

detect deepfake videos. The GRU performs well in tasks of sequence learning and overcomes the gradient vanishing and explosion problems of the standard recurrent neural network [16]. The proficiency of the attention mechanism has been proven in several areas including machine translations, image captioning, question answering, speech recognition [17], and event detection [18]. A comparative study with recent state-of-the-art deepfake detection methods is conducted in terms of accuracy, Area Under Receiver Operating Characteristic (AUROC) curve metric, precision, recall, F1-score, sensitivity, and specificity.

The rest of this work is organized as follows: Section II presents the literature review for deepfake video detection methods. Section III presents the newly proposed method for deepfake video detection. Section IV is dedicated to the experimental results and analysis. The conclusion and future work are presented in Section V.

II. LITERATURE REVIEW

The progress of AI-based video and voice generation methods raised the ease of creating natural and highly realistic deepfakes that can never be distinguished. Since deepfakes violate security and pose a real threat to society, several researchers have directed their interest to create methods for detecting deepfakes. However, they concentrate on detecting the deepfakes either in video frames or audio modality.

Some of the existing deepfake visual video detection methods spot the manipulation by targeting specific spatial and temporal artifacts that are generated during the fake creation process. Some other detection methods are data-driven that do not target any specific artifacts and distinguish the manipulation by classification [3]. The deepfake visual video detection methods can be categorized into Convolution Neural Network (CNN)-based methods [19, 20, 21, 22], methods that are based on CNN with a temporal network [23, 24, 25, 26, 27], handcrafted feature-based methods [28], and handcrafted feature-based methods with deep networks [29, 30]. This is illustrated in Fig. 1.

The work of [19] detected the deepfakes by exploiting artifacts left by the generation methods when warping the target image to be consistent with the source video. It used four pre-trained CNN models for detecting fake contents; ResNet101, VGG16, ResNet50, and ResNet152. Since deepfake videos suffer from inconsistency among the inter-frames, Hu et al. [20] introduced two branches that are based on CNNs to capture those local and global inconsistencies and then detect deepfakes. Rana and Sung [21] proposed a deep ensemble learning method for detecting deepfake videos. Their method depended on combining several deep base-learners and then training a CNN on these learners to build an ameliorated classifier. In [22], a fine-tuned InceptionResNetV2 model followed by the XGBoost model was employed to capture discrepancies in the spatial domain of fake videos and then individuate deepfakes. The FakeApp creates forged videos that had intra-frame and temporal inconsistencies between frames.

Such inconsistencies were detected using InceptionV3 CNN and long short-term memory (LSTM) models [23]. As AI-generated fake videos lack normal eye blinking, Li et al. [24] introduced the VGG16-LSTM to capture the temporal regularities in the eye blinking process and then distinguish the deepfakes. Most deepfake videos are created frame-by-frame where each forged face is created independently. This causes incoherence in the temporal domain of the face region; discontinuity and flickering. As a result, Zheng et al. [25] introduced a fully temporal convolution network that aimed to learn the temporal discrepancies while removing spatial ones. Then, a temporal transformer encoder followed by a multi-layer perceptron was employed to learn the long-range inconsistencies along the time dimension, and then distinguish the deepfakes. In [26], a 2D CNN-based Spatio-temporal learning model was introduced to learn and capture spatial and temporal inconsistencies of forged videos. This temporal inconsistency was captured from both vertical and horizontal directions over adjacent frames and helped in detecting the fakes. The work of [27] introduced a fine-tuned EfficientNet-b5 model followed by the bidirectional LSTM model and densely connected layer. It aimed to discover the Spatio-temporal inconsistencies in deepfake videos and then distinguish the authenticity of videos. Deepfakes were created by joining the generated face into the source image. This produced errors in facial landmark locations that were detected by estimating the 3D head poses for real and deepfake videos. Then, the estimated difference of head poses was fed into the Support Vector Machine (SVM) for deepfake detection [28]. Khalil et al. [29] proposed a model that employed the local binary patterns descriptor to analyze the texture of real and fake videos. Additionally, a CNN-based enhanced high-resolution network was used to automatically capture informative multi-resolution representations of these videos. Then, the output of both was fed into the capsule network to individuate deepfakes. Ismail et al. [30] introduced a hybrid method in which two feature extraction methods were employed to learn and extract enriched spatial features from the detected face frames of video. These methods were a CNN that was based on the Histogram of Oriented Gradient (HOG) method and the improved XceptionNet. Their outputs were merged to be fed into GRUs sequence to extract the spatiotemporal features and detect the fake videos.

The deepfake audio detection methods can be categorized into handcrafted feature-based methods [31, 32], methods that are based on low-level features with CNN [14, 33, 34, 35], methods that rely on using low-level features with CNN and temporal network [37, 38], and end-to-end deep networks-based methods [39]. This is presented in Fig. 2.

The work of [31] extracted several low-level-features; Constant-Q Cepstral Coefficients (CQCC), Cepstrum, Mel-Frequency Cepstrum Coefficients (MFCC), inverted MFCC, Linear Predictive Cepstral Coefficients (LPCC), and LPCC-residual features. These features were utilized along with the Gaussian Mixture Model (GMM) to detect the forged audio.