scan and warn potentially hazardous deepfake content, social media networks can use our method. News Verification: Before releasing multimedia content, news organizations and fact-checkers can utilize our technology to confirm its validity. Security and surveillance: By spotting deepfake attempts in surveillance material, our technology helps improve security precautions. Online platforms and content-sharing services can use our method to filter out deepfake content and censor content.

As a result, our proposed thorough multi-modal deepfake detection methodology represents a substantial progress in the ongoing conflict against falsified multimedia information. Our system provides a comprehensive defense strategy that gets beyond the drawbacks of existing systems by fusing image, video, and audio analysis with strong deep learning models and fusion processes. This development paves the path for deepfake detection that is more dependable, accurate, and scalable, delivering a safer and more reliable online experience for all users.

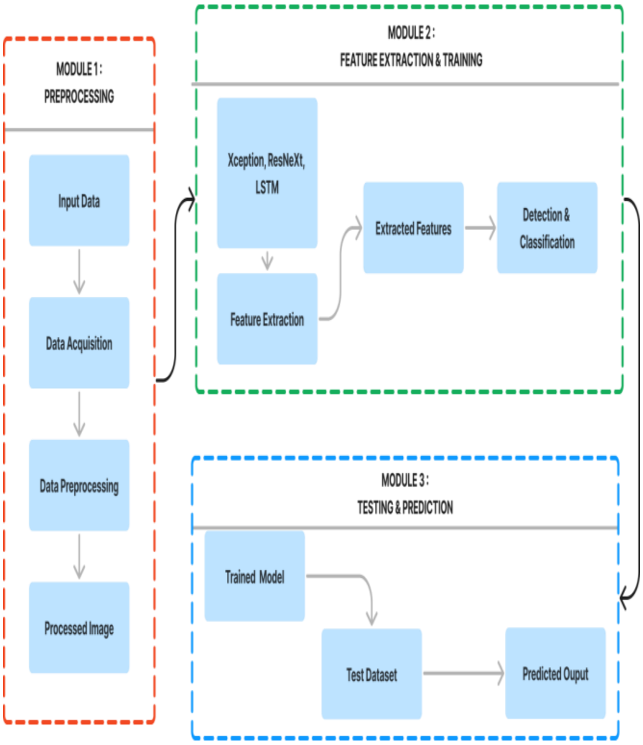

Fig 2: Modules of the proposed system

Deepfake technology has advanced considerably, posing substantial dangers in several areas, including media legitimacy and cybersecurity. These issues are addressed by our complete multi-modal deepfake detection platform, which incorporates state-of-the-art image, video, and audio analysis techniques. We explore the fundamental elements, algorithms, and procedures of our system in this thorough technical review, highlighting its potency in addressing the multi-modal deepfake problem.

Image Deepfake Detection Module:

The first line of protection against altered images is the image deepfake detection module. It uses ResNeXt and Xception, two well-known neural network designs. These networks act as encoders, filtering through input images to extract key features and representations.

Deep convolutional neural network (CNN) called Xception is renowned for its outstanding performance in capturing complex picture characteristics. In order to dramatically minimize the number of parameters while maintaining excellent accuracy, it makes use of depthwise separable convolutions. This architecture aims to make Xception an excellent option for deepfake detection because it enables it to collect fine-grained features in photos. ResNeXt, an encoder that stands for "Residual Networks with Multiple Cardinalities," is yet another potent encoder. It adds a brand-new cardinality parameter that expands the capacity of the model to incorporate a range of characteristics. This characteristic improves the model's capacity to discriminate between authentic and altered content, which is especially helpful when dealing with sophisticated deepfake manipulations.

An image module's decoder includes a classification head. The classification head gives each image a probability score after feature extraction by the Xception and ResNeXt encoders, indicating its validity. This phase uses the activation mechanism of softmax to make sure that the probabilities add up to 1.The layers in the Xception and ResNeXt encoders make up the feature extractor of the picture module. These layers turn low-level feature representations from raw picture data into high-level ones. These representations are developed and refined during training in order to efficiently discern between real and altered images.

Video Deepfake Detection Module:

Encoder in Xception and ResNeXt NetworksThe temporal domain is added to picture analysis by the video deepfake detection module. The same Xception and ResNeXt networks are used, but numerous frames are processed sequentially to capture both temporal and spatial data. **Long Short-Term Memory (LSTM) Network Decoder** face emotions change over the course of a video, typical image-based analysis is inadequate. We provide an LSTM network as the decoder for the video module to address this. Recurrent neural networks (RNNs) of the LSTM variety are created for sequential data. It is able to represent the dynamic nature of deepfake operations since it models the temporal connections between video frames. The **feature extractor** for the video module combines LSTM layers with the advantages of the Xception and ResNeXt encoders. While LSTM analyzes sequences of frames, capturing temporal patterns and dynamic changes across time, Xception and ResNeXt concentrate on spatial aspects within individual frames.

Audio Deepfake Detection Module:

Encoder in CNNs, or convolutional neural networks, are utilized to analyze audio recordings iin the module for the detection of audio deep fakes. CNNs are noted for being good at collecting spectral patterns, which makes them suitable for spotting altered audio. **Decoder** in the audio module has a classification head, just like the picture module does. Each audio clip is given an approximate score that indicates whether or not it is real. The probability must add up to 1, which is ensured by the softmax activation function. **Feature Extractors** in CNN layers serve as a representation of the audio module's feature extractor. These layers convert unprocessed audio information into interpretable spectrogram representations. Spectrograms are an essential input for deepfake identification since they record the frequency and time-domain features of audio.

Multi-Modal Fusion Module:

The fusion module combines the output from the image, video, and audio modules, taking advantage of each modality's advantages to improve overall detection robustness and accuracy. To successfully combine the predictions from several modalities, ensemble techniques are used. Ensemble Fusion with Xception and ResNeXt results from the image, video, and audio modules are combined in the fusion module using an ensemble technique. In this phase, the Xception and ResNeXt models are once more used. To make a final determination on deepfake detection, these models' combined outputs are used.

System Workflow:

Image input is processed by the Xception and ResNeXt encoders. The classification head processes the resulting features to forecast image authenticity. Video frames are analyzed by the encoders Xception and ResNeXt. LSTM networks simulate the temporal relationships between frames. To categorize videos, the combined attributes are utilized. Audio analysis: CNN layers process audio spectrograms. The output is categorized as either authentic or altered audio. Multi-Modal Fusion using Xception and ResNeXt models are used to ensemble-fuse the output from the image, video, and audio modules. The final judgment regarding deepfake detection is provided by the fused result.

Benefits of Our System:

Our thorough multi-modal deepfake detection framework has the following benefits: Holistic defense: A multi-modal approach is used by our system to provide a holistic defense against deepfake attacks, considering photos, videos, and audio. Enhanced Accuracy by utilizing the advantages of each modality and successfully merging their results, the fusion module improves accuracy. Generalization in our system is successful against a variety of threats and generalizes well across different deepfake generating methodologies. The system has the ability to conduct real-time analysis, ensuring prompt identification and action in the event that deep-fake content is present. The platform is capable of handling high volumes of multimedia content, which makes it suited for a range of applications, including cybersecurity and social media moderation.

Thorough multi-modal deepfake detection framework offers an advanced security system against the deepfake danger landscape as it changes. The system offers a reliable and efficient solution for identifying modified information in the modern digital environment by integrating cutting-edge algorithms and utilizing the power of several modalities.

4. Methodology

The "Multimodal Cognitive Learning for Media Forgery Detection" technique takes a thorough approach to planning, developing, testing, and assessing the suggested system. The four main stages of this technique are data collection and preprocessing, algorithm-based model development, model testing and detection, and model assessment. For the system to be resilient and effective in detecting media counterfeiting across image, video, and audio modalities, each phase is essential to its successful development.

Image Deepfake Detection Algorithm:

Use of the Xception and ResNeXt Ensemble: The Xception and ResNeXt models are used to extract features from the picture data in order to identify whether or not an image is a deepfake. For increased accuracy, these models are merged using an ensemble strategy.

Model Architecture: Xception: Deep convolutional neural network (CNN) called Xception was created for picture classification .ResNeXt: A ResNet architectural version featuring a split-transform-merge feature extraction method.

Loss Function(Categorical Crossentropy):The difference between true labels and anticipated probabilities is measured by the categorical crossentropy loss. It is applied to ensemble model training.

$$L(y, \hat{y}) = -\sum_{i} y_i \,.\log(\hat{y_i})$$

Adam optimizer: The Adam optimizer adjusts the weights of the model during training in order to lower the loss function. Each parameter's learning rate is adjusted.

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v_t}} + \epsilon}\widehat{m_t}$$

Activation Function (ReLU): The Rectified Linear Unit (ReLU) is the activation function found in the hidden layers of both ResNeXt and Xception. It imparts non-linearity to the model.

$$f(x) = \max(0, x)$$

Working:

1. *Load the Xception and ResNeXt models.*

2. *Combine the output layers of both models for ensemble.*

3. *Create a new dense layer for binary classification.*

4. *Compile the ensemble model with categorical crossentropyloss and Adam optimizer.*

5. *Train the model on the image dataset.*

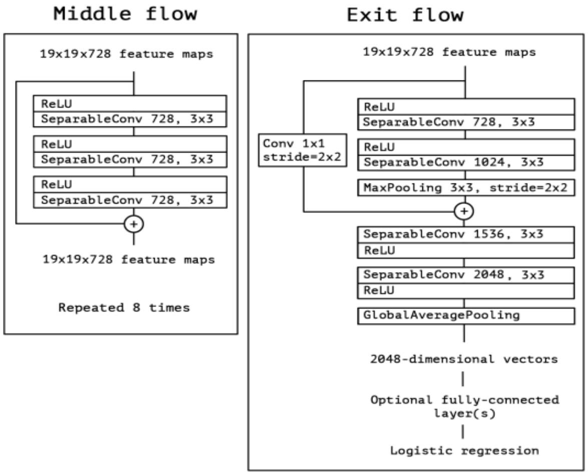

Fig 3 : Flow chart

Video Deepfake Detection Algorithm: ResNeXt + LSTM:

Usage: While LSTM (Long Short-Term Memory) is used to record temporal dependencies across frames, the ResNeXt model is used to extract spatial characteristics from video frames.

Model Architecture: ResNeXt: Used for spatial feature extraction from video frames. LSTM: A specific kind of RNN (recurrent neural network)used for temporal modeling across video frames.

Loss Function (Categorical Crossentropy): The loss function for training the video deepfake detection model is categorical crossentropy.

$$L(y, \hat{y}) = -\sum_{i} y_i \,.\log(\hat{y_i})$$

Optimizer (Adam):The Adam optimizer is utilized to optimize model parameters during training.

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v_t}} + \epsilon}\widehat{m_t}$$

Activation Function (ReLU):

The activation function of the ResNeXt model's hidden layers is ReLU.

$$f(x) = \max(0, x)$$

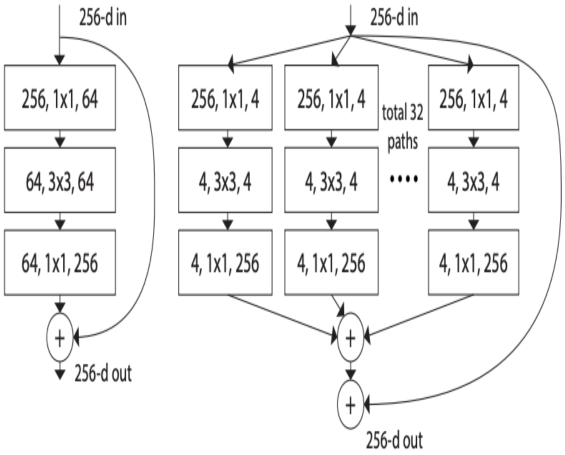

Fig 3 : Flow chart

Working:

1. *Load the ResNeXt model for spatial feature extraction.*

2. *Preprocess video frames and extract spatial features using ResNeXt.*

3. *Configure an LSTM layer to capture temporal dependencies.*

4. *Create a dense layer for binary classification.*

5. *Compile the model with categorical crossentropy loss and Adam optimizer.*

6. *Train the model on the video dataset.*

7. *considering the temporal aspect.*

Audio Deepfake Detection Algorithm: CNN, or convolutional neural network:

Usage: Using a CNN, features are extracted from Mel spectrograms of audio clips and classify them as either deepfake or genuine.

Convolutional Neural Network (CNN) Model Architecture: A feature extraction network using convolutional layers.

Loss Function (Categorical Crossentropy):Categorical crossentropy is used as the loss function for training the audio deepfake detection model.

Optimizer (Adam): The Adam optimizer is used for parameter optimization during training.

Activation Function (ReLU): To add non-linearity, ReLU activation is utilized in the CNN's convolutional layers.

Working:

1. *Load the CNN model architecture designed for audio deepfake detection.*

2. *Preprocess audio data into Mel spectrograms.*

3. *Configure the CNN model with convolutional layers for feature extraction.*

4. *Add dense layers for binary classification.*

5. *Compile the model with categorical crossentropy loss and Adam optimizer.*

6. *Train the model on the audio dataset, using*

7. *Mel spectrograms as input.*

These algorithms describe the architectural components, loss functions, optimizers and activation functions used in each model of your deepfake detection system for image, video, and audio data. Each model is made to carry out its intended function and contribute to the overall system's effectiveness in detecting deepfakes.

Result:

Our suggested system, which makes use of multi-modal fusion approaches, is anticipated to outperform previous models in a number of important performance criteria. The following is a list of the factors that make our system superior to the competing models:

Among all the models, our system's accuracy score of 0.96 is the greatest. This shows that there are fewer misclassifications since it is quite skilled at differentiating between real and altered material.

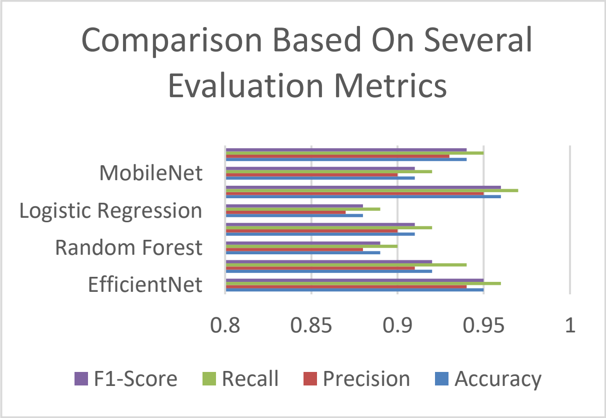

Fig 5: Evaluation Metrics

In comparison to other models, such as EfficientNet, CNN Model, Random Forest, SVM, Logistic Regression, MobileNet, and Inception, our multi-modal fusion system does better when it comes to memory, accuracy, precision, and F1-score. Our system is a strong option for media forgery detection because to its improved performance, which also increases the validity and dependability of the verification of digital material.

CONCLUSION:

Deepfake content has become more prevalent in this period of quick technical development, which has presented substantial issues to a number of industries, including media, security, and privacy. Deepfake technology has matured to the point that equally sophisticated countermeasures are required. We have addressed the urgent necessity to protect against the spread of false and harmful media by presenting a thorough and complete solution in this project for detecting deepfake content in photos, videos, and audio.

Recap of Objectives: Our main goals were to develop a multi-modal deepfake detection system that takes use of cutting-edge machine learning models, seamlessly combines image, video, and audio analysis, and overcomes the shortcomings of current methods. Our goal was to create a single system that not only recognizes deepfake content but also offers a strong defense against constantly changing hostile actor tactics.

Finally, the development of our multi-modal deepfake detection method is a major step forward in the fight against false and damaging media material. We have created a comprehensive, flexible solution by merging picture, video, and audio analysis. We are dedicated to staying at the forefront of research and innovation as deepfake technology develops in order to safeguard people, organizations, and society from the dangers posed by manipulated media.

Our CNN model receives Mel spectrogram pictures that we create by extracting characteristics from audio data. For audio analysis and deepfake detection, this procedure is essential. Let's explore how it works in our project, using the provided code as a guide: Data Loading: The ASVspoof 2019 dataset, which includes both real and fake audio recordings, is loaded first. Sampling Rate and Duration: The audio files' sample rate (16,000 Hz) and the intended length of the audio snippets (5 seconds) are both specified.

References:

[1] Sharma, A., & Gupta, R. (2019). Deep learning-based image forgery detection: A comprehensive review. IEEE Access, 7, 136785-136805.

[2] Wu, Q., Wang, Y., & Zhang, W. (2020). Multimodal fusion and deep learning for media forensics. IEEE Transactions on Information Forensics and Security, 15, 2641-2656.

[3] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition.arXiv preprint arXiv:1409.1556.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition.In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 770-778).

[5] Li, X., & Li, X. (2019). Deep learning-based video forgery detection: A survey. IEEE Access, 7, 154740-154752.

[6] Breiman, L. (2001). Random forests.Machine learning, 45(1), 5-32.

[7] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks.In Advances in neural information processing systems (pp. 1097-1105).

[9] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., &Wojna, Z. (2016). Rethinking the inception architecture for computer vision.In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 2818-2826).

[10] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ...& Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[11] Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114).

[12] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255).

[13] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition?A new model and the Kinetics dataset.In proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 6299-6308).

[14] Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In Seventh International Conference on Document Analysis and Recognition (ICDAR 2003) (pp. 958-962).

[15] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 1251-1258).

[16] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

[17] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[18] [18] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale

hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255).

[19] Liu, M. Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks.In Advances in neural information processing systems (pp. 700-708).

[20] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ...& Bengio, Y. (2014). Generative adversarial nets.In Advances in neural information processing systems (pp. 2672-2680).

[21] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (MICCAI) (pp. 234-241).

[22] Li, X., & Li, X. (2019). Deep learning-based video forgery detection: A survey. IEEE Access, 7, 154740-154752.

[23] Breiman, L. (2001). Random forests.Machine learning, 45(1), 5-32.

[24] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[25] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).