International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 11 | Nov 2024 www.irjet.net p-ISSN: 2395-0072

# Decoding Deepfakes: An LSTM-Driven Approach with Attention Mechanisms and Grad-CAM Explainability

**Aditya Aiya[1], Nishant Wani[1], Mayur Ramani[1]**

[1] *School of Computer Science Engineering and Applications, DY Patil International University, Pune- 411035*

---

***Abstract:-****Deepfake technology has alarmed a number of security and truth campaigners around the world because it makes possible to spread misinformation. Deep fake videos are increasingly becoming a problem and they can have serious consequences for how authentic our digital content is. In this work, we propose a deep learning architecture that fuses Bidirectional Long Short-Term Memory (Bi LSTM) networks with attention mechanisms so as to accurately determine whether videos are real or fake. In this work, we utilize the Celeb-DF (Celeb-Deepfakes) Dataset which consist of deepfake videos with high-quality to train and evaluate our model. By utilizing Grad-CAM (Gradient-weighted Class Activation Mapping), accompanied with the explanation of AI ethics, transparency and explainability all these measures we aim to provide visual explanations in order for help audience observe models' decision-making processes leading towards being more accountable. It can complete an accuracy of 91% and AUC:90% when robust havoc precision, recall, F1-scores. The results of this study showcase potential for explainable AI systems in deepfake detection through LSTM-based models with attention mechanisms and reinforce the necessity to prioritize interpretability as an attribute when designing reliable and fair AI systems.*

***Keywords:*** **DeepFakes, Celeb-DF, BiLSTM-Attention, Grad-Cam, Explainability**

## 1. INTRODUCTION

The advent of deepfake technology in recent times has blown huge worries for many fields like digital communication and also media and security. Mostly videos, deepfakes are modified content designed with deep machine learning methods to look as though they are real, even though the entire video is fake. These technologies use deep learning models, mostly Generative Adversarial Network (GANs) and autoencoders, to change or add various features such as voice, emotion, facial expression, etc., making it progressively harder for us to tell the difference between genuine and fake videos. Consequently, it is also an increasing threat that deepfakes would be adopted as malicious behaviour from monetary crimes, privacy invasion, misinformation and political manipulation. That is why there is a need for such deepfakes detection system which is extremely reliable and precise [1]

Addressing these issues has led the research community to propose numerous approaches for deepfake detection. Neural networks, including Convolution Neural Networks (CNN), Recurrent Neural Networks (RNNs), and hybrid methods. These models are based on spatial and temporal feature analysis of manipulated media in a video for dissimilarity detection. The development of deepfake detection is highly competitive and achieving both high accuracy and interpretability is overwhelming. This is particularly the case with the arms race of deepfake development that keeps advancing, thus finding it more and more complicated to trace.[2]

The study proposes a hybrid Bidirectional Short-Term Memory (LSTM) network with an Attention mechanism, which is able to replicate long-term dependencies, and also, emphasizes the most important components of an entire sequence. This research used BiLSTM-Attention model because LSTM can learn the temporal dynamics of sequences especially in the case of video data temporal changes is an important feature for manipulation detection. The architecture of this model helps in improving accuracy as well but also interpretability, which is essential for getting verification of how the detection system comes to a decision in the first place. The Attention mechanism enables the model to pay attention to video frames it deems relevant to the task, providing an interpretability of the media that is most contributing to the detection decision [3][4].

The Celeb-DF dataset, which is used in this study, includes a wide variety of deepfake and natural videos of celebrity faces. This dataset is well-known for having difficult, high-quality deepfake samples and is therefore a top choice for benchmarking detection algorithm robustness. This particular dataset was chosen since it is the one which the study intends to use in general: to test various models on advanced manipulations such as those that change facial motions and temporal consistency across long videos. FaceForensics++ [5] and DeepFake Detection Challenge [6]have been previously used to evaluate many detection algorithms, and curational benchmarks[7][8].

Deepfake detection has been key to the development of explainable AI (XAI). Transparency and interpretability are important, especially in the cases in

which false claims detection has legal or social consequences or should be justifiable based on the analytical results. For example, in judicial cases, forensic experts must be able to justify how they reached their conclusion on why something in a video was detected as tampered. We incorporate interpretability into our work through the use of Grad-Cam along with the LSTM-Attention model to visualize which frames were most relevant to our decision-making. This is consistent with the latest trends in explainable deepfake detection, where model transparency is increasingly considered as an appropriate representation of trust between AI systems and human users [4].

The paper presented is organized to thoroughly review the process and outcome of deepfake detection research. First, a Literature Review assesses previous methods demonstrating the effectiveness of other proposed models. Then an EDA and Methodology section that explains the Celeb-DF [9]dataset and preprocessing steps as well as the architectural design of the LSTM-Attention model and the optimization approach. The Results and Discussion section discusses the evaluation metrics of the model and compares them with the available state-of-the-art results, making use of visual explanations from the Attention mechanism. The work also discusses its limitations and potential future research directions e. g. multi-modal data fusion and improving adversarial robustness. It wraps up with a brief summary of findings and the importance of explainable deepfake detection, and AI, while expressing sincere appreciation for the contributions from collaborators and funding sources, and aims to express insights that are of value to AI and cybersecurity researcher and practitioners.

## 2. Literature Review

Konstantinos Tsigos, et al. (2024)[10]. "Towards Quantitative Evaluation of Explainable AI Methods for Deepfake Detection " This work aims to systematically compare many existing explanations and evaluates performance of explanation methods in identifying important regions in images that contribute to the detection. LIME is found to be the best method and, being the highest influencing predicator amongst adversarial attacks improves deepfake detection interpretability. Suk-Young Lim, et al. (2022) [11]. "Detecting Deepfake Voice Using Explainable Deep Learning Techniques". This research demonstrates detecting Deepfake Voice Using Explainable Deep Learning techniques. Identifying the use of the Explainable AI Techniques from Image Classification-digital fingerprinting of audio. This method uses attribution scores to make the model decisions in a human perception aligned way to show the transparency of the models and build trusts in audio-based deepfake detection models. Mudit Arya, et al. (2024) [12]. "A Study on Deep Fake Face Detection Techniques". This review discusses the development of deepfake face technologies and their impact on society. It discusses current approaches for detection using neural networks and machine learning, with advantages and disadvantages. Conclusion- This paper gives an importance to the interdisciplinary solutions (blockchain and explainable AI) required for robust deepfake detection.

Samuel Henrique Silva, et al. (2022) [13] . "Deepfake Forensics Analysis: An Explainable Hierarchical Ensemble of Weakly Supervised Models". This paper presents an explainable hierarchical forensics algorithm that incorporates human decision-making to assess whether or not an image is a deepfake. It uses attention based CNNs and Grad-CAM for visual explanations to yield high accuracy on the DFDC dataset, proving its robustness in detecting deepfakes and more importantly, providing clear and interpretable results. Fatima Khalid, et al. (2023) [14]. "DFGNN: An interpretable and generalized graph neural network for deepfakes detection". The authors propose an interpretable and generalizable graph neural network model for deepfake detection across different techniques and datasets. By transforming images into patches (i.e., nodes) and creating a nearest-neighbour graph, the model employs a pyramid structure to extract multi-scale features. It obtains high AUC scores which means it works effectively on identifying tampered facial images in different datasets. Pavel Korshunov, et al. (2022) [15]. "Custom Attribution Loss for Improving Generalization and Interpretability of Deepfake Detection". In this paper, the author proposes a novel training method that relies on custom Triplet and ArcFace losses to improve both the accuracy and interpretability of deepfake detection. The model also distinguishes between deepfake attacks by attribution, which helps deepen the generalization across many datasets. The Xception net architecture outperforms others in investigating properties of the deepfake space relative to original videos, according to the study.

Tao Luan, et al. (2024) [16]. In " Interpretable DeepFake Detection Based on Frequency Spatial Transformer", proposes the Find-X network to improve the interpretability of DeepFake detection via unsupervised learning. The model consists of a forgery trace generation network (FTG) and a forgery trace discrimination network (FTD), which are able to extract inconsistent forgery traces from frequency and spatial domain respectively. It is found that the method outperformed the state-of-art approaches on common Benchmarks while showing explanatory maps of the traces to us. Merel de Leeuw den Bouter et al. (2024) [17]. "ProtoExplorer: Interpretable forensic analysis of deepfake videos using prototype exploration and

refinement". The authors introduce ProtoExplorer, a Visual Analytics system for prototype-based deepfake detection. To further improve interpretability and discover prototype similarity issues, this system lets users visually analyse the prototypes and iteratively sharpen their predictions. Forensic expert evaluations confirm that ProtoExplorer stabilizes interpretation, decreases bias but keeps detection accuracy.

Majed M. Alwateer, et al. (2024) [18]. "Explainable Deep Fake Framework for Images Creation and Classification." We introduce an explainable deepfake creation and classification framework, consisting of Instant ID for image generation, Xception for classification, and Local Interpretable Model (LIME) for interpreting model predictions. It attains 100% precision on both of these tasks to demonstrate the efficiency of the proposed method on the creation of and classification of deepfakes. Md. Shohel Rana, et al. (2021) [19],"Deepfake Detection Using Machine Learning Algorithms." The critical thing this study conveys is that traditional ML based methods work very effectively compared to other deep learning (DL) based methods achieving 99% accuracy. The authors achieves high accuracy on Face Forensics++ and discuss results on a variety of datasets and conclude that ML can result in better performance at lower computational cost and higher interpretability. Fatima Khalid, et al. (2023) [20] . "DFP-Net: An explainable and trustworthy framework for detecting deepfakes using interpretable prototypes.". The authors propose DFP-Net that leverages prototype-based representation learning and Grad-CAM to provide more explainable deepfake detection. The framework achieves state of the art performances om benchmark datasets by generating representative images and producing heatmaps to indicate the important features, which also improve its trustworthiness in the view of forensic experts.

Yogesh Patel, et al. (2023)[21] . "Deepfake Generation and Detection: Case Study and Challenges" This broad review investigates the progress of GANs in generating and detecting deepfakes. They review the existing methods, cite challenges of implementation, and suggest opportunities for future research. Xiaorong Ma, et al. (2024) [22]. "Explainable Deepfake Detection with Human Prompts." To enhance interpretability, the study presents a human-assisted detection approach in which human prompts are fused with the latent representation of the model to develop an interpretable detection framework. Offering produced signs of regions probable to manipulation that decrease the gap between machine and human comprehension enabling improve in applied locations by means by achieving a high AUC score at FaceForensics++ dataset.

Samuele Pino, et al. (2021) [23], "What's wrong with this video? Comparing Explainers for Deepfake Detection.". This work establishes, analyses, and compares other explanation techniques such as SHAP, Grad-CAM, and self-attention models for interpreting deepfake detection. In a user survey, the authors evaluate the utility of various explainers, prompting the justification that we need to explain deepfake classifiers but not explainable deepfakes. Ying Xu, et al. (2022)[24] . "Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection" In this paper, the authors propose a model with the application of supervised contrastive loss, so that it enhances the generalization capability of the method to the generation type of the deepfake and some unseen attacks. From the representation space where manipulated content differs from genuine content, the model learns to tell one from the other. They analyse the explainability of the features learned and combine scores of their contrastive model with the Xception network, getting high accuracy with the SupCon model and with the fused approach whilst providing reproducible research with the corresponding code available to the public. Sunkari Venkateswarulu, et al. (2024) [25]. "Deep Explain: Enhancing DeepFake Detection Through Transparent and Explainable AI model". In this paper, the authors propose Deep-Explain with a combination of CNNs and LSTMs that have explainability features in the deepfake partially explainable AI model. Using DFDC dataset, Deep-Explain performs accurately in the key metrics such as recall and precision. With explainability methods like Grad-CAM and SHAP it can not only detect deepfakes but also explain how the decision was made, increasing the level of trust in digital media verification.

## 3. Exploratory Data Analysis

The Dataset used in this research consists of video samples divided into binary classes, such as "FAKE" (manipulated) and "REAL" (authentic), Dataset organization is done in a formal method; labels are kept in.CSV file, which provides 1-to-1 mapping between video files and their authenticity labels. This structure helps expedite the loading of the data making it easier to step to the next steps of feature extraction, modelling and evaluation procedures. To assess any presence of dataset-related bias, we performed an in-depth examination of the class distribution within the dataset, as this may affect model performance. The class distribution is plotted using a count plot in Figure 1; 66.09 % of our dataset are FAKE videos, 33.91 % are REAL videos. The distribution is highly imbalanced, which indicates that weighting mechanism is required while training the model, as FAKE class is heavily dominating the other. The dataset shows that it will require the utilization of balancing techniques to be able to train the model properly

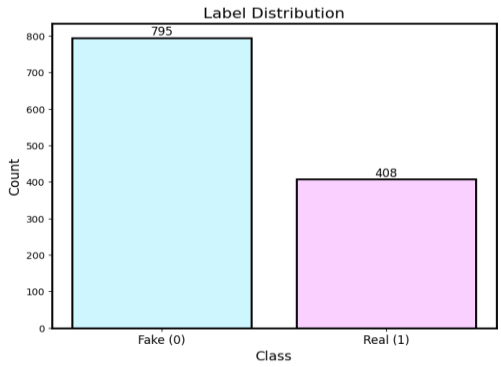


**Fig-1**. Class Distribution of Videos

In order to gain a comprehensive insight into how deepfake manipulation can be expressed visually, we conducted a detailed analysis at the frame-level. The effective differences between FAKE and REAL videos were compared at 300ms intervals by extracting sample frames at equally spaced intervals on the time axis. As shown in Figure 2, the representative frames from each of the class shows differentiating visuals. As pointed out in our study, FAKE videos often have these subtle discontinuities at the facial boundaries, signifying a sudden change of skin texture and gradients of colours. Furthermore, the lighting tends to be inconsistent if you examine the interaction between the facial structure and environment lighting in manipulated content. However, a frame-wise analysis can show unnatural transitions in facial expressions and blinks (becoming apparent in sequences where the head is moving quickly).

**Fig -2**: Sample Frames for Fake videos

**Fig- 3**: Sample Frames for real videos

However, real videos are very different. REAL content is spatially coherent regarding layout of facial features and has smooth transitions in texture and colour. Face geometry interacts with natural environmental illumination, generating consistent shadows and highlights. Intermediate frames exhibit fluid transitions of facial expressions and movements, with motion patterns that are continuous with corresponding human facial motion. The sample frames for real videos are displayed in Figure 3. This unique difference of FAKE/REAL content create significant visual signature difference for our detection system.

Based on these visual patterns we observed, we then developed a feature extraction strategy and selected EfficientNetV2B0 as the feature extractor model. This architecture has shown the ability to extract local texture patterns at face boundaries between real and fake faces, global-local lighting relations, and temporal consistency signs. The multi-level feature extraction capabilities of the network in a hierarchical manner aligns with the discriminative traits pointed out which helps achieve a strong detection of the manipulation artifacts while being computationally efficient. This multi-scale characteristic of the architecture is particularly advantageous for the algorithm to capture both subtle local inconsistencies and more global contextual features that are often present in manipulated content.

This exploratory analysis led to a few key insights that informed methodological design in the following. The insights from these findings gave us a foundation based on which we could formulate a strong detection system that utilized spatial and temporal features that were unique and distinctive for deepfakes which ended up contributing to our proposed approach as well. The basic flow of the model is represented in Figure 4 below.

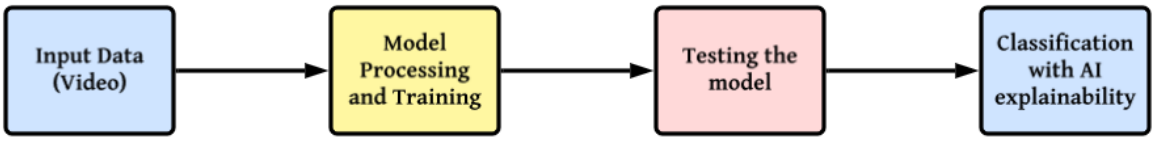


**Fig-4**: Flow of the model

## 4. Methodology

The proposed model shown in the below flow chart in Figure 5, combines an LSTM based architecture with an attention mechanism for a better classification performance to identify FAKE and REAL video frames. This architecture efficiently represents temporal dependencies across feature sequences, while the attention layer enables the model to focus on relevant features, sharpening the learning. Moreover, hyperparameter tuning was done on a large scale, with Keras Tuner's Hyperband optimizing the number of LSTM units and dropout rate, which are among the most significant aspects, to maximize performance.

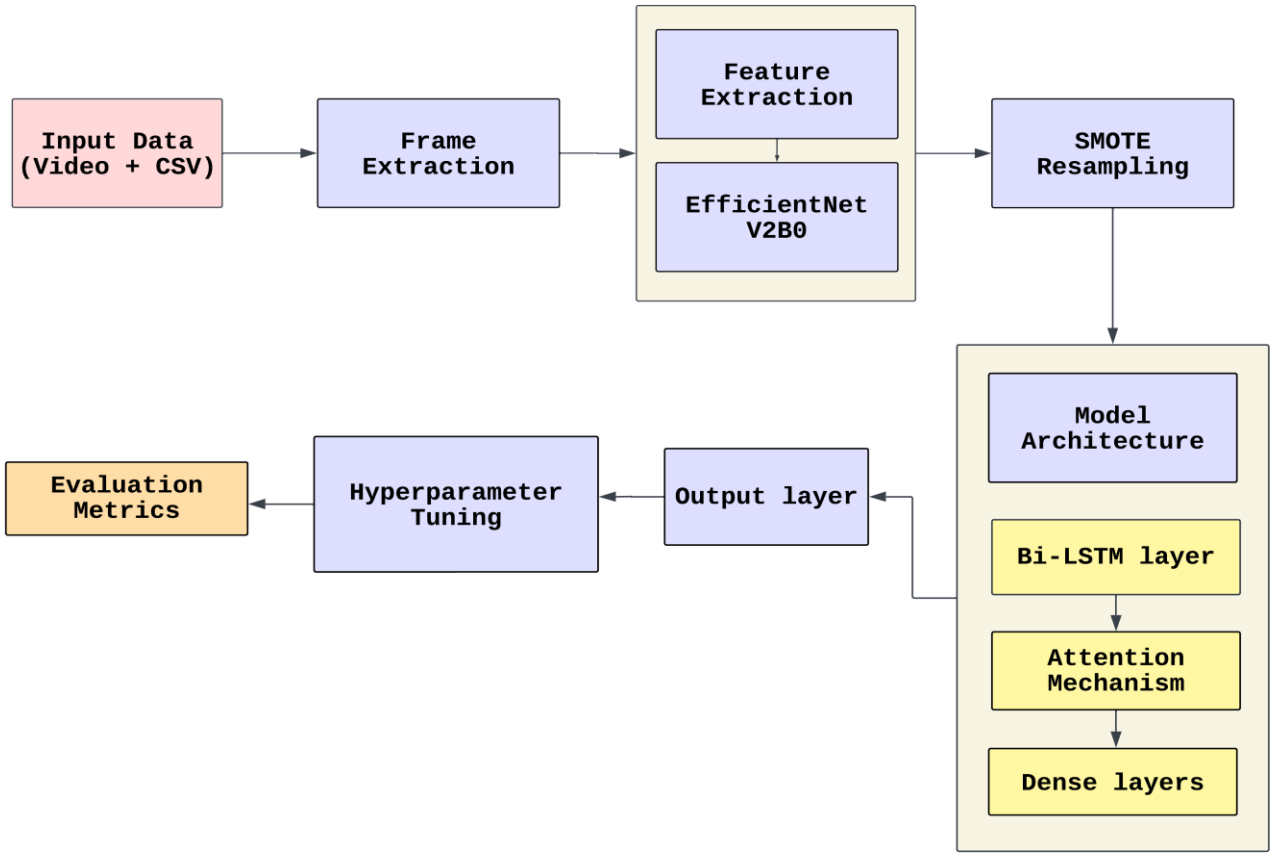


**Fig -5**: Model Flow chart

As represented in the flow chart, the model architecture has an input layer that takes a feature sequence from the frames of the video. This sequence of features acts as the primary data input for the temporal analysis. At the heart of the model is a Long Short-Term Memory (LSTM) layer, which is ideal for capturing long-range dependencies in sequential data—that is, video data, where the content of earlier frames may inform classification. An LSTM layer - LSTM layer with units (from 64 to 256) tuneable range & it has L2 regularization to solve the overfitting problem as shown in the flow chart.

The attention mechanism which is also an important part shown in the flow chart follows the outputs of the LSTM layers. By dynamically assigning weights to temporal features according to their relevance, the attention mechanism improves the model performance on focusing on certain feature(s). The selectively concentrated on relevant regions of an input sequence, resulting in a weighted representation that gives higher importance to critical parts of the input sequence (in this case, the frames of video in each input sequence) that leads to higher accuracy in classification. After the attention layer, there is a fine-tuneable dropout which is used to regularize the model from overfitting. By randomly skipping certain neurons during training, this dropout rate after being cautiously tuned through hyperparameter tuning helps to strike a balance between learning capacity and generalization.

Following the attention layer, the output is then reduced along the temporal dimension using a mean function which flattens the sequence into a tiny, summarized vector that contains only the most important features as illustrated. This aggregated representation is then forwarded to an additional fully connected dense layer containing 64 units, a batch normalization layer, and another dropout layer. This dense layer with activation called 'Relu' is followed by L2 regularization that guarantees minimal risks of overfitting before it classifies the former transition of temporal to space using dense layers along with the L2 regularization. The architecture ends in a sigmoid activated output layer for binary classification with the output score filling in a probability for the input sequence being classified as either FAKE or REAL.

Hyperparameter tuning was critical, and was performed to optimized the performance of the model. Once again, we used Keras Tuner in the Hyperband configuration to choose the best value for LSTM units and dropout rate based on the validation accuracy. Hyperband has an adaptive tuning mechanism, as illustrated in the flow chart, economizes the computational resources by executing 50 epochs/run (maximum) per trial with a factor of 3. This process was coupled with early stopping to watch out for validation loss and preventing overfitting. The winning model was fitted to the complete training set after tuning but included a ReduceLROnPlateau callback to allow a more gradual learning rate based on the validation loss.

Further, after training the model completely, emphasizing more towards transparency and explainability, the architecture as shown in Figure 6, employs an XAI (Explainable AI) model, Gradient-Weighted Class Activation Mapping **(**Grad-Cam), which helps in explaining the reason behind the decision taken by the trained model to classify the video as FAKE or REAL

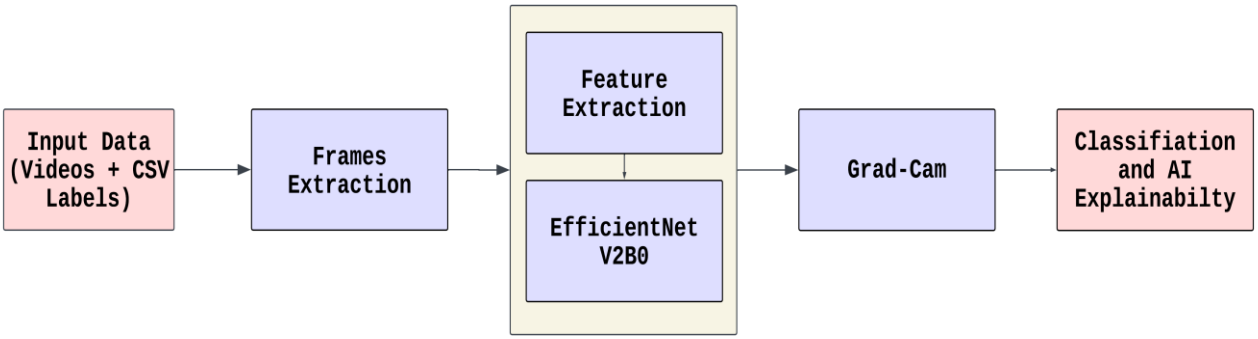


**Fig-6**: Testing model flow

## 5. Results And Discussion

Using Celeb Face dataset, the proposed method of LSTM-attention boosted deep fake detection framework provides excellent performance on distinguishing real contents from fakes. By combining the EfficientNetV2B0 for spatial feature extraction and LSTM-attention for temporal modelling, the model reaches a test set accuracy of 91.0%. Detection tasks present harsh conditions nowadays, and thus this performance is incredibly promising. Table 1 displays the complete classification report for the trained model

Table 1. Classification report

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **FAKE** | 0.87 | 0.96 | 0.92 | 159 |
| **REAL** | 0.96 | 0.86 | 0.91 | 159 |
| **Accuracy** | | | 0.91 | 318 |
| **Macro avg** | 0.92 | 0.91 | 0.91 | 318 |
| **Weighted avg** | 0.92 | 0.91 | 0.91 | 318 |

The framework demonstrates complementary, but asymmetric, performance characteristics across classification tasks. The model shows 87% precision and 96% recall when detecting FAKE video, which reflects a high sensitivity to tampered content and a decently high precision as well. On the other hand, the model achieves 96% precision and 86% recall for REAL video classification, meaning that it is very confident in its identification of real content. On the other hand, the powerful F1-scores of 0.92 (FAKE) and 0.91 (REAL) demonstrate that our model excels at both detection and verification, achieving consistent results for each category as well. The close distance between the metrics for training and for validation as seen in the learning curves (Figure 7 and Figure 8) demonstrate successful regularization and strong feature learning.

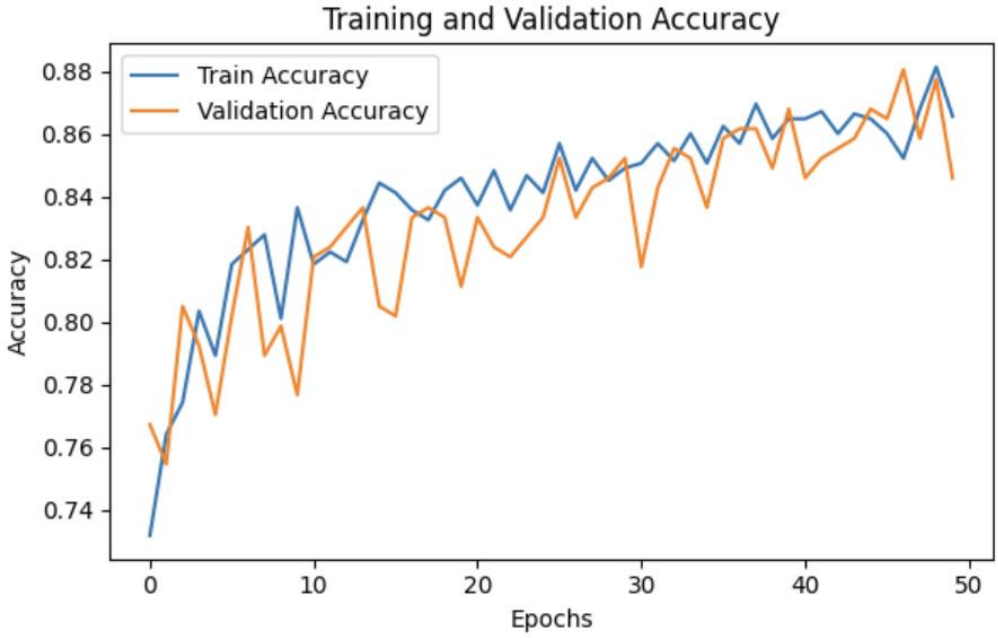


**Fig-7**: Training vs Validation Accuracy Curve

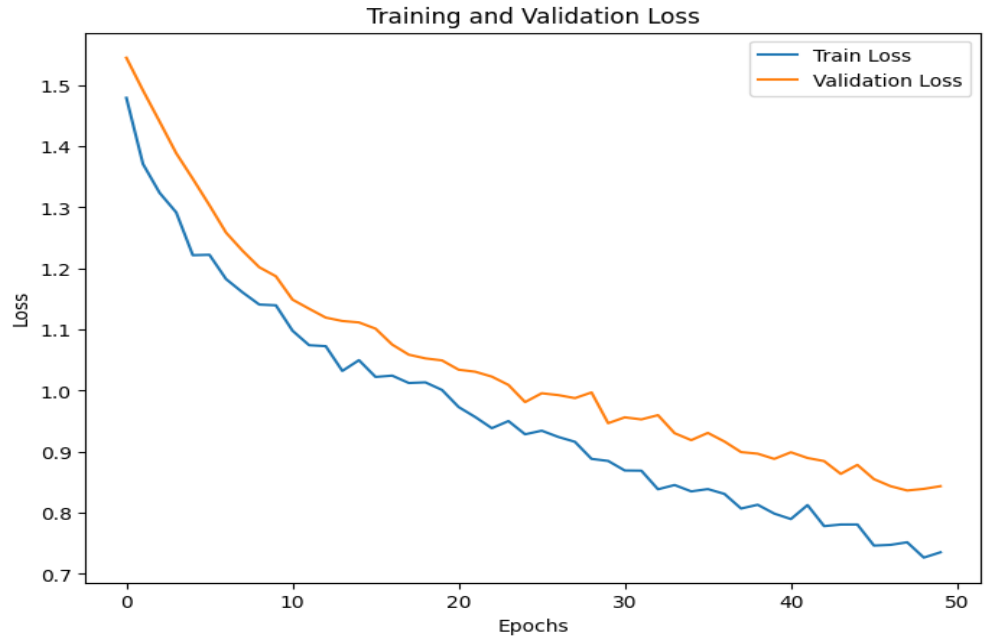


**Fig-8**: Training vs Validation Loss Curve

The high efficiency of the results are further advocated by Confusion matrix showing high percentage of correct predictions for both the classes in Figure 9 along with ROC-AUC graph achieving value of 0.9 as shown in Figure 10

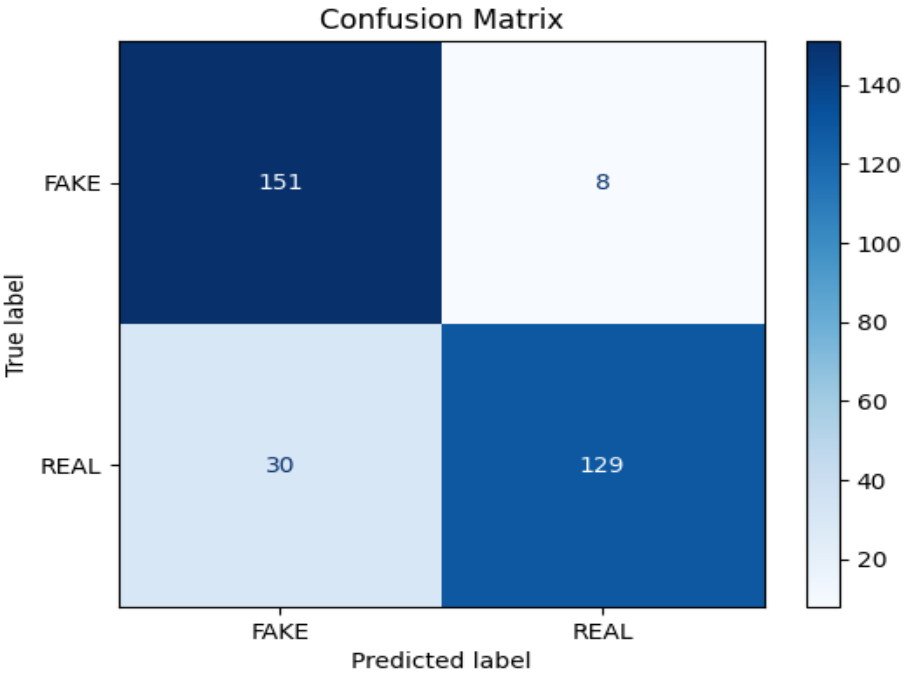


**Fig-9**: Confusion Matrix

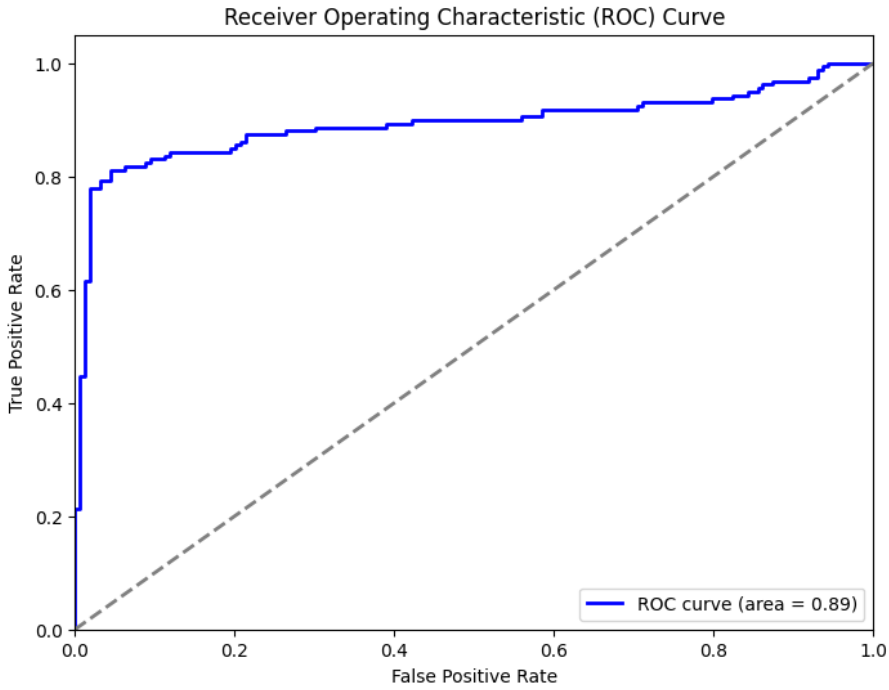


**Fig-10**: ROC-AUC graph

For FAKE, we observe that attention weights are highly active when there are visible manipulation artifacts in FAKE videos that are correctly classified, especially when the expression transitions and movements are fast. This indication of temporal sensitivity implies that specific patterns of manipulation times have been learned. For REAL videos, the more uniform attention weights distribution among frames suggests that the model has learnt coherent natural priors from real data.

Using Grad-CAM, we can further validate the model learning approach through visualization, where we show separate parts of the image being activated for separate classes. On the other hand, for FAKE videos, the model consistently attends to facial contours, spatio-temporal inconsistencies, and regions of temporal discontinuity as shown in Figure 11(a) (b) . In contrast REAL video activations have a much more even distribution over facial features, particularly during transitions between natural expressions. The activation patterns are consistent with existing knowledge of specific deepfake artifacts, providing evidence that the model has learnt relevant discriminative features.

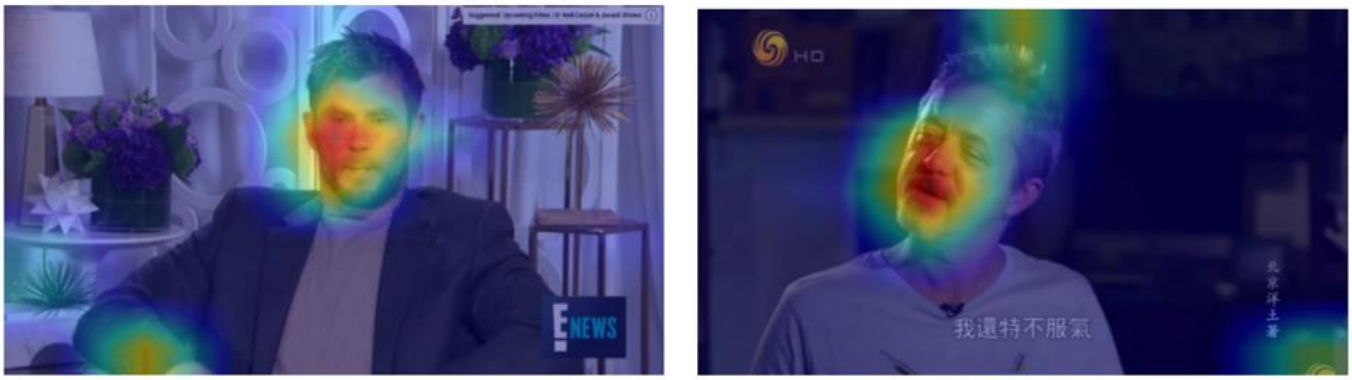


**Fig-11(a):** Grad-Cam Output

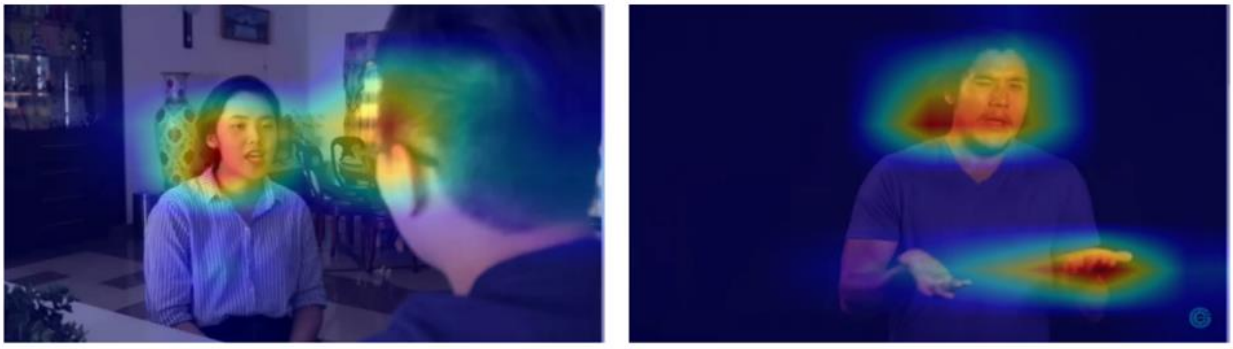

**Fig-11(b):** Grad-Cam Output

Failure cases give useful insights about the behaviour of the model that can be extracted from failure analysis. The lower precision for FAKE (0.87) indicates some slight misclassification of real videos, which might happen if the lighting conditions of the videos are quite different from the training data, or if we are using unexpected facial expressions. Nevertheless, FAKE videos present much higher recall (0.96), meaning, the model has very high sensitivity towards manipulation artifacts, and was able to identify nearly all manipulated content, but at the expense of a few false positives. This swap seems a sensible compromise for real applications where recovering manipulated content is more damaging than a false positive. The model was tested by inputting a video and Figure 12 displays the result of explainability along with text classification and probability of it being either REAL or FAKE.


**Fig-12.** Grad-cam Heat map along with text classification and probability

## 6. Limitation

Though the model architecture of this study provided good results for the classification of FAKE & REAL video frames within the Celeb Face dataset, there are some significant shortcomings. The reliance of the model on high-quality extracted features poses challenges in two aspects: First, noisy or low-quality features can severely affect the classification accuracy. In addition, the incorporation of LSTM and attention mechanisms in the architecture makes it computation intensive, as it requires a large number of resources for training and hyper parameter tuning which may make it less accessible to researchers without high-performance computing facilities.

This also present a limitation in respect to the model ability to retain past information over long video sequences. Although LSTM layers are designed with short to medium-range temporal dependencies in mind, longer-range dependencies are not completely grasped, and this might be important for long videos. Furthermore, the model was only trained on the Celeb Face dataset, thus, generalizability to other datasets of over 231 deepfake datasets is not yet demonstrated; it would be needed to test on datasets with different image scale, frame rate, and manipulations for robustness confirmation [17].

Another possible issue is overfitting. Although dropout and L2 regularization were implemented to overcome overfitting, overfitting can still be a concern due to high model complexity, particularly with small or imbalanced datasets. Third, the interpretability of the model is relatively weak, as it is still not easy to know which specific features drive the decision (even though we have used the attention mechanism). Benefits: Advanced explainability methods like Grad-CAM are composite in this sense they could help give explanation of the model decisions and practicality for model transparency more accurately.

## 7. Conclusion

Here we propose LSTM based model architecture and attention mechanism to classify frames as FAKE and REAL of videos in Celeb Face dataset. The model achieves high classification performance by capturing temporal dependencies and focusing on the informative features using a dynamic routing mechanism. Extensive Hyperparameter Tuning A thorough hyperparameter optimization, performed using Keras Tuner's Hyperband, optimized the most critical model parameters to increase accuracy and robustness overall. Finally, regularization techniques such as dropout and L2 regularization decreased the risk of overfitting even further, reinforcing the model's ability to generalize here.

Nonetheless, some limitations were noted, such as computational cost, risk of overfitting for small datasets, interpretational issues and dependence with long-range interaction. This limitation leads to some perspectives for future work: testing on wider datasets, exploring lightweight or Transformer-based architectures, and considering more sophisticated explainability

techniques. In spite of these challenges, the proposed architecture is a strong framework for video-based deepfake detection and will help in creating reliable tools in digital forensics and media authenticity verification.

## References

[1] K. Jayakumar and N. Skandhakumar, "A Visually Interpretable Forensic Deepfake Detection Tool Using Anchors," *7th International Conference on Information Technology Research: Digital Resilience and Reinvention, ICITR 2022 - Proceedings*, 2022, doi: 10.1109/ICITR57877.2022.9993294.

[2] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1335–1348, 2023, doi: 10.1109/TIFS.2023.3239223.

[3] B. Pinhasov, R. Lapid, R. Ohayon, M. Sipper, and Y. Aperstein, "XAI-Based Detection of Adversarial Attacks on Deepfake Detectors," Mar. 2024, Accessed: Nov. 08, 2024. [Online]. Available: https://arxiv.org/abs/2403.02955v2

[4] B. Malolan, A. Parekh, and F. Kazi, "Explainable deep-fake detection using visual interpretability methods," *Proceedings - 3rd International Conference on Information and Computer Technologies, ICICT 2020*, pp. 289–293, Mar. 2020, doi: 10.1109/ICICT50521.2020.00051.

[5] Andreas R\"ossler and Davide Cozzoli, "Face{F}orensics++: Learning to Detect Manipulated Facial Images," 2019, pp. 1–11.

[6] B. Dolhansky *et al.*, "The DeepFake Detection Challenge (DFDC) Dataset," Jun. 2020.

[7] J. Jung, S. Lee, J. Kang, and Y. Na, "WWW: Where, Which and Whatever Enhancing Interpretability in Multimodal Deepfake Detection," Aug. 2024, Accessed: Nov. 08, 2024. [Online]. Available: https://arxiv.org/abs/2408.02954v1

[8] R. U. Maheshwari and B. Paulchamy, "Securing online integrity: a hybrid approach to deepfake detection and removal using Explainable AI and Adversarial Robustness Training," *Automatika*, vol. 65, no. 4, pp. 1517–1532, Oct. 2024, doi: 10.1080/00051144.2024.2400640.

[9] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 3204–3213. doi: 10.1109/CVPR42600.2020.00327.

[10] K. Tsigos, E. Apostolidis, S. Baxevanakis, S. Papadopoulos, and V. Mezaris, "Towards Quantitative Evaluation of Explainable AI Methods for Deepfake Detection," *ACM International Conference Proceeding Series*, pp. 37–45, Jun. 2024, doi: 10.1145/3643491.3660292.

[11] S. Y. Lim, D. K. Chae, and S. C. Lee, "Detecting Deepfake Voice Using Explainable Deep Learning Techniques," *Applied Sciences (Switzerland)*, vol. 12, no. 8, Apr. 2022, doi: 10.3390/app12083926.

[12] M. Arya, Priyanshu, Upwan, Akash, U. Goyal, and S. Chawla, "A Study on Deep Fake Face Detection Techniques," *Proceedings of the 3rd International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2024*, pp. 459–466, 2024, doi: 10.1109/ICAAIC60222.2024.10575149.

[13] S. H. Silva, M. Bethany, A. M. Votto, I. H. Scarff, N. Beebe, and P. Najafirad, "Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models," *Forensic Sci Int*, vol. 4, p. 100217, Jan. 2022, doi: 10.1016/J.FSISYN.2022.100217.

[14] F. Khalid, A. Javed, Q. ul ain, H. Ilyas, and A. Irtaza, "DFGNN: An interpretable and generalized graph neural network for deepfakes detection," *Expert Syst Appl*, vol. 222, p. 119843, Jul. 2023, doi: 10.1016/J.ESWA.2023.119843.

[15] P. Korshunov, A. Jain, and S. Marcel, "CUSTOM ATTRIBUTION LOSS FOR IMPROVING GENERALIZATION AND INTERPRETABILITY OF DEEPFAKE DETECTION," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2022-May, pp. 8972–8976, 2022, doi: 10.1109/ICASSP43922.2022.9747628.

[16] T. Luan, G. Liang, and P. Peng, "Interpretable DeepFake Detection Based on Frequency Spatial Transformer," 2024. [Online]. Available: https://ijetaa.com/article/view/108/

[17] M. de L. den Bouter, J. L. Pardo, Z. Geradts, and M. Worring, "ProtoExplorer: Interpretable forensic analysis of deepfake videos using prototype exploration and refinement," *Inf Vis*, vol. 23, no. 3, pp. 239–257, Jul. 2024, doi: 10.1177/14738716241238476/ASSET/IMAGES/LARGE/10.1177_14738716241238476-FIG6.JPEG.

[18] M. M. Alwateer and M. M. Alwateer, "Explainable Deep Fake Framework for Images Creation and Classification," *Journal of Computer and Communications*, vol. 12, no. 5, pp. 86–101, May 2024, doi: 10.4236/JCC.2024.125006.

[19] M. S. Rana, B. Murali, and A. H. Sung, "Deepfake Detection Using Machine Learning Algorithms," *Proceedings - 2021 10th International Congress on Advanced Applied Informatics, IIAI-AAI 2021*, pp. 458–463, 2021, doi: 10.1109/IIAI-AAI53430.2021.00079.

[20] F. Khalid, A. Javed, K. M. Malik, and A. Irtaza, "DFP-Net: An explainable and trustworthy framework for detecting deepfakes using interpretable prototypes," *2023 IEEE International Joint Conference on Biometrics, IJCB 2023*, 2023, doi: 10.1109/IJCB57857.2023.10448842.

[21] Y. Patel *et al.*, "Deepfake Generation and Detection: Case Study and Challenges," *IEEE Access*, vol. 11, pp. 143296–143323, 2023, doi: 10.1109/ACCESS.2023.3342107.

[22] X. Ma, J. Tian, Z. Li, Y. Chai, L. Zang, and J. Han, "Explainable Deepfake Detection with Human Prompts," *Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2024*, pp. 3023–3029, 2024, doi: 10.1109/CSCWD61410.2024.10580761.

[23] S. Pino, M. J. Carman, and P. Bestagini, "What's wrong with this video? Comparing Explainers for Deepfake Detection," May 2021, [Online]. Available: http://arxiv.org/abs/2105.05902

[24] Y. Xu, K. Raja, and M. Pedersen, "Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection." [Online]. Available: https://github.com/xuyingzhongguo/deepfake_

[25] V. Sunkari and A. Srinagesh, "DeepExplain: Enhancing DeepFake Detection Through Transparent and Explainable AI model," *Informatica (Slovenia)*, vol. 48, no. 8, pp. 103–110, May 2024, doi: 10.31449/inf.v48i8.5792.