Non-Facial Video Spatiotemporal Forensic Analysis Using Deep Learning Techniques

Premanand Ghadekar, Vaibhavi Shetty*, Prapti Maheshwari, Raj Shah, Anish Shaha, Vaishnav Sonawane

Department of Information Technology, Vishwakarma Institute of Technology, Pune, India

Received 11 July 2022; received in revised form 15 September 2022; accepted 04 October 2022

DOI: https://doi.org/10.46604/peti.2022.10290

Abstract

Digital content manipulation software is working as a boon for people to edit recorded video or audio content. To prevent the unethical use of such readily available altering tools, digital multimedia forensics is becoming increasingly important. Hence, this study aims to identify whether the video and audio of the given digital content are fake or real. For temporal video forgery detection, the convolutional 3D layers are used to build a model which can identify temporal forgeries with an average accuracy of 85% on the validation dataset. Also, the identification of audio forgery, using a ResNet-34 pre-trained model and the transfer learning approach, has been achieved. The proposed model achieves an accuracy of 99% with 0.3% validation loss on the validation part of the logical access dataset, which is better than earlier models in the range of 90-95% accuracy on the validation set.

Keywords: transfer learning, mel-spectrogram, forgery, data augmentation

1. Introduction

Surveillance cameras are now found in almost every location, such as banks and businesses, where the recordings are used to reduce crime. However, due to the availability of video editing software like Adobe, the process of video editing has become simple [1]. Currently, videos are commonly considered the most extensively utilized communication and entertainment medium. Hence, this kind of vogue surely emphasizes the utilization of automated perusal and video content understanding using technology. This is referred to as the major goal of computer vision [2].

Several methods have been developed for detecting image forgeries, most of which rely on the extraction of specific image modifications in the output image or examination of discrepancies compared to a regular camera pipeline [3]. Based on the modification domain, these adjustments can be classified as intra-frame or inter-frame [4]. This study concentrated on forgeries in the video along directions of inter-frame which are ubiquitous in surveillance videos and difficult to detect. A significant gap has been found in existing work in direction of inter-frame forgeries due to the lack of a temporal video forgery dataset. A dataset that consists of various temporal forgeries is created and published on Kaggle [5]. The main objective of the proposed research is to come up with a model trained on this dataset that can identify temporal forgeries with an average accuracy of 85% on the validation data.

This study also helps identify false audio, reduce the spread of rumors and hate speech, make better-informed decisions, and master the art of fake audio detection. Digital authentication and forensics are the conformation and examination of audio for validation of its uniqueness (identify forgery, if any), and it also has a lot of applications [6]. Copy-move, deletion, insertion, replacement, and splicing are all methods for audio forgery [7]. For the audio forgery detection of text-to-speech

* Corresponding author. E-mail address: vaibhavi.shetty19@vit.edu

(TTS) and voice-conversion (VC) frauds, a ResNet-34 using the transfer learning approach is implemented. After successful training of the model, the prediction of audio files and their classification into three categories: real, spoof_TTS, and spoof_VC have been done. The ASVspoof 2019 dataset has data balance problems [8]. Hence, the proposed model introduces a framework to fill that gap, leading to help the models generalize to a wider range of inputs.

2. Literature Review

Richard and Roussev [9] discussed many digital forensic image/video analysis techniques, incorporating deep learning object identification structure using the YOLO method, and chromatic and pattern techniques for object recognition approaches. To accomplish digital video analysis in a forensic context, their study not only covers various forensic visual data analysis issues and resolutions but also describes several unique graphic data analytic techniques. Several experimental results for picture enhancing techniques and object recognition methods are shown, demonstrating how YOLO in particular may be used to find numerous criminal suspects and crime scene objects, then establish a link between some of them. "YOLOv3: An Incremental Improvement," [10] made transparent YOLO upgrades. The findings of multistage and single-stage object detectors are compared in this article. In terms of speed and accuracy, the numbers confirmed that the YOLOv3 object detector outperforms other object detectors.

The goal of virtual/digital media forensics is to establish systems that can automatically assess visual integrity. In the literature, feature-based [11-12] and convolutional neural network (CNN)-based [13-14] integrity analysis approaches had been investigated. Most of the proposed techniques for video-based digital forensics attempt to identify computationally cheap alterations, such as dropped or duplicated frames or copy-move operations [15]. Ways that differentiate computer-found faces from genuine faces are used to detect face-based interventions. And a two-stream network was proposed to identify two distinct face-swapping manipulations [16]. A new dataset by Rossler et al. [17] was especially relevant to practitioners, which has around half a million modified photographs created via feature-based face editing.

Hinton et al.[18] talked about the limits of CNNs for inverse graphics applications, laying the groundwork for a more vigorous "capsule" design. However, due to the absence of an optimization algorithm and the limits of technology at the time, this complicated architecture could not be executed properly. Instead, CNNs that are simple to create have become popular. Sharma and Singh [19] proposed a combined technique of image classification that employs transfer learning for feature selection and principal component analysis (PCA) for feature reduction. Capsule networks have now been created with impressive early results due to the introduction of the expectation-maximizing routing algorithm along with the dynamic routing algorithm [20]. According to Sabour et al. [20], stratified pose relationships amongst the pieces of objects are well characterized using the output of a dynamic routing algorithm, i.e., the accordance between capsules.

Many machine learning algorithms have been specially designed for video forgery detection. To discover counterfeiting, Saddique et al. [21] proposed adopting discrete texture analysis in successive frames. Christlein et al. [22] analyzed the effectiveness of characteristics for copy-move watermarking that used a multitude of conventional feature sets, including scale-invariant feature transform (SIFT) and speed-up robust features (SURF), and block-based features such as PCA, discrete wavelet transform (DWT), discrete cosine transform (DCT), and kernel principal component analysis (KPCA).

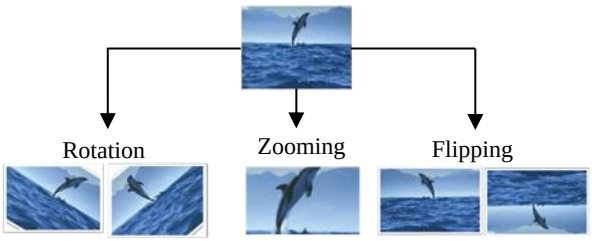
Using the concept of a near-neighbor-dense field, D'Amiano et al. [23] suggested a patch-match-based copy-move detection approach. However, while confronted with huge amounts of data, this method failed miserably. Wu et al. [24] suggested a method for detecting frame duplication and frame deletion in vector flow picture sequences by observing velocity and discontinuity peaks. In the moving picture expert group (MPEG) [25] videos, it presented a video forgery copy-move detection algorithm. To determine the optical flow coefficient for each region, their method separates each video frame

into suspected cleared sections. When an uncommon trend in the optical flow coefficient object is found, it indicates counterfeiting. Singh and Singh [26] proposed a passive blind approach that uses the correlation coefficient and coefficient of variation to detect duplicate frames.

Wang et al. [27] offered discrete wavelet packet deconstruction and singular point analysis of speech data, to identify audio tampering of time-domain such as audio recognition, addition, replacement, and slicing. It provides a technique for measuring reverberation length for identifying indicators of tampering in audio tapes. They compared pitch to format sequences to detect copy-move forgeries in audio recordings. To identify places of copy-move fraud in a video, a histogram is computed using LBP and a comparison technique is applied. Detailed analysis of image and video forgery along with fake video datasets used for tampering has been demonstrated [28].

3. Dataset and Attributes

This part describes the dataset used for both video and audio forgery detection. The attributes of the datasets such as the number of files, description of files, etc. are mentioned. The representation technique of the audio files is researched and an explanation for choosing the mel-spectrogram has been given.



(a) Frame augmentation before insertion

(b) Inserting augmented frames in a random position

(c) Deleting the original frame from a random position

Fig. 1 Temporal forgery techniques implemented in dataset creation

(1) Video forgery: A custom dataset for temporal forgery detection has been developed by modifying the dashcam dataset containing 1544 videos [29].

(2) Creating a custom dataset: The creation of a dataset for forgery detection has been achieved by introducing seven types of important temporal forgeries in the dashcam dataset, like insertion, deletion, duplication, flipping, rotations, and zooming forgery. Fig. 1 shows some of the forgery techniques implemented during dataset creation. Training data contains 9448 videos of the dashcam dataset containing non-tampered and tampered videos that are forged by the mentioned forgery techniques. Test data contains 2904 videos of the same type.

(3) Audio forgery: ASVspoof 2019 was created for the third automatic speaker verification spoofing and countermeasures challenge. Table 1 shows the number of audio files present in the ASVspoof 2019 logical access (LA) dataset according to their labels.

Table 1 ASVspoof 2019 data distributions

-	**Train**	**Dev**	**Eval**
Bonafide	2580	2548	7355
Spoof	22800	22296	63882
Total	25380	24844	71237

3.1. Audio representation technique

3.1.1. Spectrogram

Fourier transform is used to build spectrograms from sound sources. The Fourier transform displays the amplitude of each fundamental frequency after dividing a signal into its fundamental frequencies. A spectrogram breaks the length of a sound source into tiny window segments, which are then subjected to the Fourier transform to detect the frequencies contained within each window. Next, all of those windows' Fourier transforms are then integrated into a single plot. It plots frequency (y-axis) vs. time (x-axis) and uses different colors to show each frequency amplitude. The brightness of the color that represents the signal is proportional to its energy.

3.1.2. Chroma features

Chroma-based characteristics, also known as "pitch class profiles," are a useful tool for analyzing music with usefully categorized pitches (typically into twelve categories) and tuning which approximates the equal-temperament scale. Chromatic and melodic aspects of music are captured by chroma features, which are resistant to changes in timbre and instrumentation.

3.1.3. Mel-spectrograms

A mel-spectrogram happens to a spectrogram where the frequencies exist convinced to the mel scale. It remaps the principles in hertz to the mel scale as shown in Fig. 2. The mel scale can be termed as the scale of pitches perceived by listeners to have the same distance from one another. General frequency measurement has a common reference point which can be defined by equating a 1000 Hz tone, with a pitch of 1000 mel and 40 dB greater than the listener's threshold.

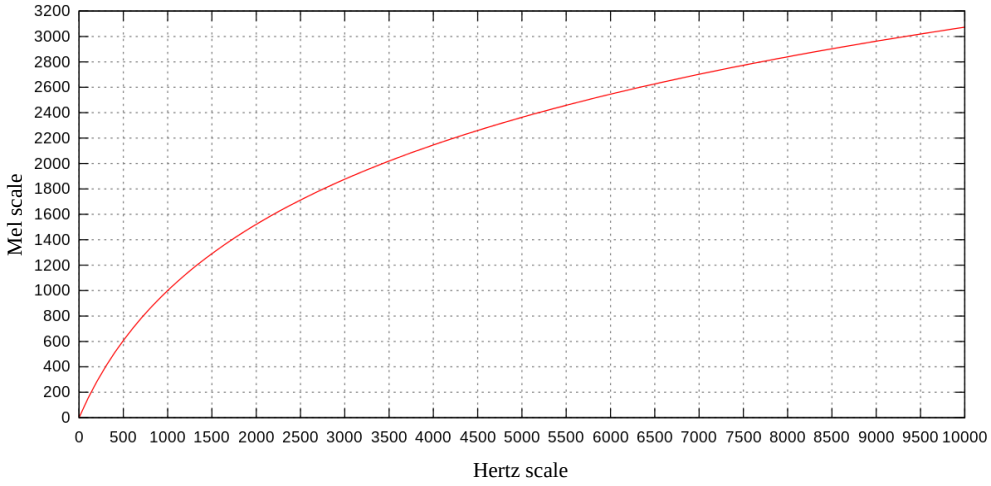


Fig. 2 Hertz vs. mel scale representation

The formula to convert frequency from hertz into mel scale can be expressed by:

$$m = 2595 \log_{10}\left(1 + \frac{f}{100}\right) \tag{1}$$

where m represents mel on the mel scale and f represents frequency in hertz.

A mel-spectrogram forms two influential changes relating to a normal spectrogram that plots frequency intersection time. It uses the mel scale as a suggestion of correction frequency in contact with the y-point around which something revolves. And the decibel scale is used as a suggestion of correction amplitude to signify banners. The proposed research uses the mel-spectrogram-based dataset and it is considered better than other audio representation techniques. The reason has also been demonstrated and experimented on in the experimental section.

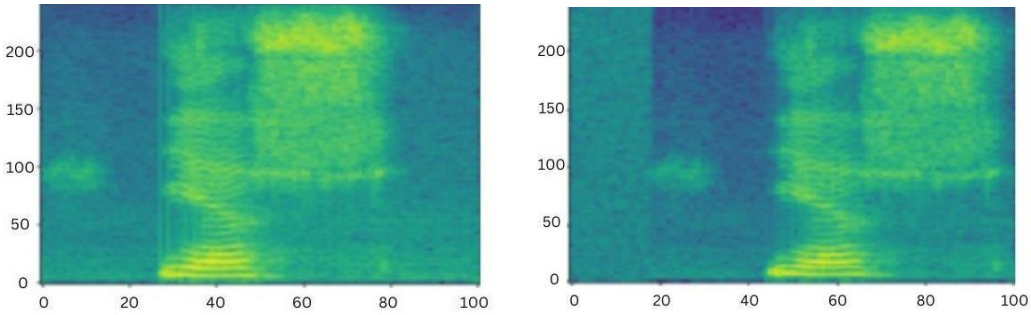
3.2. Data augmentation

The training, development, and assessment datasets of ASVspoof 2019 have a lot of data imbalance, as shown in Table 1. As a result, this research shows that developing an augmentation framework that can generate mel-spectrograms from the current datasets while also addressing the dataset's data imbalance problem, and allowing the network to learn more valuable features.

SpecAugment changes the spectrogram by distorting it in time, masking blocks of successive frequency channels, and masking blocks of utterances in time. Time shifting, time masking, and frequency masking are the three primary methods for augmenting data.

3.2.1. Time shifting

In time shifting, the audio is moved linearly from the left or right with a random second in time shifting. Here, the audio is fast-forwarded by a certain interval of x seconds, the first of these x seconds is marked as 0, i.e., silence. Then, the shift of the audio to the right (backward) for x seconds again, and the last x seconds are marked as 0, i.e., silence. Fig. 3 shows the spectrogram of the original audio and time-shifted audio.



(a) Spectrogram of original audio (b) Spectrogram of time shifted audio

Fig. 3 Spectrogram of time shifted audio

3.2.2. Frequency masking

Masked frequency channels are [f0, f0 + f]. Here f0 is selected from (0, v-f), where v denotes the number of frequency channels, and the selection of f is made from a uniform distribution ranging from 0 to masking parameter F. Fig. 4 shows the masked mel-spectrogram using frequency masking.

3.2.3. Time masking

As shown in Fig. 5, while doing time masking, the masking of t sequential steps of time [t0, t0 + t] is obtained. The t is selected within a uniform distribution ranging from 0 to a masking parameter T, and from [0, τ – t] t0 is selected. Here, τ is the length of the audio file.

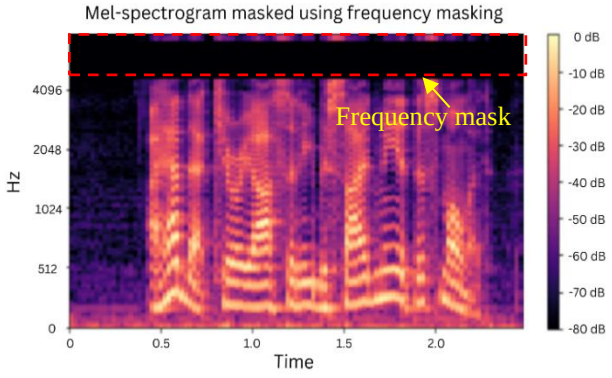


Fig. 4 Frequency masking

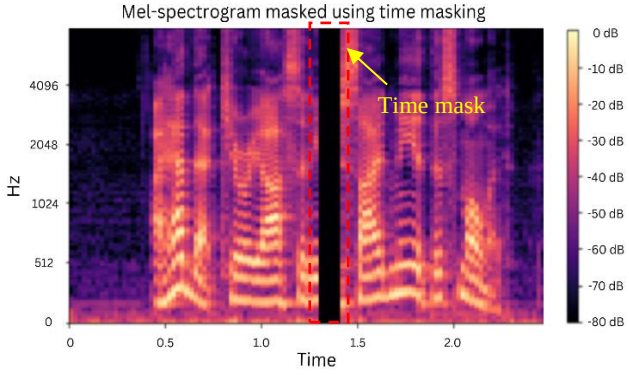


Fig. 5 Time masking

4. Methodology

4.1. Video forgery

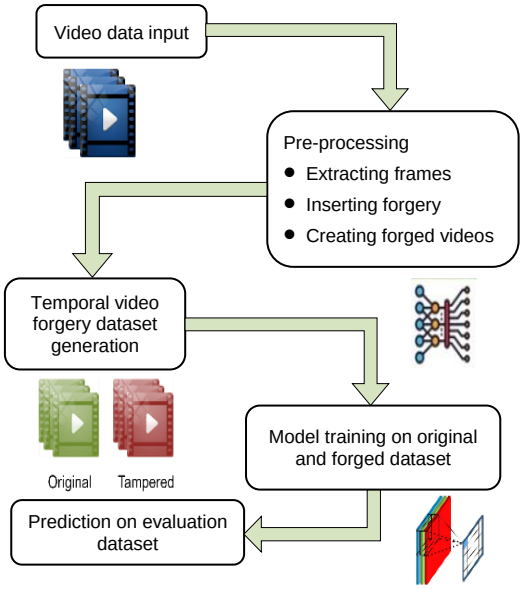


Fig. 6 Proposed methodology flow diagram

Step 1: The proposed model as shown in Fig. 6 takes data in the form of multiple sequences of images from videos. The dashcam dataset consists of 1181 videos for training and 363 videos for testing. As shown in Fig. 1, after applying the forgery techniques to the existing dashcam dataset, a new dataset with a total of 9448 videos for training and 2904 videos for testing is produced. It is used for video forgery analysis, model training, and validation.

Step 2: For training, the model takes data in the form of video and extracts some sequences of frames from the video. These multiple sequences of videos are labeled with the category of forged and original, depending upon the type of video from which these frames are extracted.

Step 3: For making the operation of frame extraction and sequencing faster, pre-extracted frames from the videos are kept in storage. The starting point of the sequence of frames is randomly chosen to avoid overfitting on specific time instances. The length of the sequence or clip is a data-dependent hyperparameter (depends upon video length). The labeled data created in this procedure is passed to the model for training and validation purposes.

Step 4: The model contains multiple convolutional 3D layers which convolve the sequence of frames to a 3D volume of features as shown in Fig. 7. From the convolved output, the model chooses important features using max pooling 3D layers. The dropout layers are added in between the series of max pooling 3D and convolutional 3D layers to avoid model overfitting. The model contains Relu as an activation function for the neurons. The output layer consists of two neurons that contribute two classes named forged and original. As the model is a classification problem, it uses a categorical cross-entropy loss function. A stochastic gradient descent optimizer is used to overcome the overhead of the gradient descent algorithm and a learning rate scheduler is used to decay the learning rate with an increase in the number of epochs. The Tesla K80 GPU is used for training this model.

Step 5: Table 2 shows the observed metrics while training the model. The accuracy and loss are 85.55% and 30.74% respectively. The precision (positive predictive value) and recall (sensitivity) values are 86.51% and 74.75% respectively. Table 3 shows the validation metrics for the model. The accuracy and loss are 82.17% and 35.87% respectively.

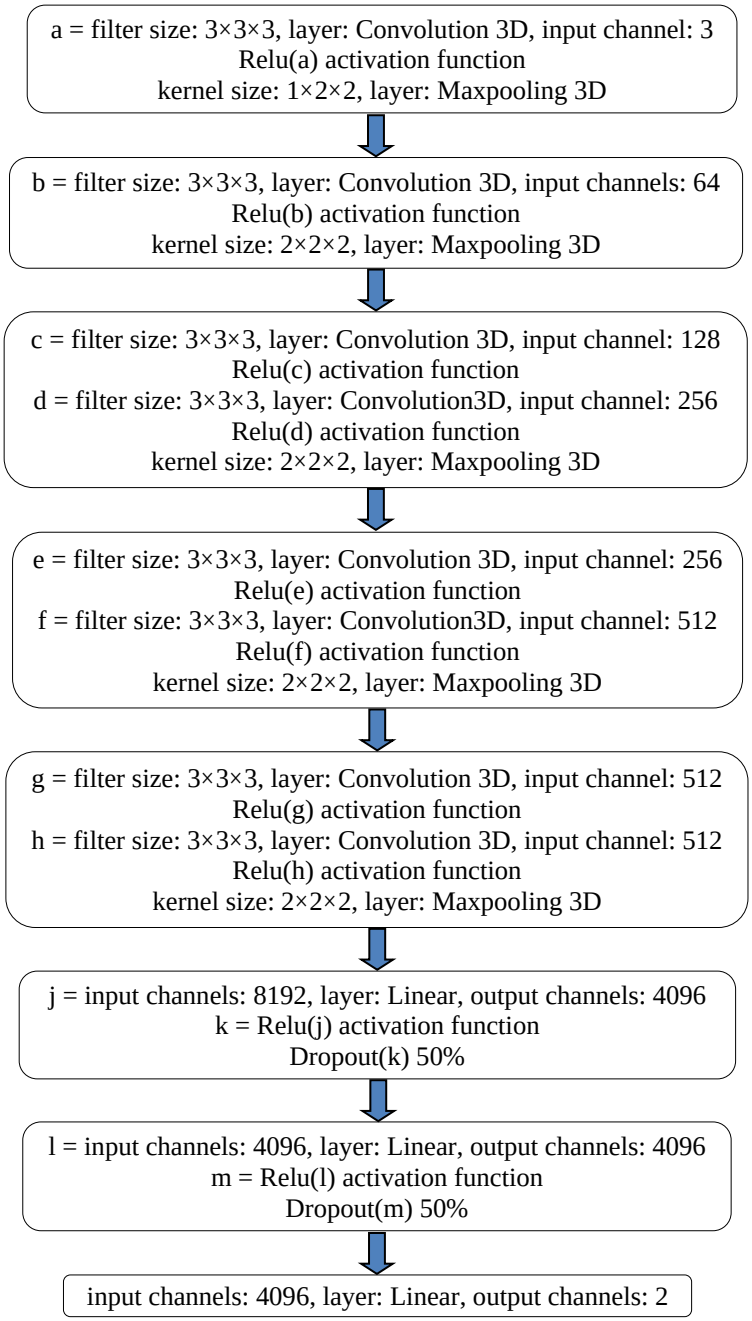


Fig. 7 Proposed model of CNN architecture with hyperparameters

Table 2 Training metrics

Accuracy	0.8555
Loss	0.3074
Precision	0.8651
Recall	0.7475

Table 3 Validation metrics

Accuracy	0.8217
Loss	0.3587

The formulas for calculating accuracy and loss are shown below:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{2}$$

$$Loss = \frac{FP + FN}{FP + FN + TP + TN} \tag{3}$$

The terms false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN) are derived from the confusion matrix in Fig. 8.

Fig. 8 Confusion matrix of the proposed model prediction on the validation dataset

4.2. Audio forgery

The proposed study takes the training audio data files from the ASVspoof 2019 LA dataset section and develops a mel-spectrogram representation of each audio file. Data augmentation has been used to address the problem of data imbalance in the dataset. Next, the ResNet-34 model applies transfer learning to the dataset and supplemented data. The data is then divided into three categories: real, spoofed_TTs, and spoofed_VC. The proposed methodology for the research is described in the next section.

Step 1: This study proposes a voice classifier based on deep convolutional neural networks for detecting spoofing attempts with the help of the ASVspoof 2019 dataset and required pre-processing.

Step 2: In the suggested technique, the train folder is used from the ASVspoof 2019 dataset and an audio time-frequency model of power spectral densities on the mel frequency scale (mel-spectrogram).

Step 3: For deeper residual training, (80-20) train and validation split (for transfer learning on ResNet-34 architecture) are designed. The fastai package and the Tesla K80 GPU are used to implement this transfer learning approach. The proposed methodology is shown in Fig. 9.

***Transfer learning*:** Given the significant computing and time resources required to create neural network models for this challenge, and the significant improvements in the skill that they provide on related problems. It helps in improving the DL models using pre-trained models as a preliminary step in computer vision.

***A residual network, or ResNet for short*:** It is an artificial neural network that uses skip connections or shortcuts to bypass some layers in the creation of a deeper neural network. Skipping enables the creation of deeper network layers without having to deal with vanishing gradients.

Step 4: On the validation split formed on the train folder files, the first epoch provides an accuracy of 92.47%. A total of 12 epochs have been executed for audio forgery detection on the ASVspoof dataset. After every 4 epochs, the learning rate is identified to get minimum loss and then changed for the further epochs.

Step 5: Better accuracy of 99% with 0.03% validation loss on the validation set is achieved after performing 12 epochs and fine-tuning the learning rate of the proposed model. Fig. 10 depicts the relationship between loss and learning rate. The final model metrics are shown in Fig. 11.

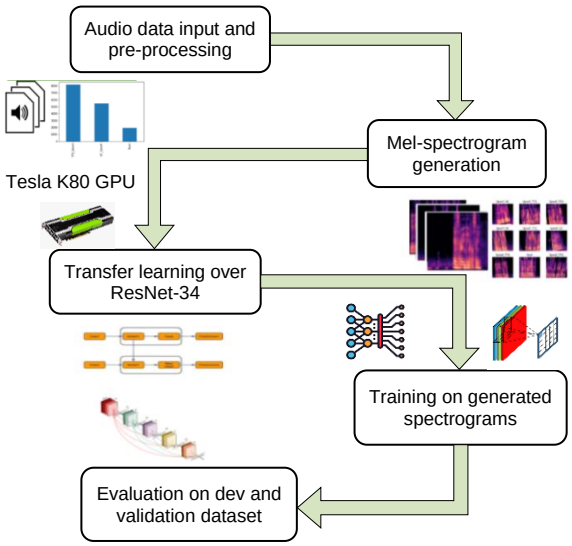


Fig. 9 Proposed methodology flow diagram

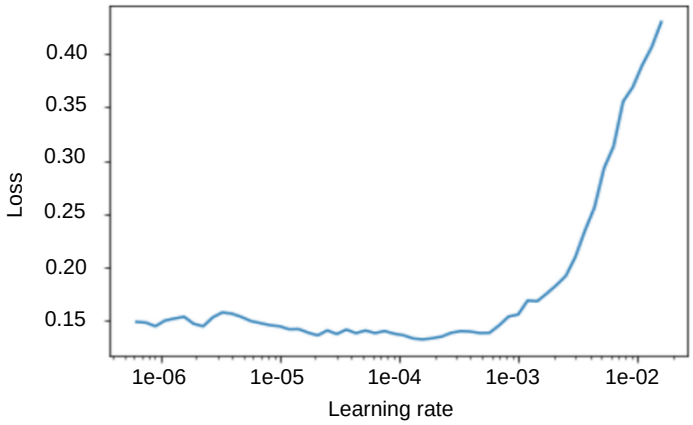


Fig. 10 Loss versus learning rate graph (for fine-tuning)

epoch	train_loss	valid_loss	accuracy	time
0	0.071467	0.016688	0.995666	09:09
1	0.038840	0.007211	0.998227	09:19
2	0.027787	0.005167	0.999015	09:10
3	0.017552	0.003989	0.999015	09:14

Fig. 11 Metrics of the final model preparation

5. Experiments

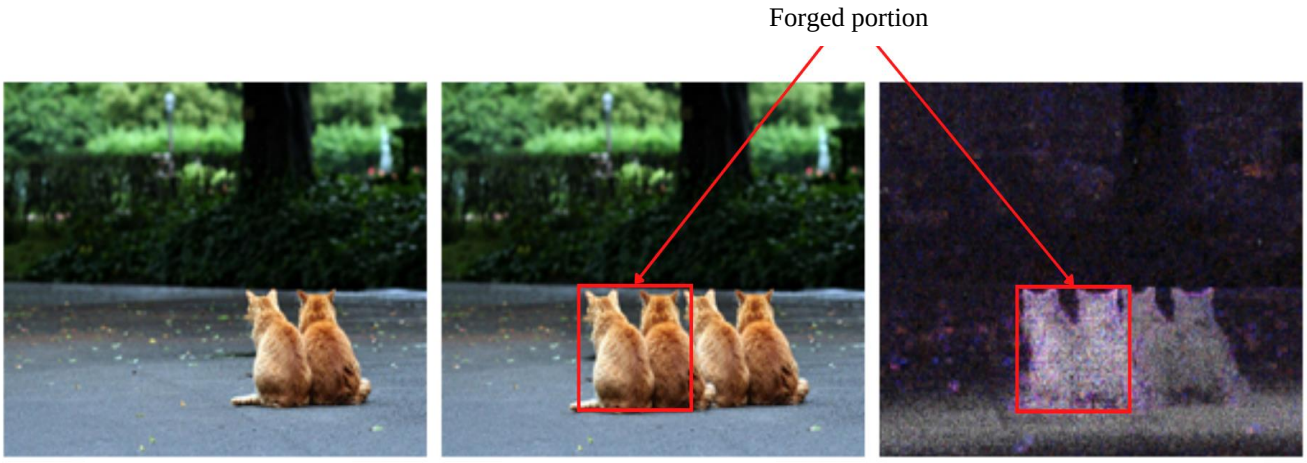
5.1. Video forgery

(1) Error level analysis (ELA)

ELA is used to recognize parts of the picture with varying compression rates. One of the major drawbacks of ELA is that it provides inaccurate recognition when low-quality JPEG images and recoloring are considered [30]. Using the ELA technique, ELA-processed images for the input images (CASIA2 dataset) are generated as shown in Fig. 12.

These newly generated images which undergo ELA processing are passed to a 2D convolutional neural network with labels attached as “original” and “forged”. The model comes up with an accuracy of 90.28% for the validation part. This experiment is done to see the performance changes in the classification of forged videos frame-by-frame. The outcome is that the performance of this ELA metric-based model is not as efficient as the proposed model which works on convolutional

3D CNN layers. Here, CASIA 2 dataset has been used. By using ELA, forgery in the spatial domain is detected. This approach can be used to detect a spatial forgery in videos by separating the frames and doing ELA analysis frame-by-frame.



Original image Forged image ELA processed image

Fig. 12 Input image versus forged image versus image after ELA processing

(2) Discrete cosine transform (DCT)

The DCT is a mathematical modification that is essential to the JPEG standard compaction. The primary objective of these procedures is to change a signal from one sort of interpretation to another. The DCT may be used to transform the signal (intra-frame information) into quantitative data ("frequency" or "spectral" information), allowing the image to be quantified and compressed.

CASIA 2 dataset is used here. DCT is applied to the frame for compression. The difference between the original and compressed frame is obtained to identify the spatial forgery in the frame. DCT coefficients are used to identify irregularities due to the spatial domain analysis caused by superimposing an image over another one.

(3) Using image processing (a non-AI approach)

The structural similarity index measure (SSIM) checks the similarity between two images by the standard deviation of pixel values of the image. These become the factors that can be used to detect some types of forgeries such as insertion, duplication, copy-moving, and removal of the region of a frame in a video.

SSIM is a perception-based model that analyzes image deterioration as a perceived change in structural information, as well as crucial perceptual appearances, such as contrast and luminance masking. Structural information means the assumption that pixels have a lot of interdependencies, especially when they're close together in a space.

In the proposed approach, SSIM and the standard deviation have been used to detect and analyze forgeries that can be embedded between consecutive video frames. This has been done by computing and analyzing the SSIM value between two consecutive frames of a video, along with calculating the difference between the standard deviation of pixel values of these two frames.

A sudden change in standard deviation values of a frame in a video sequence with a very low similarity index between consecutive frames depicts a high probability of forgery. Meanwhile, a frame window is proportionally divided into several segments. SSIM and the standard deviations are calculated for each segment and compared with corresponding segments in the previous frame. This results in more accurate forgery detection and localization of forgery. Fig. 13 shows the entire flow diagram of the process.

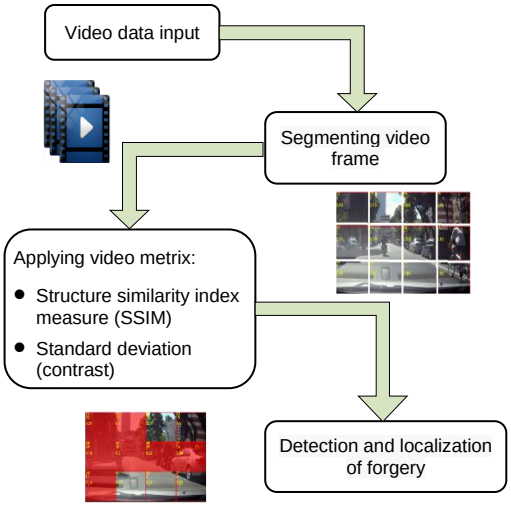
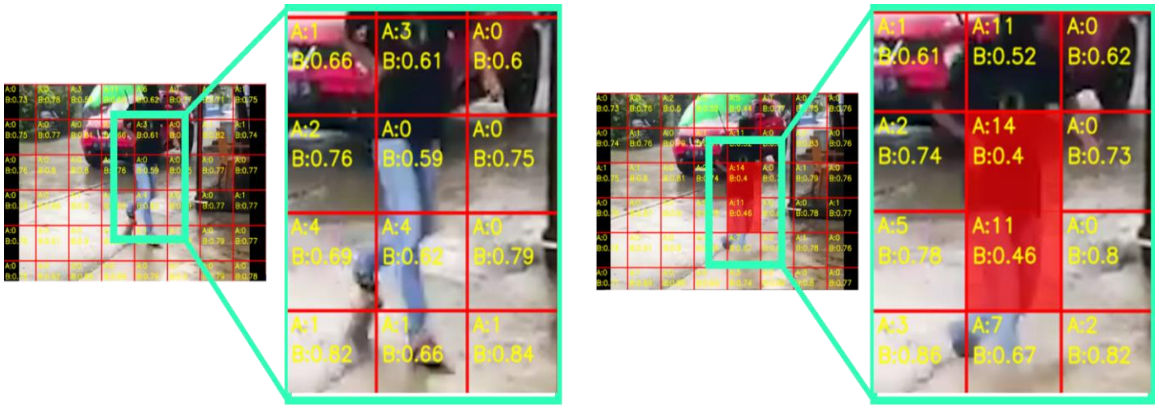


Fig. 13 Non-AI video forgery detection flow diagram



(a) Frame 1: Untampered video frame segment

(b) Frame 2: Tampered video frame segments in next consecutive frame

Fig. 14 Non-AI approach illustration to identify video forgery

In Fig. 14, "A" represents the difference between the standard deviation values of the current frame segment and the corresponding segment of the frame previous to it. "B" represents the SSIM value between the consecutive frame segments. Identification of tampered frame segments is made by setting a minimum threshold to SSIM values and a maximum threshold to the difference of standard deviation values between consecutive frame segments. The red highlighted segments indicate forgery in a particular section of the video. By using this technique, temporal forgery analysis of any video can be done.

5.2. Audio forgery

Table 4 shows the comparison of the metrics such as accuracy and F1 scores with other chart types (spectrogram, chroma STFT). The proposed model achieves a better performance under experimental conditions. By observing the chart type comparison, chroma STFT gives the least accuracy and F1 scores, whereas the mel-spectrogram gives the best accuracy and F1 scores.

Fig. 15 shows the various methods to convert the visual and audio media transmitted via radio wave signals to an image. It can be seen that the reason behind choosing the mel-spectrogram over the differing present methods is that the spectrogram gives a short "snapshot" of visual and audio media transmitted via radio waves. Therefore, it is suitable to recommend CNN-located architectures grown for management representation. Fig. 14 shows the confusion matrix generated on the validation part of the ASVspoof 2019 LA dataset.

The counts of FP and FN are quite less than TP and TN. The wrong predictions inform that type 1 errors are more than type 2. However, the error rate is very less on the whole and the errors observed in some samples were because a lot of tweaking/augmentation was done on the testing set.

Table 4 ASVspoof 2019 validation set accuracy versus graphing approach

Chart type	Accuracy	F1 score	Classes
Mel-spectrogram (proposed)	99%	99%	Real
		99%	Spoof_TTS
		98%	Spoof_VC
Spectrogram	90.33%	90.52%	Real
		95.83%	Spoof_TTS
		86.34%	Spoof_VC
Chroma STFT	82.50%	73.32%	Real
		93.52%	Spoof_TTS
		80.10%	Spoof_VC

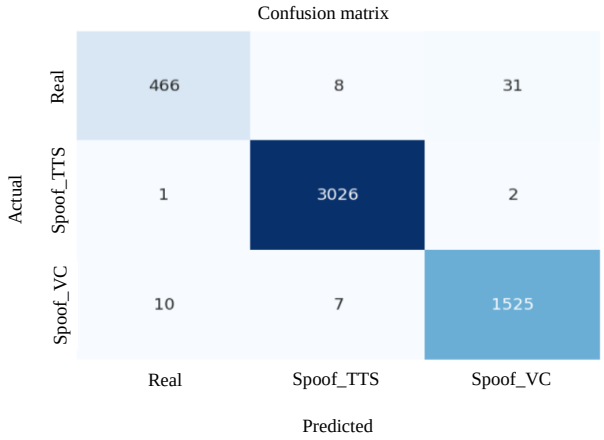


Fig. 15 Confusion matrix of the proposed model prediction on the validation dataset

6. Conclusions

To improve the accuracy and quality of video and audio forgery identifications, two models for their detection are proposed. The experiments lead to the following conclusions:

(1) This research on video temporal forgery identification fills the gap in existing work on inter-frame forgery detection due to the lack of temporal video forgery detection. It proposes the use of convolutional 3D layer model architecture with an accuracy of 85.55%. Also, a non-AI technique has been developed using metrics like SSIM and the standard deviations of the video frame segments to identify runtime temporal forgery. A comprehensive dataset for temporal forgery identification has been created for future research.

(2) In audio forgery identification, ASVspoof 2019 dataset using transfer learning is proposed. Moreover, it proposes a comparative study on various audio representation techniques and a study on why the mel-spectrogram is efficient for audio data. Augmentation of data has been done to handle the data imbalance problem.

(3) The computational complexity in CNN models utilized in the audio and video forgery algorithms, the number of parameters in each feature map is limited to a constant (usually less than 1) multiplied by the input pixels n. Convolving a fixed length filter over an image with n pixels requires O(n) time because each output is just the sum-product of k pixels in the image and k weights in the filter, and k is constant with n. Similarly, every max or avg pooling operation takes no more than linear time in terms of input size. Hence, the entire runtime remains linear O(n).

(4) Both the video and audio forgery models do not incur any computational overhead. All the processing is done as the requirements of the proposed model. Overall, the proposed models achieved the optimal accuracy performance of 99% on the validation dataset with minimal loss. Future work of this research can be directed to combining the video and audio forgery detection works. One way of doing this is by extracting audio and visual parts of video and feeding them to respective models. Outputs of both models can be combined to generate the final result.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] S. Fadl, Q. Han, and Q. Li, "CNN Spatiotemporal Features and Fusion for Surveillance Video Forgery Detection," Signal Processing: Image Communication, vol. 90, article no. 116066, January 2021.

[2] Y. B. Deshmukh and S. K. Korde, "Forensic Video/Image Analytics – A Deep Learning Approach," International Journal of Creative Research Thoughts (IJCRT), vol. 8, no. 9, pp. 411-418, September 2020.

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network,'' IEEE International Workshop on Information Forensics and Security (WIFS), article no. 8630761, December 2018.

[4] J. Xiao, S. Li, and Q. Xu, "Video-Based Evidence Analysis and Extraction in Digital Forensic Investigation," IEEE Access, vol. 7, pp. 55432-55442, April 2019.

[5] P. Ghadekar, P. Maheshwari, R. Shah, A. Shaha, V. Sonawane, and V. Shetty, "Video Forgery Dataset," https://www.kaggle.com/datasets/rajshah1/video-forgery-dataset, September 10, 2022.

[6] H. Malik and H. Farid, "Audio Forensics from Acoustic Reverberation," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1710-1713, March 2010.

[7] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital Audio Forensics: A First Practical Evaluation on Microphone and Environment Classification," MM&Sec '07: Proceedings of the 9th Workshop on Multimedia & Security, pp. 63-74, September 2007.

[8] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, et al., "ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database," [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR), https://doi.org/10.7488/ds/2555.

[9] I. I. I. Richard and V. Roussev, "Digital Forensic Tools: The Next Generation," Digital Crime and Forensic Science in Cyberspace, IGI Global, pp. 75-90, April 2006.

[10] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," University of Washington, Technical Report, article no. 1804.02767, April 2018.

[11] H. Farid, Photo Forensics, The MIT Press, February 2019.

[12] D. Güera, Y. Wang, L. Bondi, P. Bestagini, S. Tubaro, and E. J. Delp, "A Counter-Forensic Method for CNN-Based Camera Model Identification," IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1840-1847, July 2017.

[13] D. Güera, F. Zhu, S. K. Yarlagadda, S. Tubaro, P. Bestagini, and E. J. Delp, "Reliability Map Estimation for Cnn-Based Camera Model Attribution," IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 964-973, March 2018.

[14] P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, "Local Tampering Detection in Video Sequences," IEEE 15th International Workshop on Multimedia Signal Processing (MMSP), pp. 488-493, September-October 2013.

[15] D. Graupe, "Principles of Artificial Neural Networks," Advanced Series in Circuits and Systems, Vol. 7, World Scientific, 2013.

[16] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-Stream Neural Networks for Tampered Face Detection," IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1831-1839, July 2017.

[17] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces," arXiv preprint, article no. 1803.09179, March 2018.

[18] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming Auto-Encoders," Artificial Neural Networks and Machine Learning – ICANN 2011, Lecture Notes in Computer Science, vol. 6791, pp. 44-51, 2011.

[19] R. Sharma and A. Singh, "An Integrated Approach towards Efficient Image Classification Using Deep CNN with Transfer Learning and PCA," Advances in Technology Innovation, vol. 7, no. 2, pp. 105-117, April 2022.

[20] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic Routing Between Capsules," Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), pp. 3859-3869, December 2017.

[21] M. Saddique, K. Asghar, U. I. Bajwa, M. Hussain, and Z. Habib, "Spatial Video Forgery Detection and Localization Using Texture Analysis of Consecutive Frames," Advances in Electrical and Computer Engineering, vol. 19, no. 3, pp. 97-108, 2019.

[22] V. Christlein, C. Riess, J. Jordan, C. Riess, and E. Angelopoulou, "An Evaluation of Popular Copy-Move Forgery Detection Approaches," IEEE Transactions on Information Forensics and Security, vol. 7, no. 6, pp. 1841-1854, December 2012.

[23] L. D'Amiano, D. Cozzolino, G. Poggi, and L. Verdoliva, "A PatchMatch-Based Dense-Field Algorithm for Video Copy–Move Detection and Localization," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 3, pp. 669-682, March 2019.

[24] Y. Wu, X. Jiang, T. Sun, and W. Wang, "Exposing Video Inter-Frame Forgery Based on Velocity Field Consistency," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2674-2678, May 2014.

[25] G. Ulutas, B. Ustubioglu, M. Ulutas, and V. V. Nabiyev, "Frame Duplication Detection Based on Bow Model," Multimedia Systems, vol. 24, no. 5, pp. 549-567, October 2018.

[26] G. Singh and K. Singh, "Video Frame and Region Duplication Forgery Detection Based on Correlation Coefficient and Coefficient of Variation," Multimedia Tools and Applications, vol. 78, no. 9, pp. 11527-11562, May 2019.

[27] Z. Wang, Y. Yang, C. Zeng, S. Kong, S. Feng, and N. Zhao, "Shallow and Deep Feature Fusion for Digital Audio Tampering Detection," EURASIP Journal on Advances in Signal Processing, vol. 2022, article no. 69, 2022. https://doi.org/10.1186/s13634-022-00900-4

[28] F. H. Chan, Y. T. Chen, Y. Xiang, and M. Sun, "Anticipating Accidents in Dashcam Videos," Computer Vision – ACCV 2016, vol. 10114, pp 136-153, 2016.

[29] S. Tyagi and D. Yadav, "A Detailed Analysis of Image And Video Forgery Detection Techniques," The Visual Computer, 2022, in press. https://doi.org/10.1007/s00371-021-02347-4

[30] I. B. K. Sudiatmika, F. Rahman, T. Trisno, and S. Suyoto, "Image Forgery Detection Using Error Level Analysis and Deep Learning," Telecommunication Computing Electronics and Control (TELKOMNIKA), vol. 17, no. 2, pp. 653-659, April 2019.