iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XI **Month of publication:** November 2023

DOI: https://doi.org/10.22214/ijraset.2023.57056

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

Deepfake Video Detection Using LSTM and XRESNET

Varun P Shrivathsa
:*Dayananda Sagar University, Bengaluru, India varunpshrivathsa@gmail.com*

Abstract: ***As the rapid evolution of Artificial Intelligence continues, it becomes increasingly crucial to implement robust measures for monitoring and combating the proliferation of Deepfake videos. In this proposed method, frame-level features are extracted from videos using the XResNet convolutional neural network. These extracted features serve as the foundation for training the LSTM (Long Short-Term Memory) Recurrent Neural Network, enabling it to classify videos as either real or fake. Our dataset originates from Meta DFDC (Deepfake Detection Challenge) videos, selected for both the training and testing phases of our model. This model is able to predict a video with an accuracy of 83.3%.***
Index Terms: XResnet, LSTM, Recurrent Neural Networks, Convolution Neural Networks

I. INTRODUCTION

This project amalgamates two potent neural network ar- chitectures: XResNet, an evolution of ResNet designed for efficiency and performance, and LSTM (Long Short-Term Memory), known for its ability to understand temporal depen- dencies in sequential data. By combining these technologies, the project aims to create a comprehensive system capable of analyzing spatial and temporal features within video data, enhancing the accuracy and reliability of deepfake detection. XResNet is vital in the initial stages of the pipeline, utilizing its prowess in feature extraction. With a focus on facial features crucial for deepfake identification, XResNet captures intricate details and patterns, providing a strong foundation for subsequent analysis. The architecture's efficiency is par- ticularly advantageous, ensuring that deep learning models can operate effectively even in resource-constrained envi- ronments. The 128 facial landmarks serve as key points of reference, enabling a deeper understanding of facial dynamics and expressions in various real-world scenarios. In essence, the fusion of XResNet with 128 facial landmark detection advances the accuracy of feature extraction. Complementing XResNet, the project leverages LSTM to analyze the temporal dynamics of video sequences. Videos are inherently sequen- tial, and LSTM's ability to retain information over extended sequences proves invaluable. The temporal analysis augments the spatial understanding provided by XResNet, offering a holistic perspective on video content. The dataset employed in this project is sourced from the Meta Deepfake Detection Challenge (DFDC), which provides a diverse range of au- thentic and manipulated videos for training and evaluation. The neural networks are trained to discern subtle patterns indicative of deepfake manipulations, achieving an accuracy rate of 83.3%. Continuous monitoring and updates ensure the system's adaptability to evolving deepfake techniques.

II. XRESNET AND RELU ACTIVATION FUNCTION

XResNet builds upon the foundational concepts of ResNet, enhancing them with key features to optimize performance.

At the core lies the deployment of residual blocks, which have proven instrumental in training exceptionally deep net- works. These blocks introduce shortcut connections, address- ing the vanishing gradient problem and facilitating the smooth flow of gradients through the network. This enables the successful training of models with an extended number of layers.

A distinguishing feature of XResNet is the incorporation of a "stem block" at the beginning of the architecture. This block serves as the initial layer, efficiently capturing essential fea- tures from the input data and setting the stage for subsequent operations.

Convolutional layers play a central role in XResNet's ar- chitecture, responsible for learning hierarchical features from the input data. These layers contribute to the network's ability to understand complex spatial patterns and representations in the data. By focusing on the intricacies of facial patterns and expressions, it empowers deep fake detection systems with the ability to discern subtle variations indicative of manipulations.

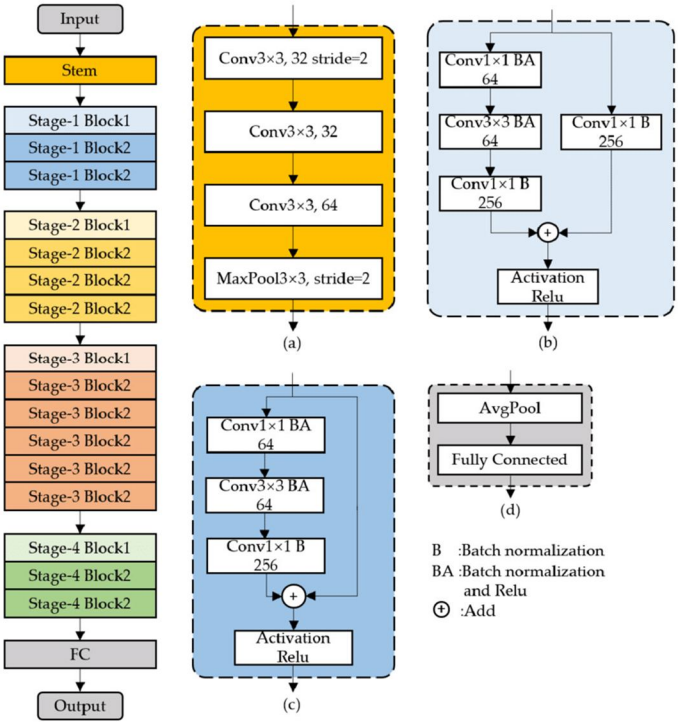


Fig. 1. Representation of XResNet Architecture

This application of the activation function is fundamental for introducing non-linearities into the network, allowing it to capture intricate features in the input data. By replacing negative values with zero, ReLU facilitates the model's ability to learn and adapt to patterns, contributing to the network's representational power. The rectification operation ensures that the gradient remains non-zero for positive values, facilitating the backpropagation of errors during training. This is particu- larly crucial in the training of deep neural networks where the flow of gradients through many layers can be a challenge. The simplicity and efficiency of ReLU contribute to the efficient learning of manipulated content. XResNet, with ReLU activa- tions, can quickly adapt to the distinctive patterns associated with deepfake content, enhancing its detection capabilities.

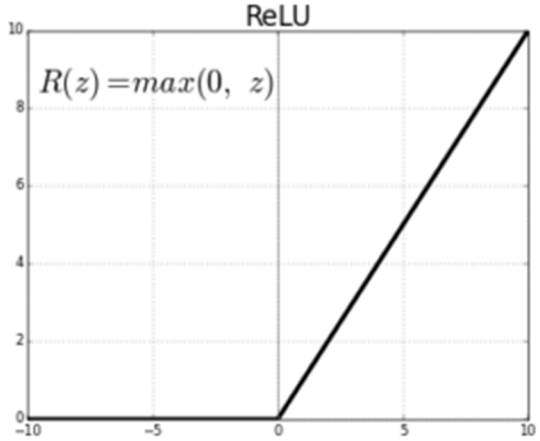


Fig. 2. Representation of ReLU function.

III. LONG-SHORT TERM MODEL

Long Short-Term Memory (LSTM) represents a pivotal advancement in recurrent neural network (RNN) architecture, specifically tailored to surmount challenges associated with capturing long-term dependencies within sequential data. At its core, an LSTM integrates a memory cell that spans the entirety of the sequence, affording the model the ability to selectively retain and retrieve information over prolonged intervals. This innovative approach effectively addresses the vanishing gradi- ent problem encountered in conventional RNNs and facilitates the nuanced learning of complex dependencies inherent in sequential datasets.

It integrates a memory cell that spans the entirety of the sequence, affording the model the ability to selectively retain and retrieve information over prolonged intervals. This innovative approach effectively addresses the vanishing gradient problem encountered in conventional RNNs and facilitates the nuanced learning of complex dependencies inherent in sequential datasets. The distinctiveness of LSTMs lies in their incorporation of three gates—forget, input, and output gates—which orchestrate the flow of information. The forget gate assesses what information from the cell state should be discarded or preserved, the input gate regulates the inclusion of new information into the cell state, and the output gate dictates what information should be outputted based on the cell state. Activation functions, such as sigmoid and tanh, play a crucial role in these operations, imparting non-linearity to the model.

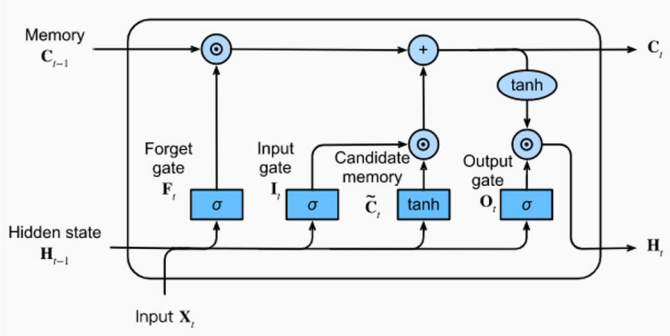


Fig. 3. Representation of LSTM Architecture.

During the training phase, the LSTM model is exposed to a dataset comprising both authentic and deepfake videos. Through the process of backpropagation, the model fine-tunes its parameters to distinguish between the temporal patterns present in genuine videos and those introduced by deepfake generation techniques.Its key contribution to deepfake detec- tion lies in its ability to detect temporal anomalies.

Once trained, the LSTM model is deployed for testing and inference on new video data. By scrutinizing the temporal dynamics of the sequence, the model provides predictions on whether the content is authentic or potentially a deepfake. Continuous monitoring and adaptation are crucial, as deepfake techniques evolve, necessitating regular updates and retrain- ing to maintain the model's effectiveness against emerging manipulation methods. Deepfake generation often introduces artifacts or imperfections in the synthesized content.

IV. DESIGN AND IMPLEMENTATION

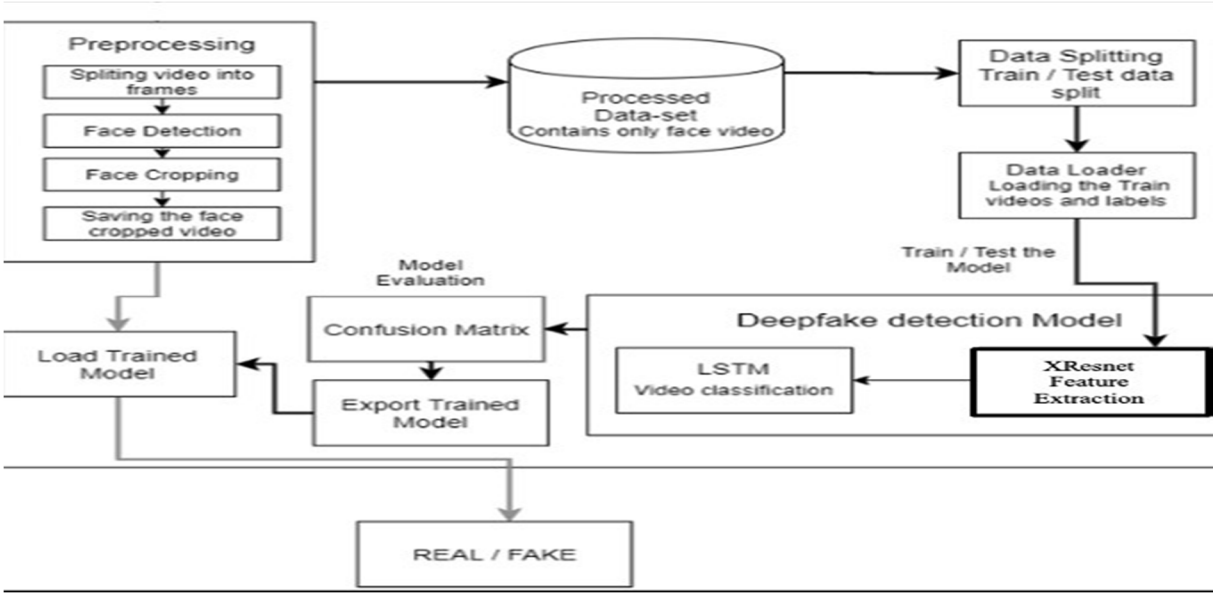


Fig. 4. Design Flowchart.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

A. Data Collection

The composition of the Meta DFDC dataset likely encom- passes a diverse array of videos, featuring both authentic and deepfake content. This diversity is instrumental for training a robust deepfake detection model, enabling it to general- ize effectively across various scenarios and against different manipulation techniques. An additional consideration is the availability of ground truth labels within the Meta DFDC dataset. These labels signify whether each video is authentic or a deepfake, serving as crucial annotations for supervised learning. Understanding the ground truth is fundamental in training a model that can effectively distinguish between genuine and manipulated content.

B. Preprocessing

Preprocessing is a fundamental stage in readying the data for effective training and inference. The first critical step involves frame extraction, breaking down videos into individual frames to establish a temporal sequence. Subsequently, face detection algorithms are applied to precisely locate and extract facial regions in each frame, and facial landmark detection is em- ployed to identify key points, facilitating consistent alignment across frames.

Fig. 5. Preprocessing and Cropping.

Once individual frames are obtained, image preprocessing steps are applied. This includes resizing frames to a consistent resolution, cropping to focus on the facial region, and normal- izing pixel values to a standard range. This normalization aids convergence during training and maintains numerical stability.

C. Training And Testing

```
[Epoch 1/20] [Batch 5 / 6] [Loss: 0.636132, Acc: 79.17%]Testing
[Batch 1 / 2]  [Loss: nan, Acc: 83.33%]
Accuracy 83.33333333333333
[Epoch 2/20] [Batch 5 / 6] [Loss: 0.473011, Acc: 87.50%]Testing
[Batch 1 / 2]  [Loss: nan, Acc: 83.33%]
Accuracy 83.33333333333333
[Epoch 3/20] [Batch 5 / 6] [Loss: 0.460424, Acc: 87.50%]Testing
[Batch 1 / 2]  [Loss: nan, Acc: 83.33%]
Accuracy 83.33333333333333
[Epoch 4/20] [Batch 5 / 6] [Loss: 0.442162, Acc: 87.50%]Testing
[Batch 1 / 2]  [Loss: nan, Acc: 83.33%]
Accuracy 83.33333333333333
[Epoch 5/20] [Batch 5 / 6] [Loss: 0.368797, Acc: 87.50%]Testing
[Batch 1 / 2]  [Loss: nan, Acc: 83.33%]
```

Fig. 6. Testing.

The training procedure includes the fine-tuning of hyper- parameters, such as learning rate and batch size, based on the model's performance on a validation set. Regularization techniques, including dropout, are incorporated to prevent overfitting and enhance the model's generalization capabili- ties. This rigorous training phase ensures that the XResNet- LSTM model is well-equipped to discern the intricate patterns associated with deepfake content

Moving to the testing phase, a separate test dataset, un- seen during training, is essential for an unbiased evaluation of the model's performance. Consistency is maintained by applying the same preprocessing steps to the test data as used during training. The model is evaluated using metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive assessment of its ability to classify authentic and deepfake sequences. Overall, this systematic approach ensures the effective training and evaluation contributing to its robustness and reliability.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

000.mp4	REAL
000_003.mp4	FAKE
001.mp4	REAL
001_870.mp4	FAKE
002.mp4	REAL
002_006.mp4	FAKE
003.mp4	REAL
003_000.mp4	FAKE

Fig. 7. CSV file for validation.

A dataset comprising 32 videos, a strategic partitioning of the data is undertaken, allocating 24 videos for the training of the model and reserving 6 videos for subsequent testing. This division is essential to gauge the model's ability to gen- eralize its learning to new and unseen data. The training set, consisting of the majority of videos, serves as the foundation for developing and fine-tuning the deepfake detection model.

```
train :  24
test :  6
TRAIN:  Real: 3  Fake: 21
TEST:  Real: 1  Fake: 5
```

Fig. 8. Splitting Data into training and testing.

In the testing phase of the deepfake detection model, which involved the evaluation of 6 videos not seen during training, the model demonstrated a commendable performance by correctly classifying 5 out of the 6 videos. This success underscores the model's ability to generalize its learning from the training set to previously unseen data, a critical aspect in assessing its real-world applicability.

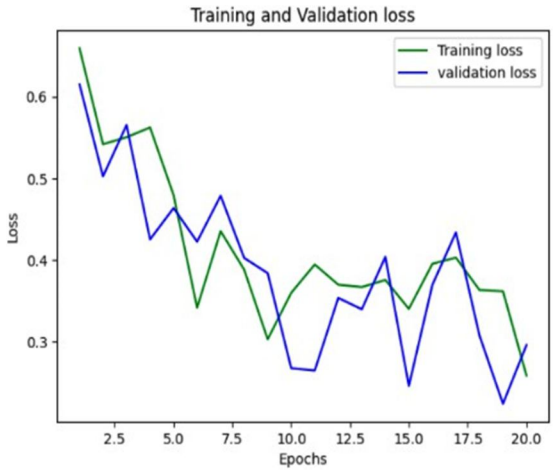


Fig. 9. Representation of Training and Validation loss.

The training loss, reflecting the error between the model's predictions and the actual labels within the training set, is fundamental for gauging how well the model is adapting to the intricacies of the training data.

A diminishing training loss over epochs implies that the model is effectively learning the features associated with both authentic and deepfake videos in the training dataset, continuously refining its predictive capabilities.the validation loss is computed on a distinct sub- set—the validation set—that the model has not encountered during training. This loss provides insights into the model's ability to generalize to unseen data. The goal is to minimize both training and validation losses, ensuring the model not only learns from the training data but also extends its predic- tive accuracy to new instances.Thus this graph states that the loss in training and validation are stabilizing simultaneously as the model is learning form multiple training sets.

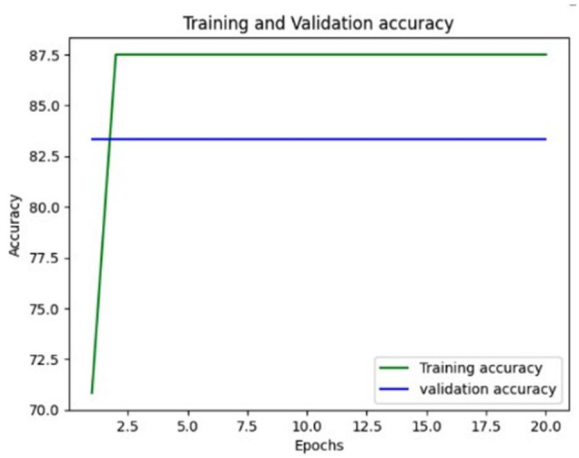


Fig. 10. Representation of Training and Validation accuracy

The training accuracy and validation accuracy graphs serve as essential tools for evaluating the performance and gener- alization capabilities of a deepfake detection model during its training process. The training accuracy, expressed as a percentage, reflects how well the model is learning to classify examples within the training dataset. An increasing training accuracy across epochs suggests that the model is becoming adept at correctly categorizing samples within the training set. model selection often corresponds to the epoch where validation accuracy is maximized while training accuracy continues to increase. This balance ensures that the model not only fits the training data well but also generalizes effectively to new, unseen examples.

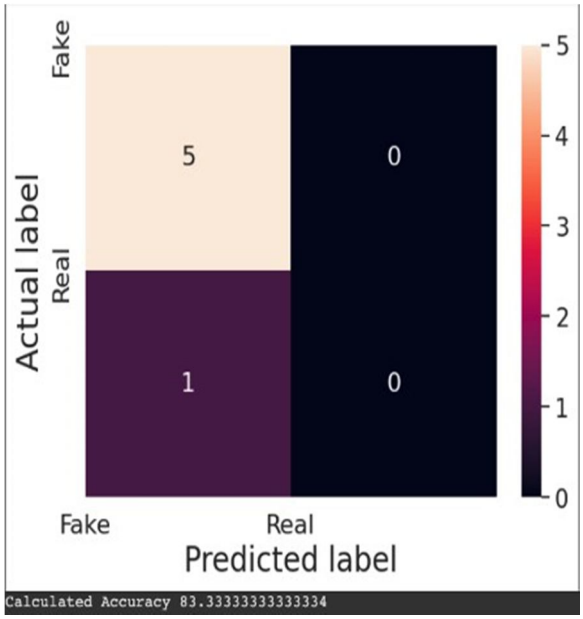


Fig. 11. Heat Map.

This heat map reveals which regions within a video frame play a pivotal role in the model's classification decision, whether it identifies content as authentic or manipulated.

```
[[5 0]
 [1 0]]
True positive =  5
False positive =  0
False negative =  1
True negative =  0
```

Fig. 12. Accuracy matrix.

An accuracy matrix, often derived from a confusion matrix, provides a comprehensive overview of a classification model's performance. It includes the following elements: True Positive (TP): The number of instances correctly predicted as posi- tive. True Negative (TN): The number of instances correctly predicted as negative. False Positive (FP): The number of in- stances incorrectly predicted as positive. False Negative (FN): The number of instances incorrectly predicted as negative.

D. Prediction

The prediction process itself involves applying the loaded model to the preprocessed video data. The XResNet com- ponent extracts spatial features, capturing intricate details in each frame, while the LSTM processes temporal patterns, considering the sequential nature of the frames in a video. The model outputs probability scores for each class (authentic or deepfake) based on these features.

V. CONCLUSION

In conclusion, the deepfake video detection project employ- ing XResNet and LSTM showcases a promising fusion of spatial and temporal feature extraction techniques for robust model performance. The utilization of XResNet facilitates ef- fective extraction of spatial features, capturing intricate details within each frame, while the LSTM contributes by modeling temporal dependencies, considering the sequential nature of video data. The project, trained and tested on the Meta DFDC dataset, achieved a commendable accuracy of 83.3%, indicating the model's ability to discern between authentic and deepfake content.

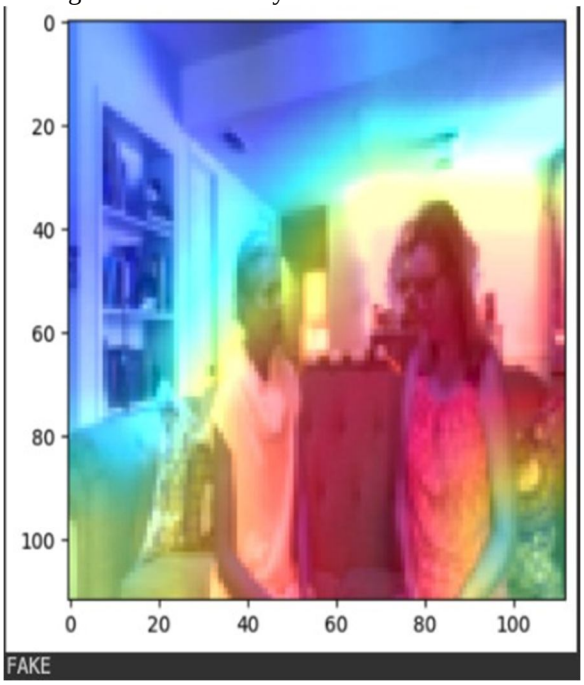


Fig. 13. Result-1.

The model is able to predict of the video is authentic or a fake and it also gives a optical flow chart indicating with red color on the presences of anomalies.Finally, the output of the LSTM, whether it be a refined feature representation or a prediction, is highlighted. The optical flow chart serves as a valuable tool for understanding the intricate changes and information dynamics within the LSTM network, providing an intuitive visual representation of its capabilities in handling sequential data.

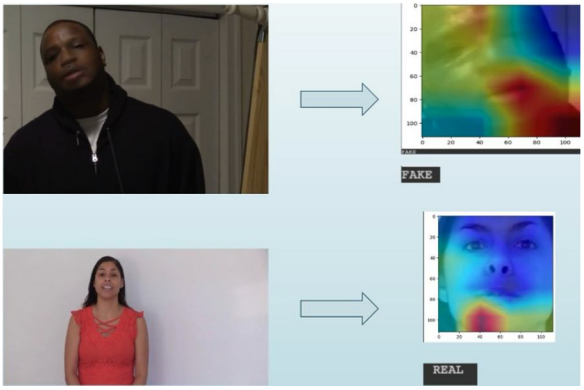
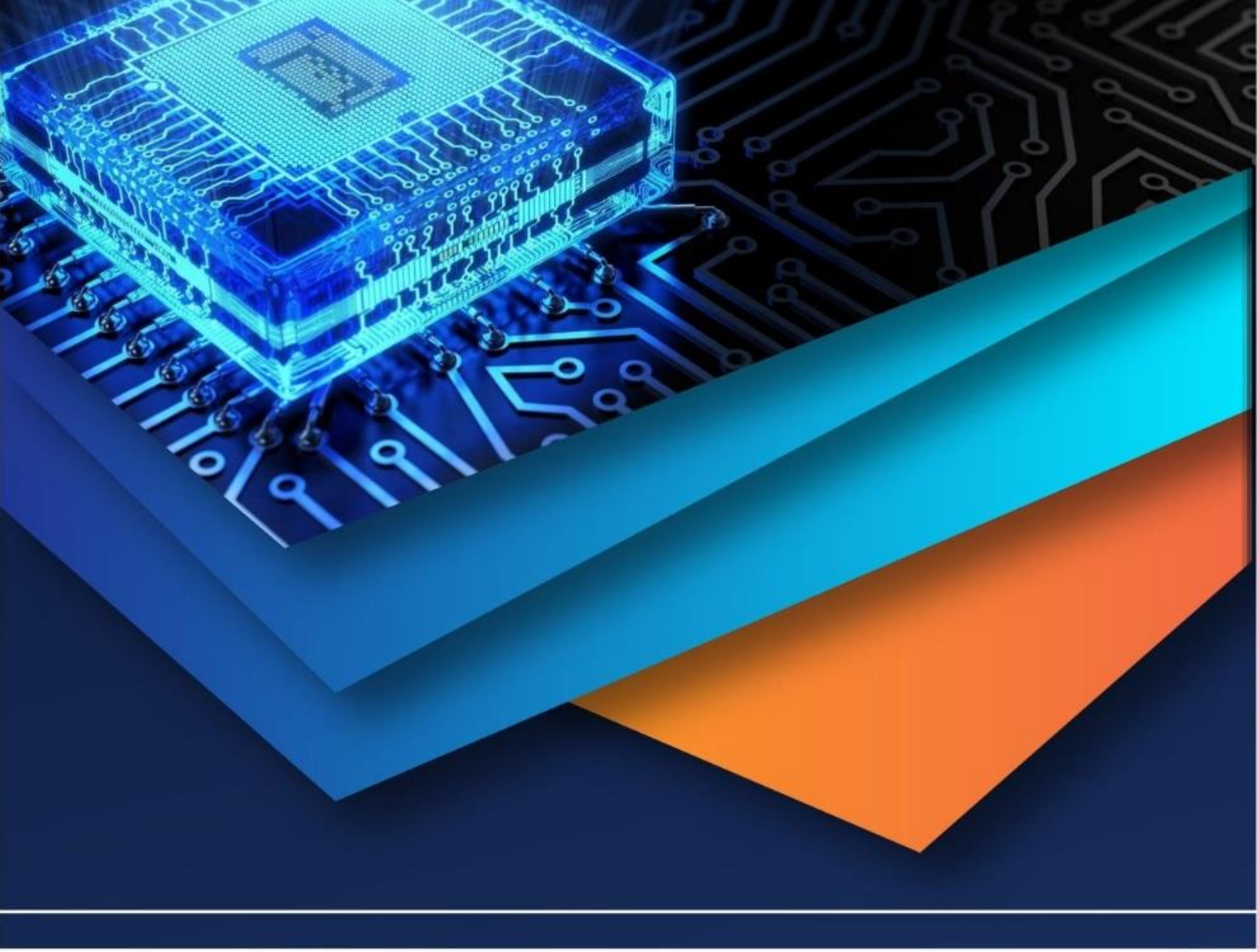


Fig. 14. Result-2

REFERENCES

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, ”MesoNet: a Compact Facial Video Forgery Detection Network,” in ACM Multimedia, 2018.
[2] C. Feichtenhofer, A. Pinz, and R. P. Wildes, ”Spatiotem- poral Residual Networks for Video Action Recognition,” in NIPS, 2016.
[3] A. G. Howard et al., ”Fastai: A Layered API for Deep Learning,” arXiv preprint arXiv:2002.04688, 2020.
[4] M. Sundararajan, A. Taly, and Q. Yan, ”Axiomatic Attribution for Deep Networks,” in ICML, 2017.
[5] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, ”Deep Anomaly Detection and Localization for Uncon- strained Face Verification,” in CVPR, 2018.
[6] Y. Li et al., ”VIP-CNN: Visual Phrase Guided Convo- lutional Neural Network,” in CVPR, 2017.
[7] I. Misra, C. L. Zitnick, and M. Hebert, ”Shuffle and Learn: Unsupervised Learning using Temporal Order Verification,” in ECCV, 2016.
[8] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, ”Face Alignment Across Large Poses: A 3D Solution,” in CVPR, 2016.
[9] S. M. Lundberg and S.-I. Lee, ”A Unified Approach to Interpreting Model Predictions,” in NIPS, 2017.
[10] Y. Qian, Y. Dong, and Y. Yang, ”DeepFake Video Detection Using Recurrent Neural Networks,” in Journal of Visual Communication and Image Representation, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY