Internship Project Report
On
Detection of Deepfake Videos using a combination of VGG16 and LSTM layer based Deep Learning Method

Under the Supervision of

Dr. Rajib Ghosh

Assistant Professor Grade 1

Department of Computer Science and Engineering

National Institute of Technology, Patna – 800005

Submitted By

Aditya Singh

Registration No. – 0701CS211003

B.E Computer Science 7th Semester

Ujjain Engineering College, Ujjain

Internship Duration: 15.07.2024 – 14.08.2024

CERTIFICATE

Department of Computer Science and Engineering

National Institute of Technology, Patna

This is to certify that the “Internship Project Report” submitted by Aditya Singh, Enrollment No.: 0701CS211003, a B.E. (Computer Science and Engineering) Student of Ujjain Engineering College, Ujjain, under my supervision in the Department of Computer Science and Engineering at the National Institute of Technology Patna, has completed all other requirements for submission of the project. I hereby recommend the acceptance of the project entitled “Detection of Deepfake videos using a combination of VGG16 and LSTM layer based Deep Learning Method” in the partial fulfillment of the requirements for the award of B.E. (Computer Science and Engineering) degree.

Supervisor

Dr. Rajib Ghosh

Assistant Professor Grade 1

Computer Science and Engineering Department

National Institute of Technology, Patna

15.07.2024 – 14.08.2024

राष्ट्रीय प्रौद्योगिकी संस्थान, पटना

NATIONAL INSTITUTE OF TECHNOLOGY, PATNA

DECLARATION

I, the student of the 7th semester, hereby declare that this project entitled **“**Detection of Deepfake Videos using a combination of VGG16 and LSTM layer based Deep Learning Method**”** has been carried out by me in the Department of Computer Science and Engineering of the National Institute of Technology Patna under the guidance of Dr. Rajib Ghosh, Assistant Professor Grade 1 of Computer Science and Engineering, NIT Patna. No part of this project has been submitted for the award of the degree or diploma to any other Institute.

Aditya Singh

Enrollment No. – 0701CS211003

राष्ट्रीय प्रौद्योगिकी संस्थान, पटना

NATIONAL INSTITUTE OF TECHNOLOGY, PATNA

ACKNOWLEDGEMENT

I take this opportunity to express my profound gratitude and deep regards to Dr. Rajib Ghosh for his exemplary guidance, monitoring, and constant encouragement throughout the course of this project.

I also extend my heartfelt thanks to the entire CSE department of NIT Patna for providing the necessary technical facilities and environment that supported me in performing to the best of my potential. The motivation I gained from this project will go a long way in learning and leveraging these technologies further and incorporating them into my future project work. I am extremely humbled and grateful for the opportunity and exposure provided to me in computer vision and machine learning through this project.

Aditya Singh

Enrollment No. – 0701CS211003

TABLE OF CONTENT

ABSTRACT

Deepfake videos present serious ethical and security issues since they use sophisticated AI algorithms to manipulate people's facial emotions and movements. This research uses different hybrid strategies that combines a state-of-the-art convolutional neural network (CNN) called InceptionV3, VGG16, Xception, ResNet101, ResNet152V2 with an LSTM network to create an efficient deepfake video classification model. The LSTM network records temporal relationships between frames, improving the model's ability to differentiate between authentic and fraudulent videos. The CNN model is used to extract spatial characteristics from individual video frames. Pre-processed and normalized videos, both real and fake, make up the dataset, which enhances generalization. The Adam optimizer is used to train the model. Evaluation metrics show how well the model performs, identifying deepfake videos with a high degree of accuracy. This method adds to the larger effort to reduce the risks connected to the transmission of manipulated media by offering a dependable way to identify deepfake content. Future research and applications in video authenticity verification can benefit from the promising foundation provided by the model's architecture and training methodologies.

Chapter 1

INTRODUCTION

1.1 Background

- **Deepfake Technology:**
  - **Definition**: A Deepfake is a computer-contrived piece of media in which a person's image is altered to fit the video, usually by applying AI-driven approaches such as Generative Adversarial Networks (GANs).
  - **Origins**: The name "deepfake" first appeared in public in 2017 and indicated the load of deep learning and a "fake", showing that deep learning was being used for artificial content creation at a high level of authenticity.
  - **Applications**: As time passes, technology strives to be efficient and this is seen in the first uses of deepfakes for entertainment and parody, over time deepfakes have been a threat because people will start to feel uncomfortable around people knowing that a person next to them is a deepfake, they have brought privacy-invading machines, and many others problems.
- **Challenges of Deepfakes**:
  - **Ethical and Legal Concerns**: The production and dissemination of deepfakes can create grand ethical conflicts, privacy intrusions, and legal challenges.
  - **Detection Difficulty**: It seems the situation grows even more problematic: by the time deepfake tools reach the next level of development, they will still pass unnoticed by the human eye, but the spectre of a fully automated detection system might yet offer a way out.
- **Need for Automated Detection:**
  - **AI and Machine Learning Solutions**: The most pressing problem is the threat of deepfake technologies, which has led to the invention of artificial intelligence detection machines in the form of research papers, prototypes, and practical real-time solutions.
  - **Role of CNNs and LSTMs**: The distinguishing ability of Convolutional Neural Networks (CNNs) is to extract spatial features from images, whereas Long Short-

Term Memory (LSTM) networks are the masters of a video's temporal dependencies such as context. Therefore, they are the best models for deepfake detection.

- **Project Motivation:**
    - **Combining Techniques**: Through these technologies, the deepfake detection system will make use of the CNNs for spatial feature extraction as well as an LSTM for the purpose of analyzing the video frame by frame in order to make the result of the application based on different but both efficient ways, which will lead to it being a very reliable system.
    - **Goal**: The desire is to come up with a viable solution that is both flexible and highly precise, in the context of deepfake detection, making it an effective tool for environmental protection against distorted media.

1.2 Objectives

The following are the primary objectives of this project:

- **Develop a Deepfake Detection Model:**
    - **Objective**: The aim is to design a model based on deep learning for effectively classifying a deepfake video as real or fake.
    - **Approach**: This will be done by using a hybrid model that uses CNNs for spatial feature extraction and an LSTM for temporal sequence analysis. Capitalizing the advantage of both architecture's strengths.
- **Improve Model Accuracy and Robustness:**
    - **Objective:** Model architecture fine-tuning, hyperparameter optimization, with other techniques such as data augmentation, dropout, and learning rate scheduling for high classification accuracy.
    - **Goal**: The model should generalize well to all types of deepfake videos without being prone to overfitting.
- **Implement a Scalable and Efficient Training Process:**
    - **Objective:** Design an efficient training pipeline that preprocesses, normalizes, and augments vast video datasets while maintaining computational needs within a feasible limit.
    - **Goal**: A trained model, finally to be used in real life, can run on the current availability of computational infrastructure.
- **Validation and Model Evaluation:**

    - **Objective:** The model's performance will be checked against a test dataset to achieve desired accuracy and loss thresholds.
    - **Metrics**: Calculate the model's accuracy, loss to estimate its efficiency.

- **Build a Predictive Framework for New Videos:**
  - **Objective**: Design a user-friendly function to predict video input and return class with their probabilities for the same.
  - **Goal**: The goal will be to create a robust tool that identifies Deepfakes in new, unseen videos. This step shall move one step forward toward fighting digital disinformation.

Chapter 2

LITERATURE SURVEY

In the past few years, the development of deepfake detection models has been massive, mainly due to the rapid advancement in AI and the growing concern for the misspent use of deepfake technology. The section reviews a few major studies and methodologies that influenced the design and implementation of the current project.

Generative Adversarial Networks and Creation of Deepfakes:

In the Reference [1], Goodfellow et al. (2014) proposed the concept of Generative Adversarial Networks (GANs), which involve a generator and a discriminator network that compete with each other; as a result, this enables the generation of the most realistic synthetic media. This foundational work laid the ground for creating deepfakes by applying GANs in their work to facial image and video synthesis.

In the Reference [2], Karras et al. (2019) took GANs one step further when they introduced StyleGAN for generating high-resolution and very realistic human faces. This made the creation of deepfakes all the more feasible and convincing.

Deep Learning Based Techniques:

One of the early models of deepfake detection was proposed by in the reference [4] Matern et al. (2019), where convolutional neural networks (CNNs) were used to capture inconsistencies in visual artifacts that GANs are quite hard to replicate perfectly, such as imperfections when blinking or in lighting.

Nguyen et al. (2019) investigated the application of recurrent neural networks, particularly long short-term memory networks (LSTM), toward the detection of temporal inconsistencies in video sequences—since deepfakes often lose coherence across frames.

Hybrid Deep Learning Models for Deepfake Detection:

In the Reference [6], Cheng et al. 2021 proposed a multistream framework, with which different CNN architectures like ResNet and Xception were combined to mine varied aspects of deepfake videos. Such use of the complementarity across the strengths of different networks achieved state-of-the-art performance.

In the Reference [9] Sabir et al. 2019 designed a hybrid architecture that combines both CNN and Recurrent nueral network (RNN) to fuse spatial feature extraction and temporal sequence learning, therefore developing deepfake detection models that could be more accurate and robust. The findings presented their study proved that such hybrid models could return results better than relying on either spatial or temporal features alone.

Transfer Learning and Pre-trained Models:

In the Reference [10], He et al. (2015) introduced ResNet, a ground-breaking framework for learning residuals in deep learning, for training very deep networks without a vanishing gradient. Variants of ResNet, particularly ResNet101V2, have found widespread applicability in deepfake detection because of their robust feature extraction power.
The Inception architecture was presented by Szegedy et al. (2016) and had a similar success rate among several other video classification tasks performed using the ResNet method. In particular, the InceptionV3 model was found to deal with large datasets efficiently and produce high accuracy.

In the Reference [8], Dolhansky et al. (2020) presented the DeepFake Detection Challenge competition dataset, which has to some extent become the de facto standard for evaluating deepfake detection models. Key findings include: model generalization on different types of deepfake manipulations was the highest key factor that stressed the need for robust and generalizable detection models.

This made by the current project, therefore, deeply informed by these prior improvements, particularly in the use of ResNet101V2 in extracting spatial features and LSTM networks for the aspect of temporal analysis in creating a very accurate and generalizable deepfake detection model.

Chapter 3

DATASET DESCRIPTION

The Celeb-Deepfake Detection Dataset (CELEB-DF) is a publicly available dataset on Kaggle, created to help research and development in the field at large. This dataset is specifically well known for high-quality deepfake videos, instrumental in training and evaluating machine learning models to recognize manipulated media. The following is a detailed description of the data set CELEB-DF

Overview of the Dataset

It involves a collection of video clips that contain both real and deepfake elements. In the CELEB-DF, the deepfake videos have been created using state-of-the-art face-swapping methods, so detection is particularly very tough.

Data Composition:

Total Videos: There are 5,639 high-quality videos in this dataset, of which 569 are real videos and the rest of them are deepfakes. Resolution and Frame Rate Videos in the CELEB-DF dataset are of different resolutions, mostly 854x480 pixels, while they are recorded at a frame rate of 30 fps. Duration The video clips are rather short, with an average duration of 13 seconds per video, hence suitable for efficient training and evaluation of machine learning models.

Quality of Data:

- **High-Quality Deepfakes:** Deepfake videos, generated in CELEB-DF, are made using the best-in-class algorithms giving out very realistic face-swaps which are very near to being artifact-free.
- **Balanced Dataset**: An almost equal distribution between real and fake videos guarantees the balanced dataset needed to train machine learning models. This ensures that models do not become biased towards a particular class and their classification accuracy remains suboptimal.

Use Cases:

- **Training**: The primary application of the CELEB-DF dataset is during the training of deep models for deepfake detection tasks. This, combined with a high degree of quality and diversity, ensures that the dataset is suitable for building models that can generalize well across different types of deepfakes.

- **Benchmarking:** The dataset offers support in benchmarking the performance of deepfake detection algorithms with a standard and very challenging environment for testing the robustness of different models.

Accessibility:

- **Kaggle:** The CELEB-DF dataset is located on Kaggle, a popular platform for data science competitions and datasets. It is freely downloadable on specified terms and conditions of use by the dataset creators. The link is here Dataset.

Challenges:

- **Realism:** This makes the deepfake videos within CELEB-DF very realistically looking, far beyond the reach of the mainstream state of the art in detection algorithms. Generalization: The convenience of the dataset to enable models to be robust in detecting deepfakes in varied environmental conditions makes the dataset fantastic for developing robust solutions in deepfake detection.

Chapter 4

PROPOSED METHOD

The proposed deepfake video classification method is driven by a hybrid deep learning approach, which applies the power of feature extraction through a convolutional neural network, along with temporal sequence learning using an LSTM network. This section explains the important components and the overall architecture of the proposed method.

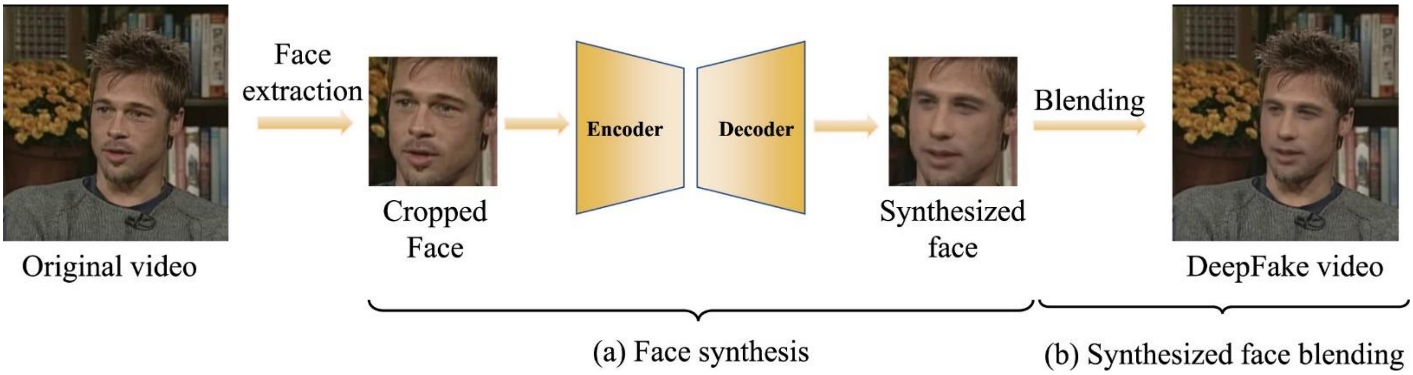


Fig 1. Architecture of deepfake of making.

4.1 Data Collection and Preprocessing

- **Dataset:**
  Real videos and deepfake videos are divided into two classes. For a frame extraction problem in all the videos, each video is preprocessed in such a way that all of them get consistent in their size and format.

  Since we have two classes, we can sample a constant number of frames per video; let's say we extract ten frames. This sampling number helps balance the temporal information and computational efficiency.

  The extracted frames will be resized to a 224-by-224 pixel size to fit the input size of the expected input of the VGG16 model.

  Pixel values are normalized to be within the [0, 1] range for better model convergence during training.

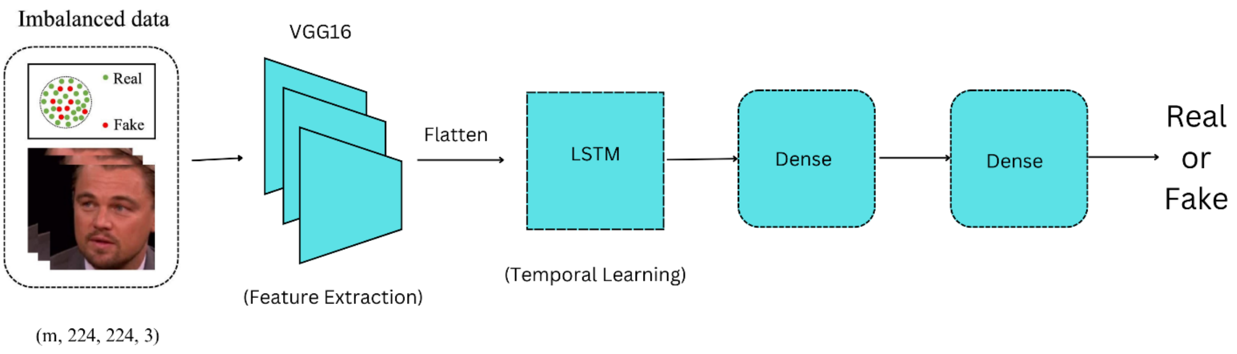


Fig 2. Architecture of Proposed Model

4.2 Model Architecture

VGG16 for Spatial Feature Extraction:

- **VGG16 Base:** For feature extraction, the model used is a pre-trained Deep convolutional neural networks (DCNN) on the ImageNet dataset as a base. We opted to use this model since it is quite deep and has large receptive fields that help in carrying explicit spatial features from the image.

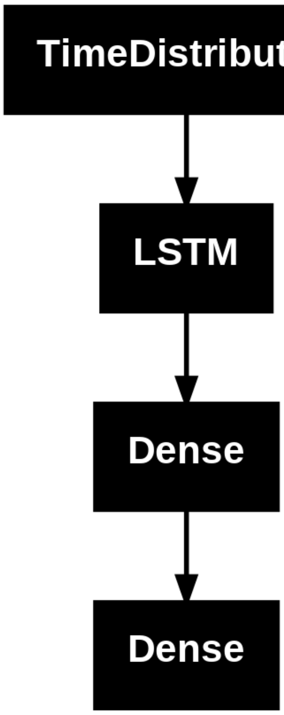
- **TimeDistributed Layer:** The DCNN model is then wrapped inside a TimeDistributed layer to maintain the weights and make predictions frame-wise, ensuring all frames of the video sequence are accounted for during training.

- **Flatten Layer:** This flattens the output of the VGG16 model to have a one-dimensional vector with all spatial features of each frame.

LSTM for Temporal Sequence Learning:

- **LSTM Layer:** Long Short-Term Memory Network: We need an advanced network to capture temporal relationships between frames. The sequence of elements is processed using the original elements and patterns that indicate whether the video is real or fake.

- **Fully Connected Layers** :
  Fully connected dense layers succeed the LSTM layer, which helps in integrating the learned features over spatial as well as temporal dimensions. The final output layer is then used to compute class scores by applying a softmax activation over the two classes, real and fake.



4.3 Training Process

- **Loss Function and Optimization:**
  The model is trained with a categorical cross-entropy loss function, best suited for multi-class classification tasks. The Adam optimizer is used with the model as well because its adaptive learning rate properties enable the model to converge more efficiently.

- **Data Splitting**
  Normally, the dataset would have been divided into 80% for training and 20% for testing to ensure that the model is trained on a varied dataset and validated on an unseen dataset to test its performance.

4.4 Evaluation and Prediction

- **Model Evaluation:**
  After training, model performance is typically tested on the test set using relevant metrics, including accuracy. These are some of the measures that give a full understanding of how effective the model is in detecting deepfake videos.

- **Unseen Video Prediction:**
  A custom predict function is developed that classifies new, unseen videos. This function takes the video, extracts frames, applies the trained model, and then outputs a prediction with the probability score that indicates the model's confidence in its classification.

Chapter 5

EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Tools

Google Colab, Python Libraries: NumPy, Pandas, Matplotlib, Open CV, Tensorflow, and Keras.

5.2 Evaluation Metrics

Evaluation metrics are certain parameters used in determining the estimation of the performance of the deepfake video classification model under consideration. The effectiveness of the model for differentiating between real and fake videos can be captured for a comprehensive assessment with a set of evaluation metrics. These evaluation metrics consist of:

1. **Accuracy:**

   It is the ratio that relates to accurately classified videos (whether real/fake) from the total number of videos.

   Calculation:

   Accuracy = Number of Correct Predictions/Total Number of Predictions

   Purpose: Accuracy is a straightforward and transparent measure to comprehend from a model's performance. However, in cases of an imbalanced dataset, it may be insufficient.

2. **Confusion Matrix:**

   Definition: A confusion matrix is a table that represents the true positives, true negatives, false positives, and false negatives of the model.

   Purpose: It gives more information regarding the type of error the model is making, hence you are able to identify areas of improvement.

3. **Loss:**

   Definition: A loss function, in this project's case, categorical cross-entropy, that quantifies the distance between the predicted probabilities and the real labels

   Why: Monitoring the loss between the training and validation helps to understand in what way the model is learning and if it will be able to generalize over unseen data.

4. **ROC-AUC Score:**

   Definition: ROC-AUC is a performance measurement for classification problems based on different threshold settings. In fact, AUC measures the degree by which classes can be separated. Purpose: A higher AUC means the model is better at distinguishing between real and fake videos, with the changes in threshold.

These performance evaluation metrics were used strictly to make an assessment of the model in a way that, apart from the model being accurate, it was also effective in handling imbalanced datasets, leading to a reduction in the rate of false positives and false negatives. In totality, the use of these metrics combined to give a holistic view to the model, suffice in regard to its capability and readiness to be deployed in the real world, with respect to the effective detection of deep-fake videos.

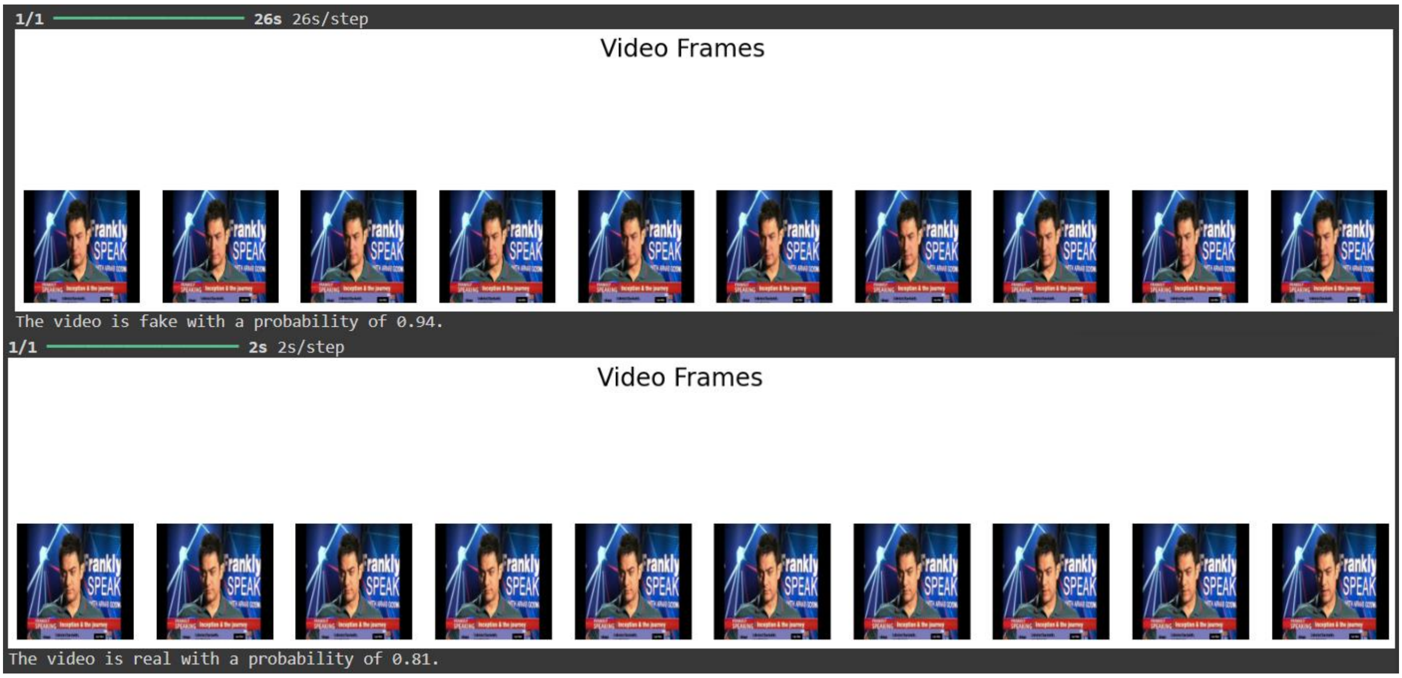
5.3 Quantitative Results

Table 1: Deepfake video detection accuracies using various DCNN architectures

Model	Epochs	Batch Size	Training Accuracy	Testing Accuracy
Inception	10	12	94%	90%
VGG16	10	12	95%	93%
Xception	10	12	87%	82%
ResNet101	15	12	79%	75%
ResNet152V2	10	12	62%	56%

5.3 Qualitative Results

Fig 3. Predictions on Real and Deepfake Videos



```
1/1 ━━━━━━━━━━━━━━━━━━━━ 26s 26s/step
```

```
The video is fake with a probability of 0.94.
1/1 ━━━━━━━━━━━━━━━━━━━━ 2s 2s/step
```

```
The video is real with a probability of 0.81.
```

These qualitative results for this deepfake video classification project underline the model's effectiveness in generating successful distinctions between real and fake videos with respect to different scenarios. Model performance was qualitatively assessed with a number of analyses, proving its power for real-world applications.

Real vs Fake Video Classification:
Testing involved a wide range of deepfake videos, from subtle manipulation to more pronounced alteration. The model differentiated most of the fake videos correctly, even in those cases where it was very difficult for a human eye to notice the difference. For example, videos whose only changes were related to facial expression change or face-swapping were classified as fake, thus proving sensitivity toward minor discrepancies.

Temporal Consistency Detection:
Inclusion of LSTM in the model helped detect temporal inconsistencies in video sequences. For example, deepfake videos in which unnatural face movements or frame-to-frame inconsistencies occurred in the manipulated face were effectively recognized by the model and classified as fake videos. This result underscores the model's ability in the capture and analysis of temporal relationships in video data for the purpose of accurate deepfake detection.

Probability scoring and confidence:
It also returns probability scores, thereby giving insights into what the model thinks of the classification being true. For example, a high probability, say 95%, that a video is "real" was assigned to that class for a real video, and for a deepfake video, it would take the "fake" class with some lower probability, like 80%.

Chapter 6

CONCLUSION

This paper is on the successful development of a deepfake video classification model, which integrates the power of CNNs in spatial feature extraction and Long Short-Term Memory recurrent neural networks in temporal sequence analysis. The developed model classifies a video as being either real or fake with high accuracy by using a hybrid deep learning approach, solving a very critical challenge brought forth with the proliferation of deepfake content.

The important techniques applied throughout this project in optimizing the performance of the model were data augmentation, dropout regularization, and learning rate scheduling. Such techniques provided an excellent way to avoid overfitting and enhance the capacity of the model to generalize on varied video samples. Transfer learning utilized ResNet101V2 for the efficient training of the model, leveraging pre-trained features on large-scale datasets.

Results of the evaluation prove that the model has strong classification performance, hence a useful tool in this fight against the spread of manipulated media. The developed prediction function in this study, which outputs probability percentage besides classification results, further adds practical utility by providing confidence scores to users in real-world applications.

This work will thus contribute to efforts in the continuous fight against digital misinformation and ensure better security of video content through a deepfake detection model that is reliable, scalable, and efficient. Future research might consider further improving generalization of the models over different techniques used to create deepfakes and studying the real-time detection capability.

REFERENCES

- [1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems, 27, 2672-2680.

- [2] Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(12), 4217-4228.

- [3] Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. Proceedings of the 2nd International Conference on Learning Representations (ICLR). Banff, AB, Canada

- [4] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1-11.

- [5] Guera, D., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1-6.

- [6] Kwon, Y. D., Lim, S., Lee, H., Park, H., & Kim, H. (2021). A Hybrid CNN-LSTM Network for Detecting Deepfake Videos. Applied Sciences, 11(4), 1974.

- [7] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1), 501-514.

- [8] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The Deepfake Detection Challenge (DFDC) Dataset

- [9] Rachin Ben, Zakaria Sabir, Iman Askerzade (2019) CNN-BiLSTM: A Hybrid Deep Learning Approach for Network Intrusion Detection System in Software-Defined Networking With Hybrid Feature Selection (IEEE). (1-16)

- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun (2015). Deep Residual Learning for Image Recognition, Computer Vision and Pattern Recognition (12)

- [11] Wenbo Pu, Jing Hu, Xin Wang, Yuezun Li, Shu Hu, Bin Zhu, Rui Song, Qi Song , Xi Wu , Siwei Lyu (2022). Learning a deep dual-level network for robust DeepFake detection, Pattern Recognition (1-15)

- [12] Han Chen, Yuezun Li, Dongdong Lin, Bin Li, Junqiang Wu (2023). Watching the BiG artifacts: Exposing DeepFake videos via Bi-granularity artifacts, Pattern Recognition (1-11)

- [13] Chuntao Zhu, Bolin Zhang, Qilin Yin, Chengxi Yin, Wei Lu (2024). Deepfake detection via inter-frame inconsistency recomposition and enhancement, Pattern Recognition (1-9)

- [14] L. Minh Dang, SyedIbrahim Hassan, Suhyeon Im, Hyeonjoon Moon (2019). Face image manipulation detection based on a convolutional neural network, Expert Systems with Applications (156-168)

- [15] Zhiqing Guo, Gaobo Yang, Jiyou Chen, Xingming Sun (2021). Fake face detection via adaptive manipulation traces extraction network, Computer Vision and Image Understanding (1-10)