POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Deepfake video detection using a hybrid ResNeXt and LSTM architecture

Hibrit ResNeXt ve LSTM mimarisi kullanılarak deepfake video algılama

***Yazar(lar) (Author(s))**: Nurcan YARDIMCI[1], Mohamed Ibrahim ABDI[2], Burhan ERGEN[3]*

ORCID[1]: 0009-0002-0476-9856

ORCID[2]: 0009-0002-7874-8740

ORCID[3]: 0000-0003-3244-2615

To cite to this article: Yardımcı N., Abdi M. İ., and Ergen B., "Deepfake Video Detection Using a Hybrid ResNeXt and LSTM Architecture", ***Journal of Polytechnic***, 29(2):290223:1-10 (2026).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Yardımcı N., Abdi M. İ., ve Ergen B., "Deepfake Video Detection Using a Hybrid ResNeXt and LSTM Architecture", ***Politeknik Dergisi***, 29(2):290223:1-10 (2026).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1721371

Deepfake Video Detection Using a Hybrid ResNeXt and LSTM Architecture

Highlights

- *Deepfake video detection using hybrid ResNeXt-50 and LSTM architecture.*
- *Spatial feature extraction with frame-based CNN, temporal feature extraction with LSTM.*
- *Performance evaluation with benchmark datasets such as DFDC, Celeb-DF, FaceForensics++, DFD.*
- *Strong generalization against various manipulation techniques with up to 95.7% accuracy.*
- *Scalable and lightweight model proposal suitable for real-world video verification systems.*

Graphical Abstract

In this study, a hybrid deep learning architecture that integrates ResNeXt-50 based spatial feature extraction and LSTM based temporal modeling is proposed for the detection of deepfake videos. The proposed system detects facial regions by segmenting videos into frames, then feature vectors are obtained with ResNeXt-50 and these vectors are transferred to the LSTM layer. The model has been tested on various datasets and has achieved accuracy rates of up to 95.7%.

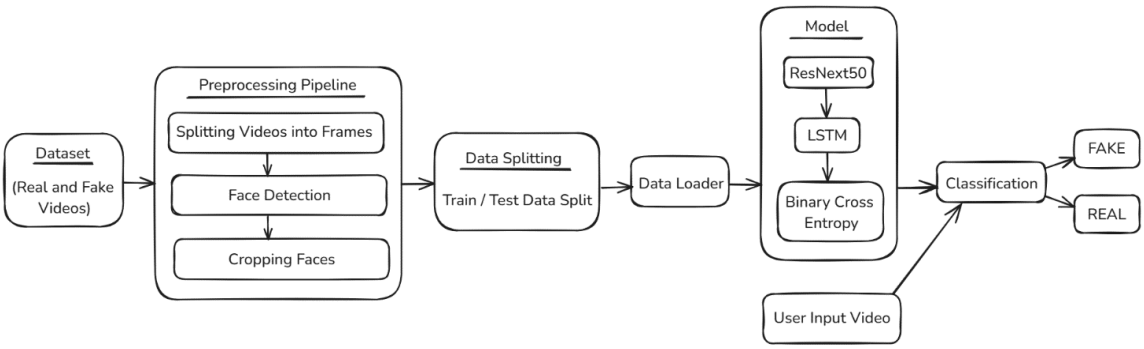

Figure. Proposed Methodology

Aim

To investigate the effectiveness of ResNeXt-50 and LSTM architecture in detecting deepfake videos.

Design & Methodology

In this study, videos were separated into frames with pre-processing steps, face regions were detected and cropped to make them suitable for the model. Then, spatial features were extracted from each frame with ResNeXt-50 and these features were transferred to the LSTM network to perform temporal modeling; the model was tested on four different datasets.

Originality

This study is one of the studies where ResNeXt-50 and LSTM architectures were used together and tested on four different deepfake datasets, showing high generalization success.

Findings

The proposed model achieved 95.7% accuracy on the DFDC dataset, 94.9% on FaceForensics++, 91% on DFD, and 88% on Celeb-DF. This success is due to the joint modeling of spatial and temporal information.

Conclusion

The hybrid architecture based on ResNeXt-50 + LSTM provides an effective, high-accuracy, and generalizable solution for deepfake video detection. The model can be integrated into real-world applications and can be used in areas such as media forensic analysis, social media monitoring, and digital rights management.

Declaration of Ethical Standards

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Deepfake Video Detection Using a Hybrid ResNeXt and LSTM Architecture

Araştırma Makalesi / Research Article

Nurcan YARDIMCI[1*], Mohamed Ibrahim ABDI[1], Burhan ERGEN[1]

[1]Fırat University, Faculty of Engineering, Department of Computer Engineering, Turkey

(Geliş/Received : 17.06.2025 ; Kabul/Accepted : 21.09.2025 ; Erken Görünüm/Early View : 28.09.2025)

ABSTRACT

The growing spread of deepfake materials presents a serious threat to individual privacy, media credibility, and public trust. Existing detection methods often struggle to generalize across various manipulation techniques and video quality levels. This study proposes a hybrid architecture based on deep learning (DL) is introduced, which leverages the spatial feature extraction strengths of ResNeXt-50 along with the temporal sequence modeling capabilities of LSTM networks. The suggested framework handles video input by initially obtaining frame-wise features via a pretrained ResNeXt-50 backbone and then examining temporal dynamics through an LSTM layer. Experimental evaluations were conducted using benchmark datasets, including Deepfake Detection Challenge (DFDC), Celeb-DF, FaceForensics++, and DFD. Findings indicate that the developed model significantly outperforms conventional CNN-LSTM combinations, attaining 95.7% accuracy on the DFDC dataset and above 90% on the other datasets. This research highlights the practical applicability of hybrid DL techniques in real-world video authentication systems and contributes a high-performance solution to the growing field of synthetic media detection.

Keywords: DeepFake, ResNeXt-50, LSTM, Deepfake detection, Hybrid deep learning.

Hibrit ResNeXt ve LSTM Mimarisi Kullanılarak Deepfake Video Algılama

ÖZ

Deepfake içeriklerin artan yaygınlığı, bireysel gizlilik, medya güvenilirliği ve kamu güveni için ciddi bir tehdit oluşturmaktadır. Mevcut tespit yöntemleri genellikle çeşitli manipülasyon teknikleri ve video kalite seviyeleri arasında genelleme yapmakta zorlanmaktadır. Bu çalışma, ResNeXt-50'nin mekansal özellik çıkarma güçlerini LSTM ağlarının zamansal dizi modelleme yetenekleriyle birlikte kullanan derin öğrenmeye dayalı bir hibrit mimari sunmaktadır. Önerilen çerçeve, başlangıçta önceden eğitilmiş bir ResNeXt-50 omurgası aracılığıyla kare bazlı özellikler elde ederek ve ardından bir LSTM katmanı aracılığıyla zamansal dinamikleri inceleyerek video girişini işler. Deneysel değerlendirmeler, DFDC, Celeb-DF, FaceForensics++ ve DFD dahil olmak üzere kıyaslama veri kümeleri kullanılarak yürütülmüştür. Bulgular, geliştirilen modelin geleneksel CNN-LSTM kombinasyonlarından önemli ölçüde daha iyi performans gösterdiğini, DFDC veri kümesinde %95,7 ve diğer veri kümelerinde %90'ın üzerinde doğruluk elde ettiğini göstermektedir. Bu araştırma, hibrit derin öğrenme tekniklerinin gerçek dünya video kimlik doğrulama sistemlerinde pratik uygulanabilirliğini vurgulamakta ve sentetik ortam tespitinin büyüyen alanına yüksek performanslı bir çözüm sunmaktadır.

Keywords: DeepFake, ResNeXt-50, LSTM, Sahte video tespiti, Hibrit derin öğrenme.

1. INTRODUCTION

Advances in machine learning and deep learning have led to the emergence of deepfake content that can replicate the appearance and behavior of individuals very closely. These manipulated media can fabricate scenarios, distort reality, and damage reputations. Public figures are especially vulnerable, as altered visuals and audio can be used to spread misinformation and undermine trust in the media [1].

As deepfake technologies become more accessible and sophisticated, the threats they pose to personal privacy, social credibility, and national security continue to grow. Therefore, robust and accurate detection systems are essential for this purpose. Traditional detection methods, which rely on handcrafted features or static analysis, are inadequate for complex and subtle manipulations. DL-based approaches, particularly those combining CNN and LSTM networks, have shown significant promise in learning spatial and temporal inconsistencies in videos [2].

CNNs are effective in extracting spatial characteristics from individual frames, like facial landmarks, textures, and boundaries, whereas LSTMs capture temporal dependencies and detect anomalies across sequences, including unnatural blinking patterns, lip-sync mismatches, and abrupt transitions [3]. However, many

*Sorumlu Yazar (Corresponding Author)
e-posta : nurcann.yardimci@gmail.com

existing models emphasize either spatial or temporal information, limiting their ability to generalize across diverse manipulation techniques and video quality [4].

In light of these difficulties, this study presents a hybrid architecture that integrates ResNeXt-50 and LSTM networks. ResNeXt-50 was chosen for its strong representational capacity, which leverages grouped convolutions and residual connections to extract high-quality spatial features. Its modular design and ImageNet pre-trained weights also facilitate efficient transfer learning, which is particularly beneficial when training data are limited [5].

LSTM networks, on the other hand, model the sequential dynamics of facial movements by analyzing frame-level embeddings over time. They are particularly effective in identifying temporal irregularities and motion-based clues indicative of manipulation [6]. Prior studies have demonstrated that integrating LSTM with CNN-extracted features improves the classification performance in deepfake detection tasks [7].

The suggested model comprises two primary components: a spatial feature extractor based on ResNeXt-50 and a temporal modeling unit using LSTM layers. Together, these enable the detection of both visual artifacts and motion inconsistencies. Evaluation on benchmark datasets, including DFDC, Celeb-DF, FaceForensics++, and DFD, demonstrated the model's effective generalization across various deepfake generation techniques and quality levels. This hybrid design offers a scalable, accurate, and robust solution for detecting deepfake content, thereby supporting practical applications like media verification, forensic analysis, and platform moderation.

Due to the limitations of current methods, there is a significant need for models that do not rely on hand-crafted features or deeply complex structures, but also offer a combination of accuracy, speed, and generalization ability.

1.1. Related Work

Saikia et al. proposed a CNN–LSTM based hybrid model for detecting deepfake videos. This method performs feature extraction using optical flow and captures motion inconsistencies in consecutive frames. Both temporal and spatial analysis are included in the process. An accuracy rate of 91.21% was achieved in FaceForensics++, 79.49% in Celeb-DF and 66.26% in DFDC, demonstrating that early detection can be achieved even with a small number of frames [8].

Koçak et al. developed a hybrid approach for deepfake video detection. In this study, face regions were extracted from video frames obtained from the DFDC dataset using the MTCNN (Multi-Task Cascaded Convolutional Neural Network) method. Then, feature extraction was performed using Xception and ResNet50 deep learning models. The resulting feature vectors were evaluated with various machine learning-based classification algorithms and eight different hybrid models were proposed. This approach has provided an effective method for detecting deepfake content, demonstrating higher success compared to similar studies in the literature [9].

Sagar and Arukonda developed a CNN–LSTM based hybrid model for detecting fake video content. Spatial features were extracted with ResNeXt50_32x4d and temporal features were extracted with LSTM. In the experiments conducted on FaceForensics++ and Celeb-DF V1 datasets, 96.67% and 91.80% accuracy was achieved, respectively. Researchers state that although the method has shown high success, it has limitations in terms of real-time use, scalability and generalizability [10].

Korkmaz and Alkan proposed a solution that utilizes deep learning algorithms. This solution aims to identify deepfake videos. In this study, the EfficientNet architecture, which has rarely been used in previous studies, was preferred for feature extraction and tested on the DFDC dataset. BlazeFace detector is used to extract faces from video frames. Different versions of EfficientNet (B0–B5) were trained and compared, and an accuracy rate of approximately 91% was achieved [11].

Battula and Rajasekaran used an AdaBoost-powered model to identify deepfake videos. In the study, facial regions were extracted from videos using the DFDC dataset, and spatial, temporal and frequency-based features were obtained from these areas. The AdaBoost algorithm created a strong model by combining weak classifiers consisting of simple decision trees, and as a result of the experiments, high precision, recall and F1 scores were reported with 86.5% accuracy. It was emphasized that the method used is more advantageous than CNN-based models, especially in detecting subtle manipulations [12].

Another hybrid ResNeXt-LSTM model achieved an accuracy of 91% on the Celeb-DF v2 dataset [13].

Further studies have examined the sequence length as a key variable. For instance, a CNN-LSTM model using 80-frame sequences achieved up to 92.49% accuracy on the DFDC dataset, revealing the impact of temporal depth on performance [14]. Zhang et al. aimed to identify fake videos by developing a method that uses coordinated movements of facial features. In this context, facial points associated with the Coordinated Motion Landmark Mining Strategy (CMLMS) were identified. Then, the Landmark Temporal Dynamic Relationship Module (LTDRM) was developed and the spatiotemporal relationships of these points were analyzed. The

presented model's performance has been tested on the FaceForensics++, Celeb-DF, and DFDC datasets. Specifically, an accuracy rate of 90.7% has been reported for the DFDC dataset [15].

Jayashre and Amsaprabhaa developed the HODFF-DD method for the detection of deep fake videos, which extracts features with InceptionResNetV1/V2, performs temporal analysis with BiLSTM and is optimized with Spotted Hyena Optimizer. The method achieved 92–95% accuracy on the FaceForensics++ and Kaggle Deepfake Detection Challenge datasets and 76% accuracy on the FakeAVCeleb dataset [16].

Selvaraj et al. proposed an adversarial training based method for deepfake video detection. This approach, based on the EfficientNet architecture, combines VAT and Two-Gen-BAT techniques. In experiments conducted on the FaceForensics++ dataset, the model's robustness increased and 88.3% accuracy was achieved [17].

Amritha Devi and Simon proposed a lightweight CNN model named DeepGuardNet inspired by MesoNet for the detection of deepfake videos. In tests using the Celeb-DF dataset, the model achieved 91% accuracy and stood out with its low computational cost [18].

Cybersecurity-focused systems, such as AuthentiScan, built on ResNeXt-50 and LSTM, achieved 90.95% accuracy using 40-frame sequences from a 6,000-clip dataset [19], further affirming the utility of this hybrid structure in practical scenarios.

Although the ResNeXt-50 and LSTM combination has been utilized in earlier studies, our approach differs in terms of design goals and evaluation strategy. Instead of introducing a novel architecture, we focused on assessing the generalization capability of this hybrid model across multiple datasets (e.g., DFDC, FaceForensics++, Celeb-DF, and DFD) without relying on additional attention or frequency-domain enhancements. This positions our work as a rigorous validation of a lightweight and practical deepfake-detection framework. By emphasizing simplicity and efficiency, our model contributes to a reproducible and scalable approach suitable for real-world deployments.

2. MATERIAL and METHOD

In this study, we propose a special two-stage DL architecture for detecting deepfake videos using spatial and temporal information specific to facial movements.

2.1. Datasets

The performance of the proposed ResNeXt-50 and LSTM based hybrid model in detecting deepfake content is evaluated on four large datasets that are widely used in the literature and represent different forgery methods. These datasets include various face manipulation techniques, image qualities and shooting conditions. The scope of each is summarized below, and the sample sizes used in the study are specified.

The Celeb-DF dataset consists of 590 real videos obtained from YouTube and 5,639 high-quality fake videos generated from these [20]. In this study, 598 real and 596 fake videos were used.

FaceForensics++ was created by applying Methods like DeepFakes, Face2Face, FaceSwap, and NeuralTextures to 1,000 original videos [21]. From this dataset, which offers rich content in terms of compression level and quality scenarios, 1,989 video samples were used in this study.

Among the most extensive datasets available for deepfake detection is the one released as part of Facebook's Deepfake Detection Challenge (DFDC), which features a collection of over 100,000 video samples [22]. For the purposes of this study, a subset comprising 3,292 videos was utilized—consisting of 1,726 authentic and 1,566 manipulated clips. These videos were divided into three subsets as 70% training (2304 videos), 20% validation (663 videos) and 10% testing (325 videos) to be used in the training and evaluation process. This distribution provided a balanced measurement of both the learning and generalization success of the model.

Thanks to the diversity in the datasets, the model was able to demonstrate its generalization capacity over different forgery methods, quality levels, and facial manipulation strategies with high accuracy. The proposed framework comprises two core modules: a frame-wise spatial feature extractor based on the ResNeXt50 convolutional neural network and a sequence modeling unit based on LSTM networks. The architecture is illustrated in Figure 1.

Proposed DL-based deepfake detection model architecture. The system begins with a dataset that includes both genuine and manipulated video content. In the preprocessing phase, the videos were divided into frames, face detection was performed, and the face region was cropped to fit the model input. The dataset is divided into different parts for the purpose of training and evaluation of the model. In the model phase, each video frame was subjected to feature extraction using the ResNext50 CNN architecture and transferred to the LSTM network to capture temporal consistency. Binary cross-entropy loss function was preferred during the training process of the model. The trained model classifies both the user-provided video input and the test examples originating from the dataset into real and fake.

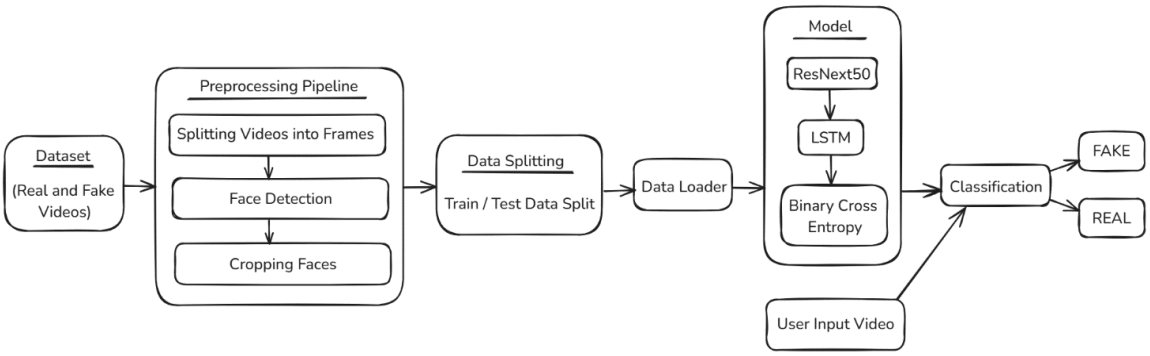

Figure 1. General Structure of DL Based Deepfake Detection Model

2.2. Frame-Level Spatial Feature Extraction Using ResNeXt50

The first layer of the proposed structure in the study derives detailed and selective spatial information from each frame in the videos. We employed ResNeXt50**,** a variant of ResNet known for its improved performance through cardinality-based aggregation of transformation paths**.** This model architecture is designed to extract subtle visual irregularities that often go unnoticed by the human eye. These include anomalies in skin texture, unnatural variations in illumination, and edge artifacts resulting from artificial face synthesis processes. Each input video was preprocessed by sampling a fixed number of frames (20 frames per video) either uniformly or randomly, ensuring that the temporal dimension of the facial expression dynamics was preserved. These frames are scaled to 112 × 112 pixels to balance computational efficiency and spatial resolution. To maintain consistency with the pretrained ResNeXt50 architecture, each image underwent normalization employing the commonly used ImageNet mean and standard deviation parameters.

Each frame was independently passed through ResNeXt50, where we discarded the final classification layer and extracted the global average pooled feature vector (typically of dimension 2048). This results in a temporal sequence of feature vectors $\{ f_1, f_2, \dots , f_{20}\}$, where each $f_i \in \mathbb{R}^d$ represents the abstract spatial features from the i-th frame.

2.3. Temporal Modeling With LSTM

While spatial information is crucial for detecting visual forgeries in individual frames, deepfake videos often exhibit temporal inconsistencies, such as subtle anomalies in motion, eye-blinking frequency, lip synchronization, or unnatural transitions between frames. To capture such temporal artifacts, we employed an LSTM network.

The LSTM receives a sequence of frame-level feature vectors as input and processes them in order to learn to model both short-term transitions and long-range dependencies in the temporal domain. The LSTM is composed of one or more layers with a configurable hidden dimension (e.g., 512 or 1024 units) and optionally incorporates dropout regularization to prevent overfitting. The output of the LSTM at the final time step (or alternatively, a pooled summary of all time steps) is used as a holistic representation of the video sequence.

2.4. Fully Connected Classification Layer

In the final stage, the model generates its predictions by passing through a fully connected layer; these predictions are distributed to the classes corresponding to the probability values by softmax transformation. The cross-entropy loss function was used in the training phase, and the Adam algorithm was preferred for optimization. Classes in the dataset were created in a balanced manner, and strategies such as early stopping were structured to be applicable to the risk of overfitting during the learning process of the model.

The temporal features obtained from the LSTM are passed through a dense layer. This layer expresses the hidden state of the video as a numerical probability value, indicating its authenticity or manipulation. To translate this value into a probability ranging from 0 to 1, a sigmoid function is employed. This output can be interpreted as;

$$y \approx 1 : \text{high likelihood that the video is real}$$
$$y \approx 0 : \text{high likelihood that the video is fake}$$

2.5. Training Objective

During training, the model was optimized using binary cross-entropy (BCE) loss, defined as;

$$\mathcal{L}_{BCE} = -[\, y . \log(\hat{y}) + (1 - y) . \log(1 - \hat{y}) \,] \quad (1)$$

Where $y \in \{ 0, 1\}$ is the ground-truth label and $\hat{y}$ is the expected probability. This loss function effectively penalizes incorrect predictions while maintaining the numerical stability during optimization.

Due to the high memory consumption of video-based input data, the batch size is set to 4. This batch size maintained the stability of the training while ensuring the model fit into GPU memory. Adam algorithm was preferred for optimization. The learning rate is set to 1×10^{-5}. This value is chosen to avoid excessive updates during fine-tuning of the pre-trained ResNeXt-50 based network. To reduce class imbalance in the dataset, CrossEntropyLoss function with weight vector [1,15]

was used. The training process was performed for 20 epochs and the validation loss was monitored for early stopping.

Additionally, various data augmentation techniques were applied during the training process to increase the diversity of the dataset and improve the generalization ability of the model. Specifically, random horizontal flipping was used to simulate mirror images of faces, random resized cropping was used to represent different zoom and framing scenarios, and color jitter was used to reflect changes in light and color distributions. These augmentation steps made the model more robust to real-world differences and helped it learn generalizable spatio-temporal features instead of memorizing training examples.

2.6. Mathematical Formulation of ResNeXt-50 Architectures

In this section, we present the mathematical structures of the ResNeXt-50 components that form the backbone of our DeepFake detection model. ResNeXt-50 was responsible for spatial feature extraction from individual video frames.

ResNeXt-50 enhances the traditional ResNet architecture by incorporating grouped convolutions that divide the input channels into *G* separate groups. Each group is convolved independently, reducing the computational complexity while enabling multi-branch feature learning.

$$Y = \sum_{g=1}^{G} X_g * W_g \quad (2)$$

In equation 2, *Y* is obtained by summing the convolution outputs of each group g, where X_g is the input feature map of the group and W_g is its corresponding filter. In simpler terms, grouped convolution splits the input channels into smaller groups, enabling efficient multi-branch feature learning while reducing computational cost [3].

Each ResNeXt block includes a residual connection that helps avoid vanishing gradient problems and enables the use of deeper architectures.

$$\mathcal{Y} = F\,(x, \{W_i\}) + x \quad (3)$$

In equation 3, $\mathcal{Y}$ is obtained by adding the input feature representation x to the transformed output $F\,(x,\{W_i\})$, here F represents a sequence of operations consisting of the sequential application of convolution, batch normalization and activation steps. In simpler terms, this residual connection allows the network to pass information directly from earlier layers to later ones, helping to prevent gradient vanishing in deep architectures [3].

The bottleneck block in the ResNeXt-50 architecture consists of three layers. First, a 1×1 convolution is applied to reduce the dimensions. Then, 3×3 grouped convolution is performed for feature extraction. In the final stage, a 1×1 convolution is used again to restore the dimensions. This structure can be expressed mathematically as follows:

$$\mathcal{Y} = W_3 * \sigma\,(W_2 * \sigma\,(W_1 * x)) \quad (4)$$

W_1, W_2 and W_3 in Equation 4 represent the weights of the convolution layers; σ represents the ReLU activation function [3].

$$z_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{G} x_i\, j_k \quad (5)$$

In equation 5, *H* and *W* denote the height and width of the feature map, while z_k is the pooled value for the kth channel. In simpler terms, GAP computes the average of all activations in each feature map, producing one representative value per channel [3].

$$p = \text{softmax}\,(W_{fc}\,.\,z + b_{fc}) \quad (6)$$

In equation 6, W_{fc} and b_{fc} are the weight and bias parameters of the fully connected layer, z is the feature vector obtained after GAP, and p is the resulting probability distribution over the output classes. In simpler terms, the softmax function converts the outputs into class probabilities [3].

$$\mathcal{L} = -\sum_{i=1}^{C} \mathcal{Y}_i \ \log(p_i) \quad (7)$$

In equation 7, $\mathcal{Y}_i$ represents the correct label, p_i represents the estimated probability for that class, and *C* represents the sum of the class numbers. In simpler terms, cross-entropy penalizes the model more when it assigns a low probability to the correct class, encouraging accurate predictions [3].

2.7. Mathematical Formulation of LSTM Architectures

Here, we detail the mathematical formulation of the LSTM units that constitute the core of our DeepFake detection framework. This module is designed to model temporal relationships across consecutive video frames. Within LSTM architectures, the forget gate identifies which pieces of information need to be discarded from the earlier cell state C_{t-1}. This mechanism utilizes a sigmoid function that processes both the current input x_t and the prior hidden state h_{t-1}, thus producing a gate vector expressed below.

$$f_t = \sigma\,(W_f\,.\,[\,h_{t-1}, x_t] + b_f) \quad (8)$$

The vector shown in equation 8 takes values in the range $f_t \in [0,1]$. When the value approaches 0, it signals that the corresponding data will be discarded; conversely, values near 1 suggest that the data should be preserved. In this way, the network can learn how much of the previous cell information to remember or to ignore.

In the LSTM cell, the input gate (i_t) controls how much new information is added to the memory. This gate is calculated via sigmoid activation using the current input x_t and the previous hidden state h_{t-1} as shown in equation 9:

$$i_t = \sigma\,(W_i.[h_{t-1}, x_t] + b_i)$$
(9)

At the same time, the candidate cell state $\tilde{C}_t$ is created with the hyperbolic tangent (tanh) function as shown in equation 10:

$$\tilde{C}_t = \tanh(W_C\,.\,[h_{t-1}, x_t] + b_C$$
(10)

The gate scales the candidate state so that only the necessary information is added to the memory. Thus, the model avoids unnecessary information overload by updating the cell state with only the important information.

The cell state update equation defines how the internal memory of the LSTM cell is updated. This process is accomplished by preserving past information and adding new data. This is expressed in Equation 11.

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$
(11)

In this context, $f_t \odot C_{t-1}$ denotes the element-wise multiplication between the forget gate (f_t) and the prior cell state (C_{t-1}), signifying the portion of memory that is retained. Conversely, $i_t \odot \tilde{C}_t$ indicates the result of combining the input gate (i_t) with the candidate cell state ($\tilde{C}_t$), representing newly added information to the memory. The resulting C_t combines the new input with past information to create an updated cell state. In an LSTM cell, the output gate (o_t) determines which parts of the updated cell state will be exported in the current time step. This gate is calculated via the sigmoid function, as shown in Equation 12, and the previous hidden state (h_{t-1}) and the current input (x_t) are used together during the process.

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o$$
(12)

Then, as shown in Equation 13, the new hidden state h_t is calculated by first applying the tanh activation to the updated cell state C_t, and multiplying the output by the output gate:

$$h_t = o_t \odot \tanh(C_t)$$
(13)

This hidden state h_t is transferred to the next LSTM cell and can also be used in subsequent tasks such as classification. Thus, the model carries past information and produces an output that is meaningful in the current context.

3. RESULT and DISCUSSION

The evaluation of the system developed in this study focuses on its capacity to distinguish between real and deepfake content. The dataset used in this context is created with videos obtained from publicly available deep fake datasets such as DFDC, FaceForensics++, Celeb-DF and DFD. In each dataset, real and fake samples are allocated in a balanced manner, and all data are randomly partitioned with 80% allocated for training and the remaining 20% set aside for testing

During model training, each video was represented by a sequence of 20 randomly selected frames. The spatial features extracted from these frames were obtained with the ResNeXt-50 architecture and transferred to the LSTM network in a sequential structure. The learning of the model was performed with the cross-entropy loss function and the Adam optimization algorithm. Accuracy, F1 score and loss values were monitored throughout the training process, and analysis indicated that the model became stable after approximately 10–15 epochs and did not show a tendency for over-learning.

Figure 2 illustrates the training and validation loss values captured during model training over 20 epochs. As noted, the training loss consistently reduced, whereas the validation loss remained low and stable after around the 10th epoch. The model trains successfully without overfitting and generalizes well to data not encountered during training. The smooth convergence of both curves suggests a well-tuned architecture and optimization strategy, where optimization was performed via Adam, and the model error was minimized using the cross-entropy criterion.

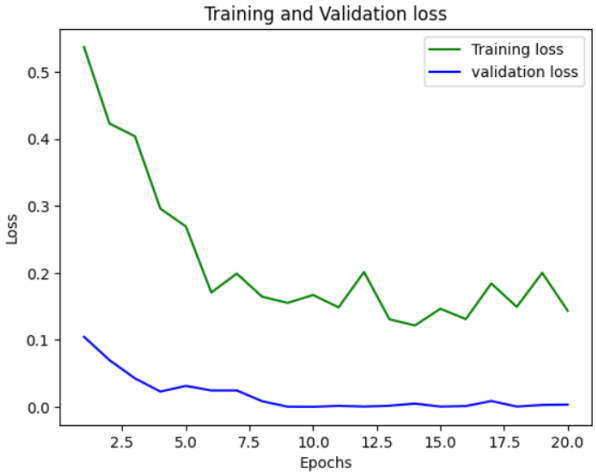

Figure 2. Loss Evaluation

The accuracy of the model observed during training is illustrated in Figure 3. The training accuracy increased rapidly, reaching 95% after approximately the 6th epoch, and then stabilized at approximately 99%. The validation accuracy remained high from the beginning and stabilized at 99% from the 10th epoch. This shows that the model has successfully learned both the training data and has a high generalization capacity on the validation data. The closeness of the respective training and validation loss trends also supports the fact that the model does not tend to over-learn.

The model's performance has been also compared with the test results of different architectural structures. Table 1 presents the accuracy values obtained by different CNN+LSTM-based models on the DFDC dataset.

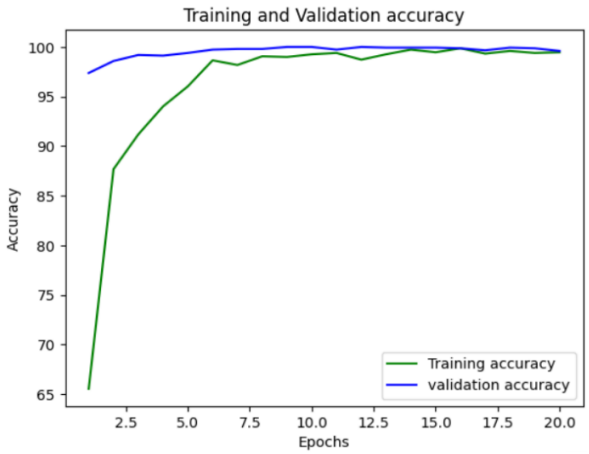

Figure 3. Accuracy Validation

Table 1. Accuracy Comparison of CNN+LSTM Based Models on the DFDC Dataset

Model	Accuracy
3D-CNN + LSTM	77%
EfficientNetV2-S + LSTM	87.3%
ConvNext-Tiny + LSTM	86.4%
ConvNextV2-Tiny + LSTM	87.6%
Proposed Model ResNext-50 + LSTM	**95.7%**

In the testing phase, the proposed ResNeXt + LSTM architecture gave the most successful results by reaching 95.7% accuracy rate on the DFDC dataset.

High success was also achieved on the other datasets. The overall performance was found to be quite satisfactory, with 88% accuracy on Celeb-DF, 94.9% on FaceForensics++, and 91% on DFD. This success was supported by the ability of the LSTM network to learn temporal inconsistencies (e.g., facial expression mismatch and eye blink anomalies). In Table 2, the accuracy values of the proposed model on different datasets are presented comparatively.

Table 2. Accuracy of the Proposed Model on Different DeepFake Datasets

Dataset	Accuracy
Celeb – DF	88%
FaceForensics++	94.9%
DFD	91%
Facebook DFDC	95.7%

In addition to the accuracy of the model, a confusion matrix was obtained to examine the performance on a class basis in more detail. This matrix shows the correct and incorrect classification of the fake and real videos by the model. Additionally, to measure the classification success, precision, recall and F1-score metrics were evaluated based on confusion matrix data. Not only the overall accuracy of the model but also its discrimination power between classes was evaluated.

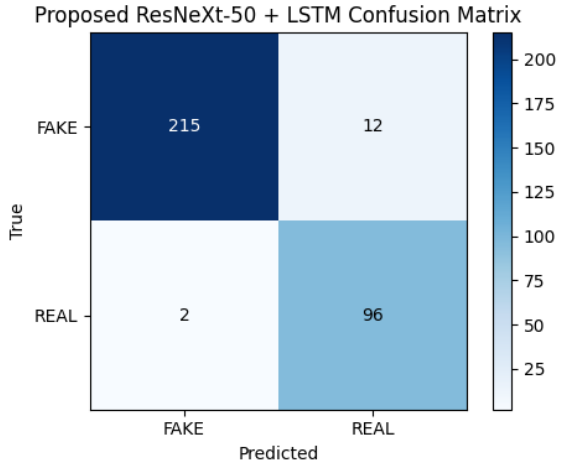

Figure 4. Confusion matrix outputs of ResNext-50 + LSTM model

Precision, recall, and F1-score values were calculated using True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP) values obtained from the complexity matrices. The equations used to calculate the above-mentioned metrics are given below.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (14)$$

$$F1\ Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (15)$$

$$Precision = TP / (TP + FP) \quad (16)$$

$$Recall = TP / (TP + FN) \quad (17)$$

According to the results obtained from 325 videos in the test set, the model achieved 95.7% accuracy, 99.1% precision, 94.7% recall and 96.8% F1-score values. These findings show that the model has high accuracy and low false positive rate in detecting fake videos.

The experiments conducted have shown that spatiotemporal DL architectures provide significant advantages in DeepFake detection compared to traditional CNN-based models. When the powerful feature extraction capacity of ResNeXt-50 is combined with the temporal learning ability of LSTM, fake videos can be effectively detected through facial expression inconsistencies, facial expressions, and transitions in time.

3.1. Comparison with Other Models

The efficacy of deepfake detection models is contingent ~~not~~ only on their architectural complexity but also on their capacity to generalize across datasets and manipulation techniques. Table 3 provides a comparative analysis of several prominent models in the literature, alongside the proposed ResNeXt-50 + LSTM hybrid architecture introduced in this study.

As illustrated in Table 3, previous architectures such as EfficientNetB4 [11], CNN + LSTM [8,14], and HODFF-DD [16] achieved competitive accuracies ranging from

Table 3. Evaluation of the Proposed Deepfake Detection Architecture Against Existing Models in Terms of Accuracy

References	Dataset	Model	Accuracy
[11]	DFDC	EfficientNetB4	91%
[19]	DFDC, Celeb-DF, FaceForensics++	ResNext-50 + LSTM	90.9%
[14]	DFDC	CNN + LSTM	92.4%
[16]	FaceForensics++, DFDC, FakeAVCeleb	HODFF-DD (InceptionResNetV1 + InceptionResNetV2 + BiLSTM, Spotted Hyena Optimizer)	94.5%
[8]	Celeb-DF, FaceForensics++, DFDC	CNN + LSTM	91.2%
Proposed Model	**DFDC, Celeb-DF, DFD, FaceForensics++**	**ResNext-50 + LSTM**	**95.7%**

91% to 94.5% across benchmark datasets. However, the proposed ResNeXt-50 + LSTM model attained the highest accuracy of 95.7%, surpassing all comparative methods. This superior performance stems from the synergistic integration of ResNeXt-50 and LSTM layers. While ResNeXt-50 enhances spatial representation through its multi-path structure, which increases the number of parallel transformations (cardinality) to extract richer and more diverse feature representations, LSTM layers effectively capture temporal dependencies such as unnatural facial dynamics and motion artifacts. Unlike earlier approaches that largely focused on either spatial or temporal domains, the proposed model unifies both aspects in a computationally efficient framework, thereby ensuring more reliable detection of deepfake manipulations across diverse datasets Furthermore, most existing models are evaluated on a single dataset or rely on limited feature representations. In contrast, the proposed model has been validated across four major deepfake datasets (DFDC, Celeb-DF, DFD, and FaceForensics++), demonstrating strong generalization capabilities. This cross-dataset robustness underscores the adaptability of the model in real-world environments, where video quality, compression, and manipulation techniques vary significantly. Finally, the novelty of this work lies not only in the architecture itself but also in its strategic optimization for deepfake characteristics: high-level spatial semantics via grouped convolutions in ResNeXt, sequential temporal modeling tailored for facial behavior inconsistencies via LSTM, balanced training on diverse datasets with varying manipulation strategies, and a practical application design that enables integration into forensic, media, and law enforcement pipelines. Collectively, this study contributes a highly accurate, generalized, and practically deployable deepfake detection system that establishes a new benchmark for hybrid spatiotemporal architectures.

This paper proposes a robust and effective hybrid DL architecture designed for deepfake video detection, which integrates the spatial extraction power of ResNeXt-50 with the temporal sequence learning capabilities of LSTM networks. By leveraging the strengths of both components, the proposed model successfully identifies subtle manipulations across video frames and temporal inconsistencies, which are often overlooked by traditional CNN-based methods. The model was evaluated on four benchmark datasets—DFDC, Celeb-DF, FaceForensics++, and DFD—where it achieved superior accuracy rates, including 95.7% on the DFDC, establishing a new performance benchmark among similar architectures.

These results are notable not only for their high accuracy but also for demonstrating the model's ability to perform consistently across different datasets and manipulation techniques. This reveals the possibility of integrating the proposed method into real-world applications, such as media forensics, social media monitoring, digital rights management, and law enforcement tools for video verification.

The modular structure of the architecture facilitates efficient adaptation and scalability, enabling its deployment in environments with limited computational resources through optimization and pruning techniques. Moreover, the use of pretrained ResNeXt-50 ensures efficient training on limited labeled data, which is challenging in deepfake detection.

3. CONCLUSIONS

In this study, a hybrid architecture combining ResNeXt-50 and LSTM networks is proposed for deepfake video detection. The model achieved 95.7% accuracy on the DFDC dataset and also showed strong generalization performance on the Celeb-DF, FaceForensics++ and DFD datasets. These findings suggest that the integration of temporal modeling with advanced spatial feature extraction offers an effective approach to detect subtle

manipulations in video sequences. In addition to its academic contributions, the proposed method has potential applications in digital forensics, social media surveillance, content verification, and law enforcement video verification applications.

The high accuracy achieved by the proposed method demonstrates its importance against current problems arising from the misuse of synthetic media. Deepfake videos carry serious risks such as misinformation, political manipulation, online fraud and violation of personal rights. In this respect, the study contributes not only to the academic field but also to the solution of social and ethical problems.

However, there are some limitations. Programming changes during the training process and model performance may vary slightly across studies. Additionally, the experiments are limited to specific benchmark datasets, which may not fully reflect the full range of manipulations in the real world.

In future studies, we plan to use transformer-based approaches instead of LSTM and model sequential attention between video frames more effectively. In addition, integrating multiple data types such as audio-video can increase detection accuracy. The development of lightweight models optimized for real-time applications is another important step that will expand the practical applicability of the method. These orientations will not only strengthen the robustness and flexibility of the proposed approach but will also contribute to reducing the social and ethical risks associated with the use of synthetic media.

ACKNOWLEDGEMENT

This work is supported by the TÜBİTAK 1002-A Rapid Support Module, project number 124E844.

DECLARATION OF ETHICAL STANDARDS

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission

AUTHORS' CONTRIBUTIONS

Nurcan YARDIMCI: Software, Writing-Original Draft, Review, Editing, Methodology

Mohamed Ibrahim ABDI: Writing, Review, Editing, Methodology

Burhan ERGEN: Supervision, Methodology, Review, Validation

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] Rani, E. G., Bhuvaneshwari, P., Darekar, R. G., and Anusha, D. "Enhanced deepfake video classification and detection: A ResNext-LSTM approach for improved accuracy". In ***Data Science & Exploration in Artificial Intelligence,*** 468-476, (2025).

[2] Petmezas, G., Vanian, V., Konstantoudakis, K., Almaloglou, E. E., and Zarpalas, D., "Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification". ***Multimedia Tools and Applications,*** 1-20. (2025).

[3] Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K., "Aggregated residual transformations for deep neural networks", In ***Proceedings of the IEEE conference on computer vision and pattern recognition,*** 1492-1500, (2017).

[4] Tan, M., and Le, Q., "Efficientnet: Rethinking model scaling for convolutional neural networks", In ***International conference on machine learning,*** 6105-6114, PMLR, (2019).

[5] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition", In ***Proceedings of the IEEE conference on computer vision and pattern recognition*** (pp. 770-778), (2016).

[6] Hochreiter, S., and Schmidhuber, J., "Long short-term memory", ***Neural computation***, 9(8), 1735-1780, (1997).

[7] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T., "Long-term recurrent convolutional networks for visual recognition and description", In ***Proceedings of the IEEE conference on computer vision and pattern recognition*** 2625-2634, (2015).

[8] Saikia, P., Sharma, A., and Yadav, R., "A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features", ***arXiv preprint arXiv:2208.00788,*** (2022).

[9] Koçak, A., Alkan, M., and Arıkan, S. M., "Deepfake Video Detection Using Convolutional Neural Network Based Hybrid Approach", ***Politeknik Dergisi,*** 28(3), 957-968, (2025).

[10] Sagar, N. K., and Arukonda, S., "A Novel CNN-LSTM Approach for Robust Deepfake Detection", ***Procedia Computer Science***, 258, 1844-1855, (2025).

[11] Korkmaz, Ş., and Alkan, M., "Derin Öğrenme Algoritmalarını Kullanarak Deepfake Video Tespiti", ***Politeknik Dergisi,*** 26(2), 855-862, (2023).

[12] Devi, B. T., and Rajasekaran, R., "Deepfake Video Detection Using Ada-Boosting on the DFDC Dataset", ***Procedia Computer Science***, 258, 1091-1101, (2025).

[13] Vamsi, V. V. V. N. S., Shet, S. S., Reddy, S. S. M., Rose, S. S., Shetty, S. R., Sathvika, S., and Shankar, S. P., "Deepfake detection in digital media forensics", ***Global Transitions Proceedings,*** *3*(1), (pp.74-79), (2022).

[14] Antad, S., Arthamwar, V. V., Deshmukh, R. K., Chame, A. U., and Chhangani, H. P., "A Hybrid approach for DeepfakeDetection using CNN-RNN", In ***2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0***, (pp. 1-6), IEEE, (2024).

[15] Zhang, Y., Niu, R., Zhang, X., Chen, S., Wang, M., and Li, X. "Exploring coordinated motion patterns of facial

landmarks for deepfake video detection", ***Applied Soft Computing***, 174, 112974, (2025).

[16] Jayashre, K., and Amsaprabhaa, M., "Safeguarding media integrity: A hybrid optimized deep feature fusion based deepfake detection in videos", ***Computers & Security***, 142, 103860, (2024).

[17] Selvaraj, P., Jagatheesaperumal, S. K., Marimuthu, K., Saravanan, O., Alkhamees, B. F., and Hassan, M. M., "Deepfake Detection Using Adversarial Neural Network", ***Computer Modeling in Engineering & Sciences (CMES)***, 143(2), (2025).

[18] Amritha Devi, N., and Simon, P., "DeepGuardNet: A Novel CNN Architecture for DeepFake Image Detection", In ***Procedia Computer Science***, 258, 811-818, Elsevier, (2025).

[19] Kumar, N., and Kundu, A., "Cyber security focused deepfake detection system using big data", ***SN Computer Science,*** 5(6), 752, (2024).

[20] Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S., "Celeb-df: A large-scale challenging dataset for deepfake forensics", In ***Proceedings of the IEEE/CVF conference on computer vision and pattern recognition***, 3207-3216, (2020).

[21] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M., "Faceforensics++: Learning to detect manipulated facial images", In ***Proceedings of the IEEE/CVF international conference on computer vision,*** 1-11, (2019).

[22] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C., "The deepfake detection challenge (dfdc) dataset", ***arXiv preprint arXiv:2006.07397,*** (2020).