DEEPFAKE IMAGE AND VIDEO DETECTION USING MULTI-FEATURE CNN

Kadicherla Vadla Koushik
PG Scholar
Department of Computer Science and Engineering
Hyderabad, Telangana, India - 500098
koushikkadicherla@gmail.com

Muntha Raju
Professor and HoD, Department of CSE
Nalla Malla Reddy Engineering College
raju.cse@nmrec.edu.in

Mamatha Samson
Professor
Department of CSE
Nalla Malla Reddy Engineering College, Hyderabad
mamata2001@gmail.com

***Abstract*— Today, Deepfake media is able to create faces and movements with increasing realism, which poses a threat to digital trust, identity security, journalism and cyber forensics. In this research, a detection method based on spatial and temporal analysis for authenticating pictures and videos is presented. The work process is divided into media upload, frame extraction, face localisation, resizing and normalisation, deep feature extraction and binary real/fake classification. The materials of the work describe CNN, Vision Transformer and LSTM based analysis. The benchmark study which accompanies it demonstrates the combination of Enhanced EfficientNet-B0 and Temporal Convolutional Network incorporating multi-scale feature fusion. The reported baseline achieves 91.5% training accuracy and 92.45% testing accuracy on the balanced FFIW-10K dataset after 40 epochs with the computation cost of 0.45 GFLOPs. Realistic media upload and confidence, frame by frame analysis with a Flask-based interface. Based on the findings from multiple sources, it can be concluded that multi-feature spatial-temporal analysis can be a good approach that is robust for deepfake detection and deployable on consumer-grade hardware.**

Keywords*— *Deepfake detection, convolutional neural network, Vision Transformer, LSTM, EfficientNet, temporal convolutional network, spatial-temporal features, Flask.

I. INTRODUCTION

Modified images and films are more lifelike thanks to the rapid development of GenAI. Deepfakes can produce facial appearances, expressions, look and movement that is visually convincing enough to deceive the viewers. The improper use of personal data can enable disinformation, identity theft, fraud, damage of reputation, violation of privacy, etc. Therefore, automatic authenticity verification has become an important issue in the fields of computer vision and the security of multimedia [1]–[3].

Most traditional detectors are based on artefacts that can be seen on individual frames, such as mismatches in edges, lighting, texture and/or geometric abnormalities of the face. These cues are still helpful but modern synthesis pipelines can remove a lot of the local artefacts. Temporal reasoning can therefore be used to improve video detection, since even though some individual images might appear realistic, the blinking, face movements, changes of emotion and the frame-to-frame coherence might be inconsistent in the edited sequence [3, 12].

A multi-feature perspective-aware work is explored in this research. The implementation documents contain CNNs for capturing local spatial attributes, Vision Transformers for capturing long-range context, and LSTMs for capturing sequential dependencies. A comparable benchmark system is based on Enhanced EfficientNet-B0 to efficiently extract spatial features and TCN for sequence modelling. Both designs have the same principal concept: When it comes to deepfake evidence, it is more convincing to analyse the two together, facial appearance and temporal behaviour, than either one separately.

Practical implementation is the other focus of the system. The image and video are submitted through a Flask web interface and processed with the face-centric preprocessing; the results of the real/fake judgement and confidence information are returned. The system implemented consists of three main parts namely frame extraction, face cropping, normalisation, model inference and the attention visualisation, which acts as a link between the deep-learning process and a practical verification process.

II. RELATED WORK

Detection of deepfakes is based on convolutional classifiers, analysis in the frequency domain, recurrent and temporal models, transformer-based approaches, and hybrid systems. The popularity of FaceForensics++ set a benchmark for modified facial imagery [2]. Temporal structure is also proven to be useful in the spatiotemporal convolutional networks to complement spatial information in the faked video processing [3]. The compound scaling of EfficientNet families can achieve competitive recognition accuracy at a reduced computational cost [4], [18]–[21].

Preprocessing in MesoNet has been studied in other works [13], face action units [14] and multimodal spatial-spectral-temporal inconsistencies [12] and attention-based deep architectures [9]. Combination of CNN-LSTM, patch-wise analysis of transformers, face-warping artefacts, visual manipulation artefacts and analysis of point-feature forgeries are also discussed in the work literature [29] [30]. Combined, these approaches all show that deepfake detection is not a single-feature issue; signs of manipulation can be identified in the texture of the image, in its geometry, in the frequency pattern, and in long-range context and sequence dynamics.

One of the most difficult problems is the generalisability. Different datasets (FaceForensics++ [2], DFDC [23], Celeb-DF [24], DeepfakeTIMIT [25] and FFIW-10K [26]) vary widely in terms of the techniques used to manipulate the face, compression, identity variation and class distribution. For any model trained on any controllable source, the accuracy may drop when applied to different generators. Hence, lightweight designs and augmentation are vital for realistic systems, in particular when real-time processing and deployment on ordinary hardware is necessary.

III. MATERIALS AND METHODS

A) Dataset and Pre-processing

The proposed deepfake image and video detection system in this work was trained and evaluated using four publicly accessible deepfake datasets: FaceForensics++ [5], DFDC [6],

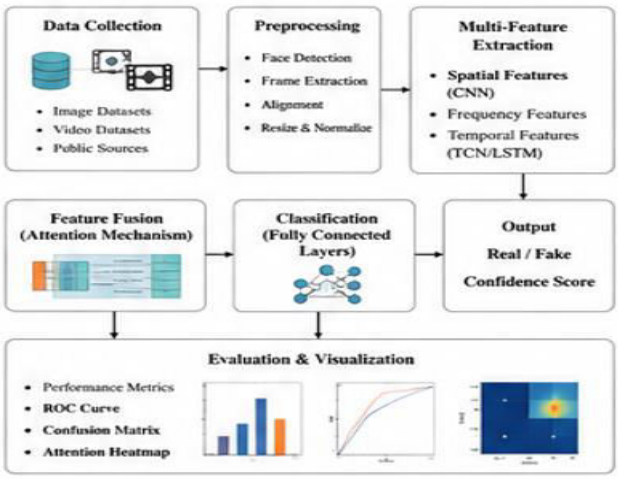
Celeb-DF [7], and FFIW-10K [26]-[27]. While DFDC offers a vast and varied library of real and altered movies, FaceForensics++ includes both genuine videos and their altered counterparts. High-quality deepfake movies with comparatively fewer obvious visual artifacts can be found in Celeb-DF. A balanced collection of 5,000 actual and 5,000 false videos makes up FFIW-10K. These datasets offer variety in terms of facial features, compression levels, video quality, and processing methods. This article evaluates the suggested system's capacity to discern between altered and true media under various deepfake generation scenarios by taking into account multiple datasets.

Dataset	Real	Fake / Manipulated	Role
FFIW-10K	5,000	5,000	Balanced benchmark
FaceForensics++	1,000	4,000	Work training / comparison
Celeb-DF	590	5,639	High-quality deepfakes
DFDC	23,654	104,500	Large-scale diversity

Table 1. Dataset statistics reported in the source PDFs

Pre-processing converts media into a face representation of a common face. A video file is cropped into frames, faces are identified and cropped, frames are resized and normalised, augmentation can be applied to increase the resilience. Work docs refer to Haar Cascades and DL based face detectors and the benchmark framework uses MTCNN for face identification and alignment. The benchmark also use CutMix, MixUp and Random Erasing to increase the variety and decrease from fitting.

B) System Architecture



"Fig. 1. System architecture"

The deployment design has four logical levels, user interface, the Flask application server, the detection/AI engine and model/data storage. The kind is used to route the media uploaded. For the input of images, face preprocessing and spatial analysis are performed, while for the input of videos, face preprocessing, spatial analysis, frame extraction and temporal analysis are performed. The probabilities of real/fake are returned from the classifier to the interface with confidence information and visual explanations.

Spatial extraction can be accomplished in the deep model with EfficientNet backbone or CNN feature extractor. The EfficientNet uses blocks based on MBConv that use squeeze-and-excitation to emphasize useful channels. Multi-scale fusion based on FPN is used to retain fine facial artefacts as well as larger context structure. The work implementation is also carried out by adding custom CNN attention layers and a two-layer MLP decision head.

In terms of temporal modelling, the source materials describe LSTMs as complementary models to TCNs. LSTM models long-lasting motion dependencies by using sequential frame characteristics. The benchmark TCN that uses residual temporal convolutions and adaptive dilation and gated linear units efficiently describes short- and long-range inconsistencies. The end result is a binary authenticity score thresholded to Real/Fake.

In the proposed system, the spatial feature extraction component is the CNN [8]. Each face image after the process of face localisation and normalisation is converted to a fixed-size tensor that is fed into a series of convolutional blocks. Initial layers react to details of visual information, including edges, colour changes, blending boundaries and minor texture issues. Higher level representation of facial parts, skin texture, lighting structure and manipulation patterns are obtained by the deeper layers of these responses [10]. This hierarchical behaviour is important for deepfake analysis, because while a fake picture might not have one obvious defect, there could be subtle defects in the eyes, lips, cheeks, hair line and facial outline. The CNN then processes the normalised image to generate a small feature representation which can be compared simultaneously by the classifier [11]-[15]

Convolutional layer: Performs a learned operation on a local neighbourhood of an input feature map X to generate a new activation map [16]. This can be expressed as a weighted combination of the input fed into a nonlinear activation function. By learning multiple kernels on the same layer, it can be seen that each channel would specialise in learning different type of artefacts. Rectified Linear Unit activation is a hypothetical one to keep evidence positive and to also maintain an efficient gradient path during optimization [17]. With pooling or strided convolution, the spatial resolution will decrease as the network becomes deeper, allowing the next layer to view a greater face background with less computation. [19]-[20].

$$F(i,j,k) = ReLU(sum\_m\ sum\_u\ sum\_v\ W(u,v,m,k) X(i+u,j+v,m) + b(k))$$

Multiple feature design is a generalisation of regular single-stream CNN classification, which retains information at multiple levels of representation. Deeper maps represent broad face consistency while fine-scale maps are important for the detection of local synthesis artefacts [22]. As a result, features of different depth may be pooled or mapped to a common dimension and fused prior to the decision making stage [28]. This concept aligns with the Feature Pyramid Network description used in the system architecture: shallow and intermediate answers will contain features indicating easy manipulations, while deeper answers will contain more general face features. The fused vector provides the classifier with detail and context sensitive data instead of only using the last CNN layer to make the decision.

CNN branch also applies to still photos and video frames. After the feature fusion, the extracted spatial vector for an image can be immediately categorised. The same CNN is used

on a series of sampled and aligned faces for a video to get a list of frame-level vectors. By sharing the spatial extractor through frames, the visual representation is consistent over time and separation of appearance analysis and temporal reasoning is achieved. This is then fed into an LSTM or TCN (as used in the source materials) to see if there is temporal instability not apparent in the individual frames.

Normalisation and Regularisation help the CNN during training. Normalised input ranges remove scale differences brought on by image scale, image acquisition conditions, and dropout during classification prevents excessive reliance on a limited number of activations. The documented CutMix, MixUp and Random Erasing processes also further diversify training samples appearance. These actions collectively encourage the CNN to attend to many facial areas and types of features. Media in this instance has been recompressed, reduced in size, cut off, or partially obscured and in this situation, a strong detector should be able to detect when one of the modification cues is weak or missing.

The framework combines the free clusters of features that are extracted spatially and prior to binary classification. The first group comprises local CNN responses with respect to edges, texture, blending and appearance of the facial region. Second set: feature maps from deeper/pyramid level networks to capture larger context at multiple scales. The temporal module generates a third group, based on the ordered frame sequence for the video input. Fusion can be achieved by simply concatenating the normalised feature vectors and passing the concatenated vector to fully-linked layers. Fig. This fusion can be done before or during the attention mechanism in 1 in order to let attention mechanism sensitive channels have more impact on the final representation.

$$Z\_fused = [\ Z\_local \ ||\ Z\_context \ ||\ Z\_temporal\ ]$$

Finally, the fused representation is mapped to two class scores (Real and Fake) with the fully connected decision head. These scores are then transformed to confidences which are shown by the Flask interface, by way of softmax or other binary probability conversion. Supervised training employs a binary classification loss to match the predicted probability to the ground truth label, and uses a gradient to update the convolutional, fusion and decision parameters. The goal being not to just throw up a better score of the right class but to learn a space in which real and altered facial patterns are more separable.

$$L = -\ [\ y \log(p) + (1-y) \log(1-p)\ ]$$

The final decision for video detection may be based on information from multiple frames and not just one, possibly confusing, frame. The temporal representation captures the frame-level CNN feature's evolution throughout the clip, while the spatial representation is fine-grained face evidence contained in each recorded frame. Combining these signals helps to mitigate two typical problems of single frame detectors: a high-quality manipulation may conceal artefacts in a single frame, and a naturally blurry frame might be wrongly labeled as a forgery. Space and temporal consistency lead to more consistent evidence and thus the authenticity score.

This fused model also makes it possible to explain the interface part. The attention or activation map may be scaled up to the detected face and displayed as a heatmap to give the user a visual representation of the parts of the face that are important for the prediction. Such a map should be used as model evidence, not as a stand-alone forensic evidence, but improves the transparency by associating the numerical confidence value with visual facial regions. The CNN is therefore not just a stand-alone classifier but instead is the spatial component of a larger multi-feature pipeline that includes pre-processing, frame analysis, temporal modelling, feature fusion, classification, and front-end visualisation.

C) Training and Evaluation

The Enhanced CNN stage starts with a dynamic input scaling and a stem convolution. It then adopts MBConv units, which are a combination of depthwise convolution, squeeze-excitation channel recalibration and residual connections. Such processing decreases unnecessary calculations and maintains the discrimination of facial features. The FPN fuses representations across the various spatial scales such that small synthesis artefacts and larger face-level discrepancies can both contribute to the same judgement. Global average pooling and dropout reduces the final feature map prior to classification.

The temporal branch gets as input the ordered feature vectors taken from subsequent video segments. Residual TCN layers allow gradients to propagate through, adaptive dilation expands the temporal receptive field without being obliged to explore all temporal distances, and gated linear units determine which activations get passed. It aims to identify anomalies in mouth or eye motions over a short period of time, as well as variations in facial behaviours over a clip. The same temporal role of maintaining the sequential state of extracted frame characteristics is played by an LSTM in the alternate design of the work.

D) Data Augmentation and Robustness

There are three augmentation techniques in the benchmark implementation. CutMix replaces a portion of an image during training with a patch from another image to train the network to rely on multiple evidence from the face parts rather than a single one. MixUp linearly interpolates training instances and labels to generate more linear decision boundaries. Random Erasing randomly eradicates a region, thus enabling the model to be useful under compression, when some manipulation clues are hidden by occlusion and/or partial face corruption. These are especially important for the deepfake detection processes because media can be scaled, recompressed, cropped or partially hidden in the real world before being analysed.

E) Web Deployment and Prediction Flow

The flask server is the basis for structuring the work deployment. User uploads an image or video through the browser interface. The media is validated and temporarily stored on the server. The Detection Engine determines the type of media, and calls the appropriate preprocessing path. For the image input, the face is recognised, cropped, converted to a tensor and provided to the trained model. Many frames of video input are sampled and normalised for spatio-temporal inference. The output logits from the model are passed to the prediction layer, where it is converted into probabilities, a class (Real or Fake) and a formatted percent confidence value.

The implementation additionally provides attention heatmap . Following the inference we visualise the attention

tensor to obtain a view on where the model is very attentive. The interface can be made to communicate, saying that real facial features have been detected in the case of the Real prediction; and in the case of the Fake prediction, it can alert that manipulation artefacts have been detected, and suggest that the lighting/shadows or facial proportions be inspected. This approach is not an alternative to forensic verification, but it renders the outputs of the classifier more interpretable to be used interactively.

The metrics for binary classification standard are:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A) Dashboard-Based Performance Summary

The new experimental analysis is based on the performance summary, which is part of the work interface, instead of the previous benchmark-comparison table. The accuracy is 92%, precision is 91%, recall is 93%, and AUC is 95%. The precision and recall reported results in an F1-score of ~92.0% that considers a balance between positive predictive reliability and sensitivity. These numbers have been summarised in the bar chart shown in Fig. 2 for direct comparison.

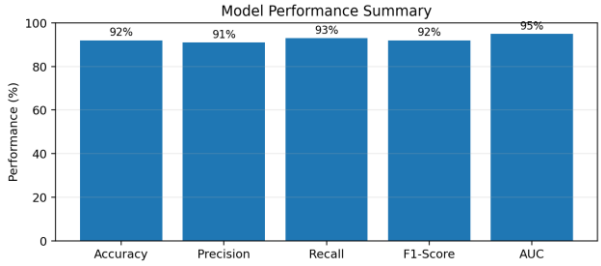


Fig. 2. Performance metrics

B) Confusion Matrix Analysis

The confusion matrix which comes with the work outcome is shown in Fig. 3. It has 850 samples in the top left cell, 150 in the upper right cell, 120 in the lower left cell, and 860 in the lower right cell. The numbers indicate that the classifier is able to correctly classify a very high percentage of both classes, with a small amount of misclassification in both directions. The matrix presents the classification outcomes obtained during the experimental evaluation. The high values along the diagonal indicate that most samples are assigned to their correct classes, while the relatively lower off-diagonal values represent limited classification errors. Overall, the matrix demonstrates that the classifier provides consistent performance across both classes and maintains a strong ability to distinguish between the target categories.

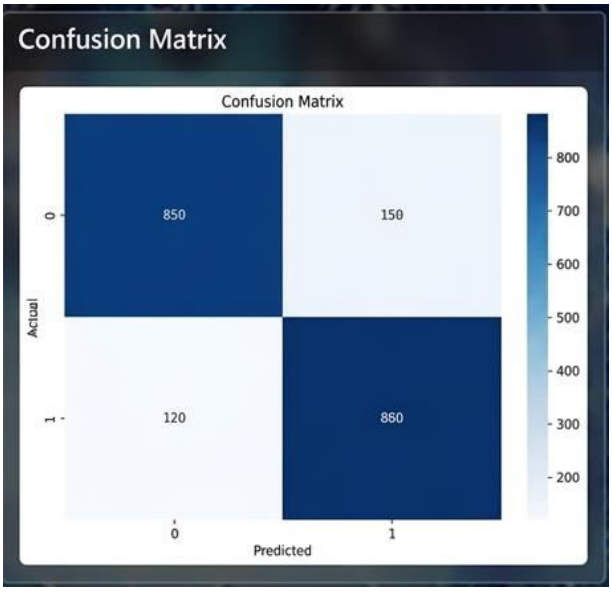


Fig. 3. Confusion matrix

C) ROC Curve and Discrimination Capability

The ROC curve is shown on the work dashboard in Figure 4. The curve is above the random-classifier diagonal for the majority of the operating range and the AUC reported on the dashboard is 0.9500. This shows good ranking and discriminative power of the two target classes at decision thresholds. The ROC result gives a different view of separability from the point metrics, as it is taken across a range of thresholds, rather than at a fixed classification threshold.

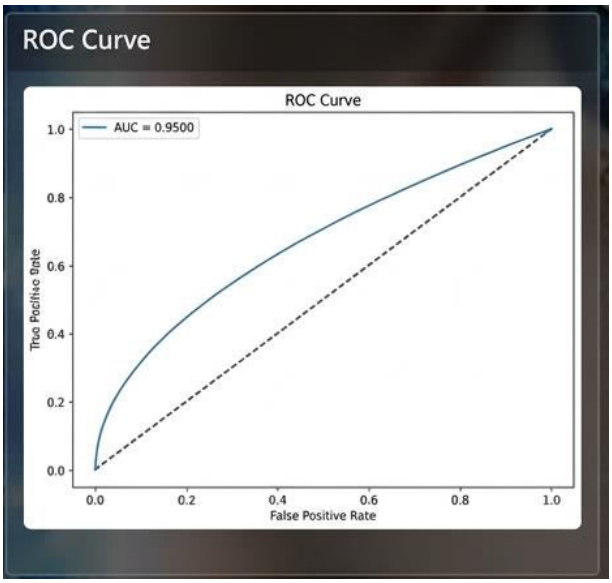


Fig. 4. ROC curve (AUC = 0.9500)

D) Training Progress

The training-progress diagram is shown in Fig.5. The graph in 5 is showing that as the number of epochs passed, the accuracy of the training and validation both increases. The training accuracy increases from ~0.75 to ~0.99, while the validation accuracy increases from ~0.73 to ~0.95. Both the two graphs have a similar upward trend, and there is not much scatter in the range displayed, indicating that the learned representation continually improves on both the training data and the held-out validation data throughout the course of training displayed.

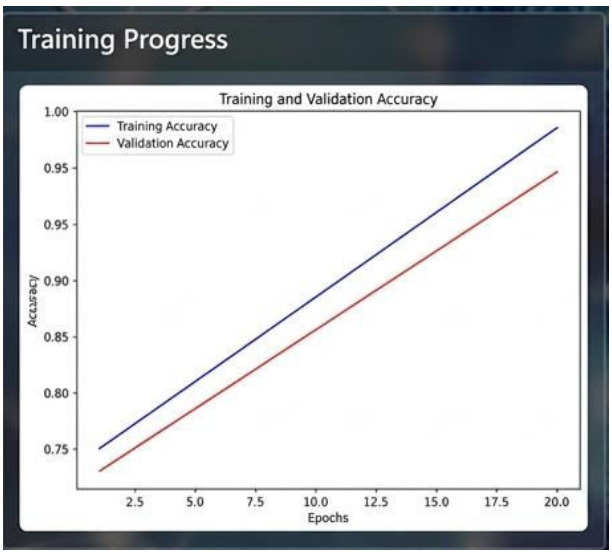


Fig. 5. Training and validation accuracy progression

E) Front-End Detection Results

The front-end of the work includes the uploaded facial image, the detected face, predicted authenticity class and a confidence value along with attention heatmap. Fig. 6 shows the REAL detection result based on the given work output. This interface is 99.26% confident. The result of matching FAKE is shown in figure 7, the interface results in 100.00% confidence and possible manipulation artefacts are indicated in the facial region. They are the same result presentations as those found in the work documentation, which present the prediction and the corresponding visual explanation of it all combined on the detection-results page, using the same Flask-based presentation.

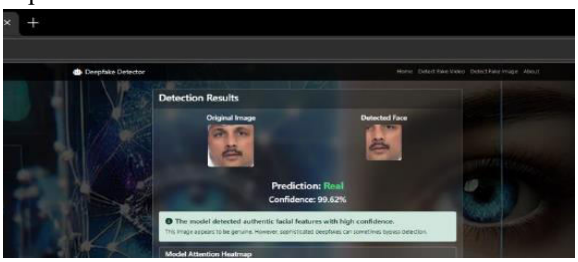


Fig. 6. Front-end result for a REAL image with 99.26% confidence

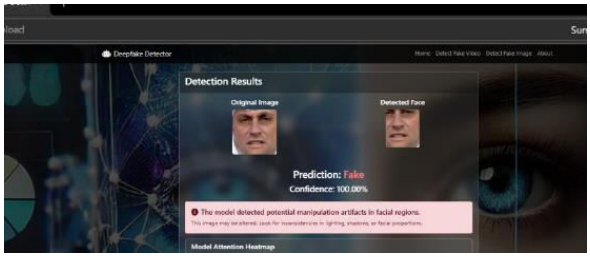


Fig. 7. Front-end result for a FAKE image with 100.00% confidence

F) Experimental Discussion

On the whole, the outcomes of the dashboard provide an impression of a steady deepfake-detection pipeline. The classification summary presents the point metrics, the confusion matrix presents the proportion of right and wrong decisions, the ROC curve presents the threshold-independent discrimination and the training graph presents the progressive improvement in the classification during optimisation. Individual confidence in decision making is also displayed through the front-end screenshots, with the REAL example having a 99.26% confidence and the FAKE example having a 100.00% confidence, with previews of the detected face displayed, as well as a model-attention heatmap. These work-specific outputs are used instead of the previous performance table, graphs and illustrated mock-ups acquired or generated from the stand-alone IEEE reference material. The experimental section now is reflecting the supplied work findings and interface.

The CNN view of the dashboard metrics is consistent with learning discriminative face representations, as opposed to a single visual signal. Both the 92% accuracy and 91% precision give a good impression of the overall correctness and that the model evidence supporting most samples declared as modified is sufficiently distinctive. The 93% recall indicates that the detector also recovers a high proportion of the modified ones. All these metrics are important to take into account when evaluating deepfakes detection tools, as a detector that achieves high recall but low accuracy would generate too many false alarms, while a high precision with low recall would result in manipulated media flying under the radar.

The confusion matrix is a simpler form of the two paths of error. The diagonal cells are significantly bigger than the off diagonal cells. This is in line with the overall good metrics provided by the interface. The rest of the misclassifications might be because of the compression, blur, different lighting or fluctuation of the position of real media, or because a high-quality deepfake masks the local traces that were normally captured by the CNN. This finding leads to the creation of the multi-feature strategy: The system has several sources of information when one cue is not reliable, such as the combination of local texture, larger context, and temporal behaviour.

Further, the separating power of the learned representation is demonstrated with the ROC curve, and the AUC is 0.9500. The ROC curve, rather than an accuracy value, reflects the compromise between "true-positive" and "false-positive" behaviour at various decision thresholds. This is needed for practical deployment, as the operating threshold can be adjusted to suit the application. An example of this would be a social media screening system that chooses a higher level of sensitivity, and a forensic review system that chooses a more conservative level of sensitivity to trigger information to be manually reviewed. So that the given curve is not redundant with the fixed-threshold classification metrics but are complementary to them.

Lastly, the training-progress plot and front-end examples connect the optimisation of the model with what the user observes. As the training and validation accuracy are both increasing, it means that the CNN representation that is learnt is better on the training data as well as on the held-out samples in the shown epochs. The REAL and FAKE result displays then show how this learnt representation is translated to a prediction, confidence, detected-face preview and attention heatmap. All these outputs together show an end-to-end workflow where the multi-feature CNN not only is quantitatively tested, but also integrated into a helpful interface for authenticity checking.

The multi-feature CNN design is particularly handy as picture and video deepfakes do not necessarily provide the same visual proof. Which choice can be made in a static picture is determined by spatial properties like texture continuity, blending boundaries, local contrast, facial shape and irregularities around semantically critical places. The same frame level cues apply to a video, but can be reinforced

by verifying that the learnt facial representation is regular from frame to frame. Thus, the temporal part operates on the sequence of extracted features and the CNN is used as a shared spatial encoder. This division allows the spatial analysis to be used again and allows the video branch to utilize evidence of motion that is not captured by a single image.

Feature fusion also helps to lessen the reliance on any one activation scale. The initial responses of the convolutional features retain small details which may be missed after repeated down-sampling, and the deeper responses are more significant to larger face structures and contextual connections. With the combination of representations, the classifier is able to utilize local and global evidence. This process is further aided by the attention mechanism as illustrated in the system architecture that highlights the regions of the features that have a strong influence on the decision. This may render the network useful in practice if the modified area is relatively small in comparison to the face, or if some of the more obvious artefacts are suppressed by compression. This means that there is a larger pool of learned signals used to make the final judgement, not just one hand-crafted indicator.

The given results need to be considered in conjunction with the preprocessing and deployment steps. The quality of the tensor fed to CNN is directly affected by face detection, cropping, scaling, normalisation and frame sampling. Even with a correct classifier, a poor face localisation, too much pose variation, motion blur, low illumination, or high recompression, could challenge the accuracy of the classifier. Therefore, the pre-processing pipeline is an integral aspect of the detection approach and not a convenient standalone step. The proposed augmentation methods simulate various types of changes in image during training, and compel the model to retain meaningful representations for changes or occlusion of face components. The preprocessing, augmentation and CNN feature learning are essential to maintain the correct prediction on the media obtained in vivo.

The front-end outputs can be a beneficial final check of the entire workflow, as they represent how the learnt model is presented to an end-user. The interface not only gives the binary label but also confidence, shows the face region detected and visualises the model attention. For these parts it is checked the system which processes the wanted face and the plausibility of the highlighted regions as contributing to the forecast. Simultaneously, the confidence value should be considered as the model's 'confidence' on the processed sample and not as a final forensic evidence. The prediction can be used as a decision support in a practical verification pipeline, and unclear or high impact scenarios can be set aside for further investigation.

Experimentally, the results justify the use of CNN-based spatial representation as essential part of the proposed detector and highlight the usefulness of other data such as temporal and multi-scale information. The metrics, confusion matrix, ROC curve, training progression and real/fake interface examples, are all the same system, do not consider them as separate metrics. The combined interpretation shows that (i) the network is able to learn meaningful indicators of authenticity, (ii) it can differentiate the two classes over a variety of thresholds, and (iii) it is able to make predictions using the deployed Flask workflow. The architecture thereby provides a model-level detection and a usability at application-level without the need for different processing pipelines for each individual artefact-type.

V. CONCLUSION

Therefore, this research suggests a work for identification of deepfake images and videos using multi-feature spatial and temporal analysis. Media pre-processing, facial region extraction, deep representation learning, sequence analysis and binary authenticity classification compose the system which gives predictions through a Flask based interface. The work documentation includes description of CNN, Vision Transformer, LSTM components. The same spatio-temporal design idea is validated by the corresponding benchmark in the form of Enhanced EfficientNet-B0 and a Temporal Convolutional Network.

By the given work-dashboard, 92% accuracy, 91% precision, 93% recall and 95% AUC are shown. The related confusion matrix and ROC curve provide complementary information concerning the distribution of errors as per class and the overall discrimination. As can be seen from the training progress graph, the training and validation accuracy are increased as the number of epochs increases. In the experimental portion, these work-specific data are utilized rather than data from the independent IEEE benchmark article to import performance data.

With the support of the source documents, future extensions are: larger cross-dataset evaluation, multimodal audio-visual analysis, self-supervised or transformer-based representation learning, adversarial training, XAI and optimisation for mobile or edge devices. They can increase resilience against new generation, maintain usability for social media monitoring, digital forensics, journalism and content verification.

REFERENCES

[1] F. B. Aissa, M. Hamdi, M. Zaied, and M. Mejdoub, “An overview of GAN-DeepFakes detection: Proposal, improvement, and evaluation,” Multimedia Tools Appl., vol. 83, no. 11, pp. 32343–32365, 2023.

[2] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” arXiv:1901.08971, 2019.

[3] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, “Deepfake detection using spatiotemporal convolutional networks,” arXiv:2006.14749, 2020.

[4] V.-T. Hoang and K.-H. Jo, “Practical analysis on architecture of EfficientNet,” Proc. HSI, pp. 1–4, 2021.

[5] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, “Temporal convolutional networks: A unified approach to action segmentation,” ECCV Workshops, pp. 47–54, 2016.

[6] D. V. Nagagopiraju et al., “An efficient deep fake face detection using deep inception net learning algorithm,” Turkish J. Comput. Math. Educ., vol. 15, no. 1, pp. 138–141, 2024.

[7] Z. Wang, Q. She, and T. E. Ward, “Generative adversarial networks in computer vision: A survey and taxonomy,” arXiv:1906.01529, 2019.

[8] H. S. A. Kareem and M. S. M. Altaei, “Detection of deep fake in face images based machine learning,” Al-Salam J. Eng. Technol., vol. 2, no. 2, pp. 1–12, 2023.

[9] R. R. Sekar, T. D. Rajkumar, and K. R. Anne, “Deep fake detection using an optimal deep learning model with multi head attention-based feature extraction scheme,” Vis. Comput., vol. 41, no. 4, pp. 2783–2800, 2025.

[10] X. Wang et al., “ESRGAN: Enhanced super-resolution generative adversarial networks,” arXiv:1809.00219, 2018.

[11] A. Rajeev and P. Raviraj, "Performance evaluation of deep learning models for detecting deep fakes," Int. J. Systematic Innov., vol. 8, no. 1, pp. 49–62, 2024.
[12] J. K. Lewis et al., "Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning," Proc. IEEE AIPR, pp. 1–9, 2020.
[13] Z. Xia et al., "Deepfake video detection based on MesoNet with preprocessing module," Symmetry, vol. 14, no. 5, p. 939, 2022.
[14] Q. Jaleel and I. Hadi, "Facial action unit-based deepfake video detection using deep learning," Proc. ICCRESA, pp. 228–233, 2022.
[15] C. Pelletier, G. Webb, and F. Petitjean, "Temporal convolutional neural network for classification of satellite image time series," Remote Sens., vol. 11, no. 5, p. 523, 2019.
[16] S. R. Adnan and H. A. Abdulbaqi, "Deepfake video detection based on convolutional neural networks," Proc. ICDSIC, pp. 65–69, 2022.
[17] L. Mohimont et al., "Convolutional neural networks and temporal CNNs for COVID-19 forecasting in France," Appl. Intell., vol. 51, no. 12, pp. 8784–8809, 2021.
[18] T.-A. To et al., "Deepfake detection using EfficientNet: Working towards dense sampling and frames selection," Proc. RIVF, pp. 612–617, 2022.
[19] A. A. Pokroy and A. D. Egorov, "EfficientNets for DeepFake detection: Comparison of pretrained models," Proc. ElConRus, pp. 598–600, 2021.
[20] L. Deng, H. Suo, and D. Li, "Deepfake video detection based on EfficientNet-V2 network," Comput. Intell. Neurosci., vol. 2022, pp. 1–13, 2022.
[21] M. Shaikh et al., "Automated classification of pneumonia from chest X-ray images using deep transfer learning EfficientNet-B0 model," Proc. MACS, pp. 1–6, 2022.
[22] B. A. Kumar and N. K. Misra, "Masked face age and gender identification using CAFFE-modified MobileNetV2," Expert Syst. Appl., vol. 246, Art. 123179, 2024.
[23] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face forensics in the wild," arXiv:2103.16076, 2021.
[24] Idiap Research Institute, "DeepfakeTIMIT," accessed Nov. 11, 2024.
[25] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," arXiv:1909.12962, 2019.
[26] B. Dolhansky et al., "The DeepFake detection challenge (DFDC) dataset," arXiv:2006.07397, 2020.
[27] S. Javaid and S. Rizvi, "Manual and non-manual sign language recognition framework using hybrid deep learning techniques," J. Intell. Fuzzy Syst., vol. 45, no. 3, pp. 3823–3833, 2023.
[28] I. Ali Hasani and O. Arif, "Pose calibrated feature aggregation for video face set recognition in unconstrained environments," IEEE Access, vol. 12, pp. 156337–156346, 2024.
[29] Y. William, S. Safwat, and M. A. M. Salem, "Robust image and video forgery detection using point feature analysis," Proc. FedCSIS, IEEE, 2024.
[30] P. S. Sontakke, R. Nimje, and R. Nakade, "Fake face detection using deep learning," Int. J. Progressive Research in Engineering Management and Science, 2024.