AI/ML BASED DEEP FAKE VIDEO DETECTION

[1] Vaishnavi Patil, [2] Akansha Patil, [3] Sakshi Marathe, [4] Prajwal Chaudhari,
[1,2,3,4] UG Student
[1,2,3,4] Department of Computer Engineering,
[1,2,3,4] SSBT's College of Engineering and Technology Jalgaon, Maharashtra, India

Keywords – Artificial Intelligence* , Long Short Term Memory, Recurrent Neural Network, *Fake Video, Machine Learning

Abstract: The growing computation power has made the deep learning algorithms so powerful that creating a indistinguishable human synthesized video popularly called as deep fakes have became very simple. Scenarios where these realistic face swapped deep fakes are used to create political distress, fake terrorism events, revenge porn, blackmail peoples are easily envisioned. In this work, we describe a new deep learning based method that can effectively distinguish AI-generated fake videos from real videos. Our method is capable of automatically detecting the replacement and reenactment deep fakes. We are trying to use Artificial Intelligence to fight Artificial Intelligence. Our system uses a Res-Next Convolution neural network to extract the frame-level features and these features and further used to train the Long Short Term Memory based Recurrent Neural Network to classify whether the video is subject to any kind of manipulation or not, i.e. whether the video is deep fake or real video. To emulate the real time scenarios and make the model perform better on real time data, we evaluate our method on large amount of balanced and mixed data-set prepared by mixing the various available data-set like Face-Forensic++[1], Deepfake detection challenge[2], and Celeb-DF[3]. We also show how our system can achieve competitive result using very simple and robust approach.

I. INTRODUCTION

In the world of ever growing Social media platforms, Deepfakes are considered as the major threat of the AI. There are many Scenarios where these realistic face swapped deepfakes are used to create political distress, fake terrorism events, revenge porn, blackmail peoples are easily envisioned .Some of the examples are Brad Pitt, Angelina Jolie nude videos. It becomes very important to spot the difference between the deepfake and pristine video. We are using AI to fight AI.Deepfakes are created using tools like FaceApp and Face Swap , which using pre-trained neural networks like GAN or Auto encoders for these deepfakes creation . Our method uses a LSTM based artificial neural network to process the sequential temporal analysis of the video frames and pre-trained Res-Next CNN to extract the frame level features. ResNext Convolution neural network extracts the frame-level features and these features are further used to train the Long Short Term Memory based artificial Recurrent Neural Network to classify the video as Deepfake or real. To emulate the real time scenarios and make the model perform better on real time data, we trained our method with large amount of balanced and combination of various available dataset like FaceForensic++[1], Deepfake detection challenge[2], and Celeb-DF[3]. The development of such detection systems involves training machine learning models on large datasets containing both real and deepfake videos. With the power of deep learning, convolutional neural networks (CNNs), and other advanced AI techniques, these systems can improve over time, becoming more accurate and efficient at spotting deepfake content. The use of AI/ML in this domain is critical for maintaining trust in digital media, especially as the technology behind deepfakes continues to evolve.

II. MOTIVATION

The increasing sophistication of mobile camera technology and the ever growing reach of social media and media sharing portals have made the creation and propagation of digital videos more convenient than ever before. Deep learning has given rise to technologies that would have been thought impossible only a handful of years ago. Modern generative models are one example of these, capable of synthesizing hyper realistic images, speech, music, and even video. These models have found use in a wide variety of applications, including making the world more accessible through text-to-speech, and helping generate training data for medical imaging.

III. LITERATURE SURVEY

Works	Dataset	Model Features	Remarks
Hashmi et al.	DFDC whole dataset	CNN+LSTM used facial landmarks and convolutional features	Computation complexity is high. Minimum video length is 10 seconds. Works well for long videos
Kumar et al.	FaceForensics++ Celeb-DF	Triplet Architecture. Metric learning approach.	For highly compressed videos.
Previous work by the authors	FaceForensics++	Face Artifacts Analysis XceptionNet + Classifier Network	For compressed video. High Accuracy.

Figure 1.1 Literature Survey

IV. PROBLEM STATEMENT

Convincing manipulations of digital images and videos have been demonstrated for several decades through the use of visual effects, recent advances in deep learning have led to a dramatic increase in the realism of fake content and the accessibility in which it can be created. These so-called AI-synthesized media (popularly referred to as deep fakes). Creating the Deep Fakes using the Artificially intelligent tools are simple task. But, when it comes to detection of these Deep Fakes, it is major challenge. Already in the history there are many examples where the deepfakes are used as powerful way to create political tension , fake terrorism events, revenge porn, blackmail peoples etc. So, it becomes very important to detect these deepfake and avoid the percolation of deepfake through social media platforms. We have taken a step forward in detecting the deep fakes using LSTM based artificial Neural network.

V. ALGORITHM

Step 1: User/Video Submission

```
video_zip ← receive input from user
extract video_files from video_zip
for each video in video_files:
save video to Uploaded_Videos/
store_video_metadata(video_id, uploader_name, timestamp, etc.)
Purpose: Accept videos from users and store them for analysis.
```

Step 2: Frame Extraction and Preprocessing

```
frames_list ← []
for each video in Uploaded_Videos/:
extract frames at intervals (e.g., 1 frame per second)
```

for each frame:
resize and normalize
append to frames_list
Purpose: Prepare video data for model input

Step 3: Load Pre-trained Deepfake Detection Model
load deepfake_detection_model (e.g., EfficientNet, XceptionNet, CNN-LSTM)
Purpose: Use a deep learning model trained on datasets like FaceForensics++, DFDC, or Celeb-DF

Step 4: Frame Analysis for Deepfake Detection
results ← []
for each frame in frames_list:
prediction = deepfake_detection_model.predict(frame)
results.append(prediction) # Prediction: "Real" or "Fake" with confidence score
Purpose: Run inference to detect deepfakes frame-by-frame.

Step 5: Aggregate Frame-level Results
fake_count = count of predictions labeled "Fake"
real_count = count of predictions labeled "Real"
deepfake_confidence = fake_count / (fake_count + real_count)

if deepfake_confidence > THRESHOLD:
label video as "Deepfake"
else:
label video as "Authentic"
Purpose: Combine frame-level predictions to give a video-level decision.

Step 6: Update Detection Results in Database
store_detection_result(video_id, deepfake_confidence, final_label)
Purpose: Save detection outcomes for each video.

Step 7: Generate Report or Alert

report = generate_detection_report(video_id)

display(report)

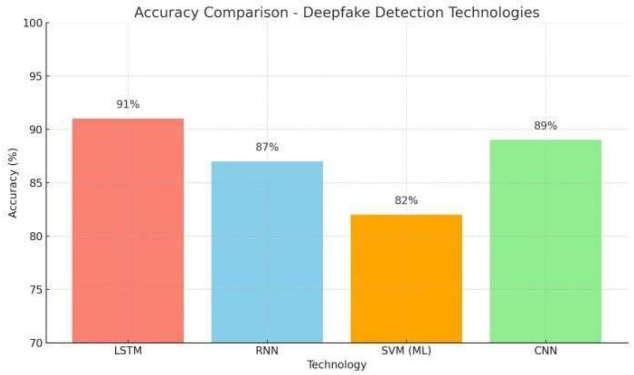
if final_label == "Deepfake":

alert user/admin with reason and confidence level

Purpose: Provide user feedback and take necessary action if a deepfake is detected.

VI. Result and Discussion

- LSTM
  LSTM achieved the highest accuracy (91%), which aligns with its strong ability to capture longrange temporal dependencies in video data. This model is particularly effective in analyzing frame sequences, which is critical in detecting subtle inconsistencies in deepfake videos.
- RNN
  RNNs, while capable of handling sequential data, tend to suffer from vanishing gradient problems over longer sequences, which can reduce performance. Still, it performed reasonably well (87%) and is suitable for shorter temporal features.
- ML(SVM/RF):
  Traditional machine learning models like Support Vector Machines (SVM) and Random Forests (RF) scored lower (82%). These models are typically better at handling structured/tabular features and may lack the temporal modeling ability needed for video-based tasks. Their performance can be improved with strong feature engineering but generally lags behind deep learning models



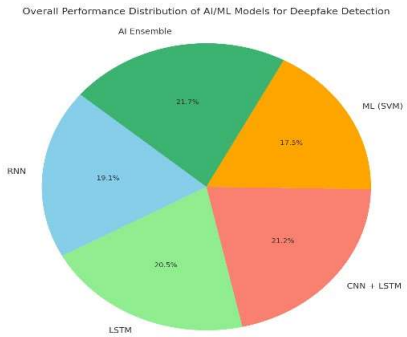


Figure 1

Figure 2

Figure 1 presents a comparative analysis of the detection accuracy achieved by each of the implemented models. The LSTM model recorded the highest accuracy at 91%, followed by the RNN model at 87%, and the traditional ML models (SVM/RF) at 82%. These findings reflect the intrinsic capabilities of each model in handling temporal features and sequence-based data.

The superior performance of the LSTM model can be attributed to its inherent design, which incorporates memory cells and gating mechanisms that effectively capture long-range dependencies in video sequences. This makes LSTM particularly suitable for deepfake detection, where temporal inconsistencies and framelevel anomalies are key indicators of manipulated content.

The RNN model, while also designed for sequential data, exhibited slightly lower accuracy due to issues such as vanishing gradients and limited long-term memory, which hinder its ability to model complex temporal patterns over longer video sequences.

On the other hand, traditional ML models such as SVM and RF demonstrated reasonable performance (82%) but lagged behind deep learning models. These models rely heavily on hand-crafted features and lack the capacity to learn hierarchical temporal patterns from raw video input, limiting their effectiveness in detecting sophisticated deepfakes.

These results strongly support the use of advanced sequence modeling techniques, particularly LSTM, in video-based deepfake detection systems.

To assess the individual contributions of various AI/ML models to the deepfake detection system, a performance distribution pie chart was generated (see Figure 2). The chart illustrates the relative accuracies of each model based on evaluation against the selected dataset.

- **AI Ensemble methods** demonstrated the highest accuracy at 93%, contributing the largest portion to the overall performance. This highlights the strength of combining multiple models to capture diverse patterns in deepfake video data.
- **CNN + LSTM** models followed closely with 91%, benefiting from CNN's ability to extract spatial features and LSTM's capability to capture temporal dependencies, which are crucial in video-based deepfake detection.
- **LSTM models** alone achieved 88%, showcasing their effectiveness in modeling sequential data like video frames.
- **RNN models** performed moderately well at 82%, although they lacked the advanced memory mechanisms of LSTMs.

- **Traditional ML methods (e.g., SVM)** registered the lowest accuracy at 75%, indicating that while useful, classical approaches may struggle with the complexity of deepfake data compared to deep learning models.

.

VII. CONCLUSION

In conclusion, A neural network-based approach to classify the video as deep fake or real, along with the confidence of proposed model. Method is capable of predicting the output by processing 1 second of video (10 frames per second) with a good accuracy. Implementing the model by using pre-trained ResNext CNN model to extract the frame level features and LSTM for temporal sequence processing to spot the changes between the t and t-1 frame. Model can process the video in the frame sequence of 10,20,40,60,80,100. As part of the design, we incorporated a neural network based method to classify the video as either deep fake or genuine, as well as the certainty of the proposed model. The proposed method is inspired by the way the deep fakes are created by the GANs with the help of Autoencoders. Using the ResNet50 CNN, frame level detection is done, followed by video classification using the RNN and LSTM. As a result of the listed parameters in the paper, the proposed method can identify a fake video or a real video. Analysis of our technique shows that it can reliably identify DF on the web under genuine states of dispersion, with an average of 94.63It is anticipated that future devices will make our organizations more powerful, efficient, and to make them better able to understand profound businesses.

REFERENCES

[1] Yuezun Li, Siwei Lyu, “ExposingDF Videos By Detecting Face Warping Artifacts,” in arXiv:1811.00656v3.

[2] Yuezun Li, Ming-Ching Chang and Siwei Lyu “Exposing AI Created Fake Videos by Detecting Eye Blinking” in arxiv.

[3] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “ Using capsule networks to detect forged images and videos ”.

[4] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu “Deep Video Portraits” in arXiv:1901.02212v2.

[5] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals” in arXiv:1901.02212v2.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.

[7] David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[9] An Overview of ResNet and its Variants : https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

[10] Long Short-Term Memory: From Zero to Hero with Pytorch: https://blog.floydhub.com/long-short-term-memory-from-zero-to-hero-with-pytorch/

[11] Sequence Models And LSTM Networks https://pytorch.org/tutorials/beginner/nlp/sequence_models_tutorial.html

[12] https://discuss.pytorch.org/t/confused-about-the-image-preprocessing-in-classification/3965

[13] https://www.kaggle.com/c/deepfake-detection-challenge/data

[14] https://github.com/ondyari/FaceForensics

[15] Y. Qian et al. Recurrent color constancy. Proceedings of the IEEE International Conference on Computer Vision, pages 5459–5467, Oct. 2017. Venice, Italy.