International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.1 (2025) pp. 886-898
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Deepfake Detection Based on Visual Lip-sync Match and Blink Rate

Homam El-Taj[1*], Fatima Alammari[2], Joud Alkhowaiter[3], Layal Bogari[4], Renad Essa[5]

[1]Dar Al-Hekma University, Cybersecurity Department, Jeddah, Saudi Arabia
*** Corresponding Author Email:** htaj@dah.edu.sa - **ORCID:** 0000-0001-7565-4760

[2]Dar Al-Hekma University, Cybersecurity Department, Jeddah, Saudi Arabia
Email: fmalammari@dah.edu.sa **-ORCID:** 0009-0000-2728-0311

[3]Dar Al-Hekma University, Cybersecurity Department, Jeddah, Saudi Arabia
Email: jialkhowaiter@dah.edu.sa- **ORCID:** 0009-0008-3735-5661

[4]Dar Al-Hekma University, Cybersecurity Department, Jeddah, Saudi Arabia
Email: lybogari@dah.edu.sa - **ORCID:** 0009-0000-8837-619X

[5]Dar Al-Hekma University, Cybersecurity Department, Jeddah, Saudi Arabia
Email: rmessa@dah.edu.sa **-ORCID:** 0009-0008-2978-2161

Article Info:

DOI: 10.22399/ijcesen.755
Received : 12 December 2024
Accepted : 13 January 2025

Keywords :

Deepfake Detection,
visual lip-sync matching,
Blink Rate,
Artificial Intelligence,

Abstract:

Deepfake technology has emerged as a significant challenge to the authenticity of digital media, necessitating innovative detection methods. This paper introduces TrueSync, an advanced application for detecting deepfake videos by integrating two critical detection features: lip-sync analysis and blink rate monitoring. Leveraging a hybrid approach combining CNN-LSTM and SyncNet models, TrueSync processes visual and temporal features to identify anomalies in lip movement synchronization and eye blinking patterns. The application utilizes a modular pipeline to analyse these features independently and then fuses the results for a comprehensive detection score. This approach enhances detection accuracy and provides users with reliable tools to combat sophisticated manipulations. By proposing this scalable solution, TrueSync addresses the increasing difficulty in distinguishing authentic videos from manipulated content, fostering trust in digital media.

1. Introduction

Deepfake technology has rapidly advanced, utilizing artificial intelligence to manipulate video, audio, and images, creating highly realistic yet fabricated media. While these advancements have proven beneficial in areas such as accessibility for individuals with speech impairments, they also present significant societal risks. The malicious use of deepfakes for deception purposes such as identity theft, and misinformation, has raised serious concerns about the integrity of digital content and its implications for privacy and security. A particularly dangerous subset of deepfakes includes those that alter lip movements to match altered audio, a form of manipulation known as lip-sync deepfakes. These videos are challenging to detect due to their subtlety, as the imperfections are localized primarily to the lip region, making it easier to produce convincing fakes. Similarly, inconsistencies in blinking patterns, which can deviate from natural human behavior, are another key indicator of deepfake manipulation. The difficulty of identifying these anomalies emphasizes the need for more robust detection systems capable of recognizing both lip-sync and blink rate inconsistencies. This paper introduces TrueSync, an innovative deepfake detection application that combines two core features: visual lip-sync analysis and blink rate monitoring. Using advanced machine learning models, including CNN-LSTM for blink rate detection and SyncNet for lip-sync analysis, TrueSync addresses the challenge of detecting deepfakes by analyzing both spatial and temporal features of the video. The CNN-LSTM model tracks the blink rate by evaluating eye states (open or closed), while the SyncNet model assesses the synchronization between lip movements and audio. This hybrid approach improves detection accuracy and provides a scalable solution for identifying manipulated content. The following sections explore

the current landscape of deepfake detection, the design and functionality of the TrueSync application, its integration of the CNN-LSTM and SyncNet models, and the results demonstrating its effectiveness. The paper concludes by evaluating TrueSync's potential to enhance trust and credibility in digital media and its future applicability in combating evolving deepfake technologies.

1.2. Problem Statement

Identifying whether a video is authentic or manipulated has become increasingly difficult. Deepfake technology can produce highly convincing videos with seemingly accurate lip-syncing and eye-blink patterns, making them appear legitimate. This creates significant challenges for individuals trying to distinguish real content from fake.

1.3. Goal

The goal is to develop a user-friendly platform that enables non-experts to upload videos for deepfake detection. By combining two advanced models—visual lip-sync matching and blink rate analysis—into a single, accessible interface, the platform aims to enhance detection accuracy.

1.4. Objectives

- To enhance detection: Improve the accuracy of distinguishing between real and deepfake videos by utilizing lip-sync and blink rate analysis.
- To reduce false positives: Minimize the misidentification of real content, thereby protecting individuals' reputations and avoiding unnecessary alarms over authentic videos.
- To mitigate false negatives: Prevent instances where the system fails to identify a video as fake, ensuring reliable detection.
- To analyze natural blink rate and lip-sync: Examine blink rate and visual lip-sync patterns to derive insights that refine detection techniques
- To create a deepfake detection application: Develop an application that integrates visual lip-sync analysis and blink rate monitoring, providing detection results as a percentage.

2. Literature Review

Deepfake technology, powered by advanced machine learning algorithms, has raised significant concerns about the authenticity of digital media. As the capability to produce highly realistic fake videos grows, the need for robust detection methods becomes increasingly critical. This review focuses on two innovative approaches for deepfake detection: visual lip-sync matching and eye-blink rate detection. Blinking, the rapid opening and closing of the eyelids, is an involuntary action regulated by the pre-motor area of the brainstem. Spontaneous blinking ensures the maintenance of an adequate tear film on the cornea, supporting eye health and visual function. However, blinking serves additional functions beyond corneal protection, as demonstrated by the differences in blinking rates among adults and infants. Blinking patterns also vary based on activities and external factors. For instance, blinking frequency increases during activities like reading aloud or rehearsing visual information but decreases during tasks requiring intense visual focus, such as silent reading.

Blink rates are influenced by multiple factors, including physical well-being, cognitive task complexity, physiological conditions, and an individual's capacity for processing information. By gathering and statistically analyzing this information, it is possible to predict the frequency and variability of eye blinks to a certain degree. In contrast, deepfake videos often show an absence or irregularity in blinking, which serves as a key indicator of manipulation. Deepfake detection also involves analyzing inconsistencies between mouth shape dynamics—visemes—and the phonemes being spoken. This approach differentiates between open and closed eye states while incorporating temporal information to detect anomalies. Evaluations using benchmarks from eye-blinking detection datasets have shown promising results, demonstrating the method's effectiveness in identifying videos generated through Deepfake technology. The integration of visual lip-sync matching and blink rate detection enhances current tools for multimedia authentication. This dual approach contributes to the security and dependability of multimedia content, addressing the growing challenges posed by Deepfake manipulation in the digital era.

The visual lip-sync match detection process includes various components for enhanced clarification, such as phoneme-viseme correlation, which examines the relationship between sound units (phonemes) and lip shapes (visemes) for detection. As our solution is based on machine learning (ML), the research primarily focuses on SyncNet models, with comparisons to alternative models for broader analysis. Additionally, the approach incorporates imitation-based visual lip-sync detection, aiming to achieve a high detection accuracy score. Phoneme recognition involves converting audio into text, which is processed by a model. Cropped audio segments are used to apply a transition model, denoted as ϕstt. The transmitted text tktk and audio xkxk are integrated using forced alignment,

represented by ϕfa, to align the text with spoken audio. Moreover, PaPa represents the phoneme units in the spoken audio, and sPtsPt shows the timing of each phoneme. Finally, the phonemes are filtered and extracted based on a predefined set of languages, as illustrated in the figure 1. This method enables precise timing calculations for every sound in the audio. It also generates an array of phonemes in the International Phonetic Alphabet (IPA) format, along with corresponding timestamps in the audio data. The IPA system further facilitates cross-linguistic comparison, revealing how sounds are perceived across different languages [1]. The lip-syncing process relies on the relationship between phonemes (basic sound units) and visemes (lip shapes). This coupling is instrumental in detecting lip-sync errors in videos and functions across multiple languages, making PhoVis an adaptable system for a variety of linguistic contexts. Phonemes, the smallest units of speech sounds, correspond to visemes, the visual representations of lip shapes during speech. Studies indicate that multiple phonemes can correspond to the same viseme [1]. For instance, the word "pet" contains the phoneme /P/, while “bell” includes the phoneme /B/. Although these phonemes differ, they appear visually similar when spoken because they are mapped to the same viseme. When phoneme alignment is consistent between the original and dubbed audio, it becomes possible to adjust a speaker’s lip movements in the video to match these sounds [1]. Is phoneme recognition possible by listening to the audio and using a model to convert it into text? If so, would a "forced aligner" tool then match the phonemes in the audio with words in the sentence, as shown in Figure 2 [1].

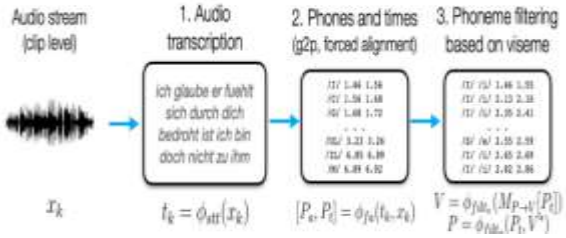

Figure 1. *Algorithm for phoneme extraction from given audio.*

The evaluation of lip-sync accuracy in multimedia applications involves focusing on phonemes that can be matched to the visemes of the target language. This approach is applicable to languages such as English, French, Italian, German, and Spanish. A 2D key-point representation is utilized to depict lip shapes during speech. These coordinates act as invisible points on the lips, capturing their movements in real-time. The process incorporates Spherical Geometric Anchor LIP and Expected Lip, which correspond to vertex shapes on a plane. Only distances calculated along the lip track are measured and matched to key points. This analysis determines which lip movements correspond to each syllable of the audio. These calculated distances, captured frame by frame, are then provided as input to a Machine Learning (ML) model, which trains the system for dyadic alignment of lip movements, enabling precise lip recognition [2].

The primary focus of this research is on the SyncNet model, which is designed to analyze audio and video by encoding short sequences into a shared space. The model reduces the distance between synchronized pairs while keeping unsynchronized pairs far apart. SyncNet finds a global shift across all frames by calculating the loss, improving its accuracy in detecting lip-sync discrepancies. SyncNet has demonstrated significant improvements in detection accuracy, increasing from approximately 76% to nearly 95% on a specific, small-scale dataset, as illustrated in Figure 2. These steps also include modifications to the training approaches used in machine learning models, such as redefining the lip-sync problem as a classification task. Additionally, more complex structures, such as transformers, are employed to evaluate the synchronization between audio and video [3, 4, 5].

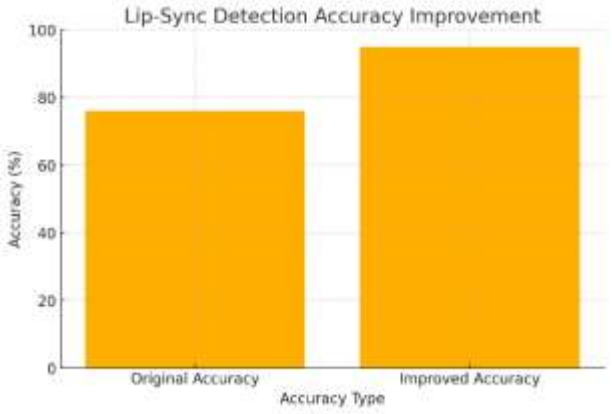

Figure 2. *Lip-sync Detection Accuracy Improvement.*

Novel approaches in this realm focus on understanding the relationship between video and audio through attention mechanisms. Audio-visual streams emphasize extracting and analyzing the information necessary to recognize the synchronization between speech and the visual appearance of lip movements. For instance, dividing a screen into smaller segments can help establish the relationship between sounds and videos. This approach is more complex and is visually represented by violin plots, as shown in Figure 2. These plots are highly effective for displaying error data, revealing that SyncNet’s estimation errors are significantly smaller compared to other systems. This ensures a high level of accuracy across varying

signal sampling rates. An alternative option, involving fixed-size window techniques, also demonstrates SyncNet's ability to handle signals of different lengths without compromising accuracy [5]. Additionally, the MSFC (Mel-Spectrogram Fed ConvNet) is a type of neural network trained to use the visual representation of sound (referred to as an electogram) as input. The GCC-PHAT (Generalized Cross-Correlation with Phase Transform) method is employed for estimating time delays. This technique accounts for the phase of signals, further improving accuracy. The training of these methods utilized three datasets—MTic, Librispeech, and MBeat—for evaluating SyncNet's performance. Results, as shown in Figure 3, demonstrate that SyncNet achieves superior accuracy and robustness compared to other methods [5].

SyncNet's performance is illustrated in Figure 3, showcasing the absolute error for SyncNet, MfCN, and GCC-PHAT across three datasets. A vertically lower distribution in the violin plots signifies superior performance. These results emphasize SyncNet's enhanced accuracy and robustness in estimating lip-sync errors when compared to other methods, as demonstrated by the visual clarity of the violin plots. Alternative methods for deepfake detection include RNN (Recurrent Neural Network) and TCN (Temporal Convolutional Network) models. TCN is particularly efficient for processing long data sequences because it avoids the looping mechanisms inherent in RNNs. This structural difference eliminates challenges such as the 'gradient exploding' problem, allowing smoother learning signals and improved data processing. Both RNN and TCN are effective for sequential data analysis; however, TCN's architecture often gives it a performance advantage over RNN [6]. LSTM (Long Short-Term Memory) models are highly effective for detecting lip-sync disparities due to their memory units' ability to process changing signals and recognize complex relationships in time-series data. They also provide clearer feedback by tracing how different inputs influence the output. This capability enhances the model's ability to learn sequential data, making it easier to extract features from speech. Moreover, LSTMs can merge lip movements across multiple languages other than English, enabling the exploration of interactions between visual lip movements and speech sounds (phonemes) in various linguistic contexts [7-9]. Similarly, GRU (Gated Recurrent Unit) models share functional similarities with RNNs but excel in determining which information to retain or discard and deciding how much past information to forget. Both GRUs and LSTMs are better suited than standard RNNs for handling sequences, as they can recall not only immediate past but also distant information, which is crucial for understanding the context. These models are particularly valuable for lip-reading, where understanding meaning from movements over time is necessary [10,11]. To enhance detection accuracy, the 3DCNN (Deep Convolutional Neural Network) model is applied to extract lip motion features by incorporating word boundary information and improving feature extraction techniques. Additionally, researchers are exploring methods to enhance lip-reading capabilities for previously unseen speakers. This involves the integration of audio-visual cues using CNN (Convolutional Neural Network) models and standard lip shapes, which help reduce variations across different speakers [11, 12]. Wang and Zang developed a system integrating RNN (Recurrent Neural Network) and CNN to implement the CRNN (Convolutional Recurrent Neural Network) model, resulting in an improved performance [12]. The CRNN model combines the spatial feature extraction strengths of CNN with the temporal dependency capabilities of RNN, making it especially effective for sequential data processing. This hybrid model improves the detection of features such as speech frames by sorting the movement of speech signals and reducing background noise [13, 14]. Table 1. shows comparison between each model. The synchronization between spoken audio and lip movements is referred to as lip-syncing. While integrating audio input with lip movement enhances detection accuracy and increases scoring precision, several limitations exist. Limitation-based deepfakes are techniques that often rely on manipulating a person's voice in a sound recording to mimic another person's modified voice. This method is designed to secure the immunity of the original speaker. For example, applying algorithms such as the Efficient Wavelet Mask (EWM) on a recording of a person, can change it. This process involves obtaining two voices: the original speaker's voice and a target speaker's voice. The algorithm modifies the original sound to resemble the target sound. The final output is often so realistic that distinguishing between the original and altered voices becomes challenging, thereby creating a credible privacy boundary that conceals the original speaker's identity [15]. The process of detecting deepfakes often relies on the characteristics of the datasets used to train or evaluate fake algorithms. Researchers examine the distribution and accuracy of training datasets, as deepfake algorithms often fail to capture the full complexity and diversity of real-world data. Additionally, reverse-engineering techniques can be applied to deepfake models to identify artifacts or signatures that reveal their real identity [16]. Preeti et al. [17] conducted a study titled "Methods to Create Deepfakes Using GAN," which utilized a

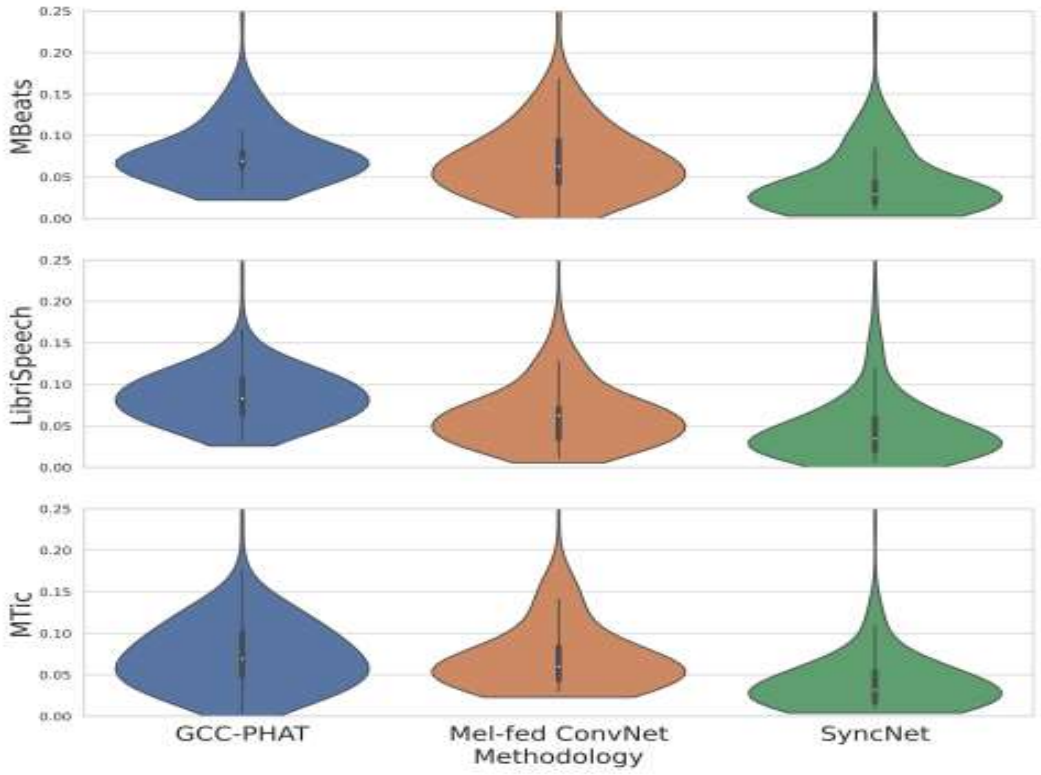

Figure 3. *Violin plots showing the absolute error for SyncNet, MfCN, and GCC-PHAT across all three datasets. A distribution that is lower vertically denotes superior performance.*

Table 1. *Comparison between each model.*

Model	**Key Features**	**Strengths**	**Weaknesses**
SyncNet	-Analyze the video and audio -Uses share spaces encoding -Learns to minimize distance between synchronized pairs	-High accuracy in lip-sync detection up to 95% -Global shift calculation	-Relatively new, and may need further validation in diverse scenarios
RNN	-Recurrent connection to remember past information	-Good for sequential data	-Struggle with long-term dependencies -Faces exploding gradient issues
TCN	-No recurrent connection -Efficient processing of long sequences	-Better gradient flow -Handl long-term patterns quickly	-May require more computational power
LSTM	-Memory call for handling changing data	-Good for complex time series relationships -Provide insight into data influence	-More complex than simpler models
GRU	- Keeps or eliminates information depending on its significant.	-Efficient like LSTM but simpler -Retains information from both the present and past	-Still share some limitations of RNN
CNN	-Capute features of lip movement at the edges of words	-Enhanced feature extraction for lip motion	-Require extensive training data and may be complex to implement
CRNN	-Combines CNN for spatial features and RNN for temporal dependencies	-Improved performance in sequential data tasks -Reduce background noise in speech detection	-Complexity in integrating Both CNN and RNN

combination of Convolutional Neural Networks (CNNs) to generate high-quality results even with small and limited datasets. Human blinking is a fundamental, involuntary action crucial for eye health and supporting visual function.

Blink rates and patterns vary due to factors such as gender, environmental conditions, emotional states, activities, and levels of concentration. For instance, people tend to blink less frequently during tasks that require focus, such as reading or working on a computer, while blink rates increase when they feel tired or anxious. Because blinking is both unique to each person and naturally consistent, it serves as a reliable measure for deepfake detection. By analyzing blink rates, researchers can identify unusual blinking patterns that may indicate manipulated or fake content. Deepfake algorithms often fail to replicate natural blinking rhythms, making blink rate detection a powerful tool for distinguishing authentic videos from AI-generated fakes.

Another crucial aspect in this research is the blink rate detection. Blinking rate/pattern has become a vital factor in the evolution of deepfake detection, emerging as a key indicator of video authenticity. Blinking is a natural, unconscious action that occurs continuously, with frequency variations influenced by factors such as individual's activity level, fatigue, and more. Typically, human blink rates range from 17 to 22 blinks per minute, though these rates can vary due to external factors such as age, gender, and time of the day [18]. In contrast, deepfake videos often exhibit blink anomalies due to the difficulty in replicating natural human behavior. Such anomalies may include extended periods without blinking, abnormally frequent blinking, or inconsistent blink durations, making them critical markers for detection. Deepfake algorithms face significant challenges in mimicking the unconscious variability inherent in human blinks. While Generative Adversarial Networks (GANs) have significantly advanced in simulating facial features, replicating involuntary actions like blinking remains a challenge [18]. Table 2 highlights how human blinks differs depending on in the individual's blinks. Moreover, the differences between natural and deepfake blinking patterns are summarized in Table 3. These findings, based on research [18], highlight the distinct characteristics of human versus algorithmically generated blinking behaviors. The selected Blink Rate Detection model leverages a hybrid architecture combining Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). This hybrid structure integrates the CNN's feature extraction capabilities with

Table 2. *How Human Blinks differs depending on the activities.*

	Male	**Female**
Baseline	9 (12)	20(16)
Tablet	4(6)	8(14)
PC100	5(8)	10(13)
PC330	7(8)	16(13)
Text (pasted over display)	3(5)	10(15)
Text (book position)	4(3)	8(14)
Text (book position)	2(7)	6(10)

Table 3. *Comparison between Normal and Deepfake Eye blinking.*

Feature	**Natural Human Blinking**	**Deepfake Blinking**
Blink Rate (per minute)	17-22	Less than 5 in many cases
Blink Duration	Varies based on activity	Often uniform, consistent across frames
Naturalness	Spontaneous, irregular	Algorithmically generated, predictable
Reaction to Cognitive Load or Fatigue	Changes with mental focus or fatigue	Often lacks such correlation
Periodicity	Irregular	Periodic or completely absent

LSTM's temporal sequence modeling strengths, making it highly effective for detecting blink patterns in video data. In this architecture, CNN processes video frames individually to capture spatial features relevant to blinking. These features include the state of the eye (open or closed), eyelid position, and pixel intensity variations across the eye region. Once these spatial features are extracted, they are passed to the LSTM layer, which tracks blink sequences and timings across frames. This enables the model to identify natural blinking rates and detect anomalies that might indicate deepfake content.

The CNN-LSTM hybrid model is highly effective because it leverages CNN's proficiency in detecting detailed image features and LSTM's ability to track temporal sequences. This combination is critical for distinguishing normal blinking patterns from irregularities often found in deepfake videos [19]. As illustrated in Figure 4, the CNN-LSTM structure processes video data in two stages: CNN layers extract frame-level spatial features, such as the eye's state (open or closed), while LSTM layers analyze these features as a time series. This allows the model to track changes across consecutive frames,

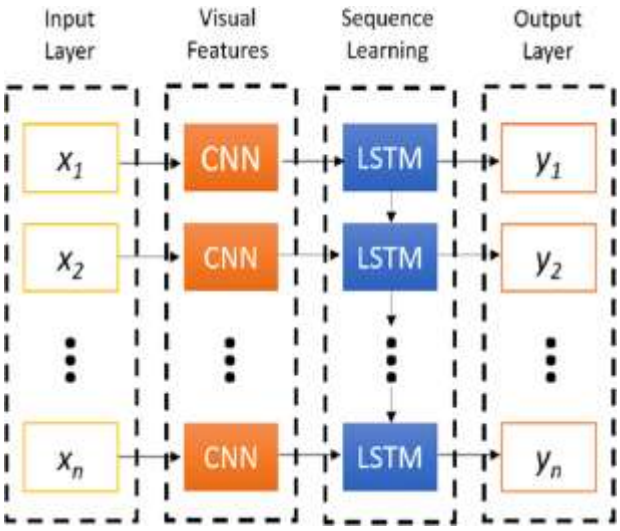

Figure 4. *CNN-LSTM model (Taşdelen, A., & Sen, B. 2021).*

accurately capturing blink frequency and duration. By integrating spatial and temporal dependencies, the CNN-LSTM model effectively monitors blink rates, detecting unusual patterns that may signal synthetic or manipulated content.The modular architecture of the CNN-LSTM model enables it to adapt to various facial dynamics, making it an effective tool in multi-feature deepfake detection tasks. Recent studies support the use of this hybrid model as they highlight its strong performance in detecting blink patterns, and its potential for integration into broader detection frameworks. This adaptability enhances the model's robustness against synthetic content [20].

Deepfake algorithms face significant challenges in accurately replicating human blink behaviour.

Blinking is a complex and involuntary action influenced by factors such as cognitive load, fatigue, and mental health. Research indicates that individuals tend to blink more frequently during conversations and less during high-focus tasks such as reading or complex problem-solving. Such subtle variations are difficult for deepfake algorithms to replicate, often resulting in unnatural blink patterns [18]. The absence or irregularity of blinking frequency in deepfake videos serves as a critical red flag for detection algorithms. For instance, many deepfake videos either omit blinks entirely or simulate them at a rate far below or above normal human behaviour [19].

A hybrid system integrates two or more different features to provide a more effective and efficient solution. These systems are designed to address diverse requirements, such as improving detection capabilities or enhancing performance in fields such as electronics and automobiles. Hybrid systems are widely adopted in industries like automotive engineering to develop more efficient and sustainable solutions.

A hybrid deepfake detection method incorporates detecting both the appearance (facial impressions) and the behavior of an individual. Appearance detection identifies visual inconsistencies, such as mismatched facial features or improper lighting, which may initially appear legitimate to users. Behavioral detection and analysis examines anomalies like facial expression mismatches, abnormal eye movements, or irregular blinking. It also evaluates the synchronization between lip movements and speech, enabling users or investigators to detect behavioral inconsistencies [21].

A hybrid detection method leverages the incorporation of machine learning mechanisms such as Long Short-Term Memory Networks (LSTM) and Multilayer Perceptrons (MLP) to detect deepfake images and videos. These methods have demonstrated high accuracy in distinguishing between real and fake media. The algorithms used for training these models involve datasets containing a large number of real and manipulated faces, publicly available on platforms like Kaggle. This dataset comprises approximately 60,000 images sourced from two different sources: the first is FlickrFaces-HQ, which features real faces, and the second is Deepfake Detection Challenge dataset, which includes deepfake faces generated using Generative Adversarial Networks (GANs). The integration of a Convolutional Neural Network (CNN) and a Multilayer Perceptron (MLP) provides an effective detection layer for accurately identifying deepfake videos. CNNs automatically extract key visual features and patterns from video frames, helping to discern real content from manipulated footage. Concurrently, MLPs focus on analyzing specific facial features such as eyes, nose, mouth, eye blinks, and lip size. Capturing such details is critical, as deepfake videos often fail to replicate these features accurately. Hybrid Optimized Deep Feature Fusion-based Deepfake Detection (HODFF-DD) is a novel technique designed for identifying deepfake videos, notable for integrating the strengths of both Inception and Residual Network architectures [22]. It combines the capabilities of two advanced deep learning models: InceptionResNetV1 and InceptionResNetV2. These models are recognized for their proficiency in analyzing image patterns and distinguishing between authentic and manipulated videos. The technique's effectiveness lies in the integration of strengths from Inception and Residual Network (ResNet) architectures [22]. The Inception component enables the model to analyze both small and large details in each video frame, identifying.

Table 4. *Comparison between each model.*

Hybrid Deepfake Detection Method	**Detection Data**	**Strengths**	**Weaknesses**
Appearance and Behavior Detection	Focuses on detecting visual facial and eye or lip inconsistency	Its ability to analyze and detect both behavior and the physical facial expressions	It has limitations in detection and may fail to identify a fake video if the video is crafted with high accuracy and realism.
LSTM and MLP Detection	It focuses on detecting the facial and pattern behavior inconsistency	Its ability to detect and analyze both images and videos accurately	Performance might be affected due to the video resolution
CNN and MLP Detection	It focuses on detecting eyeblink and lip inconsistency	It detects facial expression with high accuracy	Performance might be affected due to the video resolution and lighting
Reaction to Cognitive Load or Fatigue	It focuses on in detecting image patterns and facial inconsistency	Its ability to detect even with low video resolution	It is expensive to implement
HODFF-DD			

subtle inconsistencies that may signify manipulation. Simultaneously, ResNet incorporates "shortcut" connections that simplify the learning process and enhance the retention of critical features, even in deeper networks. By merging these two approaches, HODFF-DD effectively detects and classifies deepfakes, even under challenging conditions such as variable lighting and appearances [23-28]. Table 4 shows comparison between each modelTrueSync is an application designed for deepfake video detection, focusing on two main features: abnormal blink rates and visual lip-sync mismatches. The application analyzes videos featuring a single individual to determine the video's authenticity. The detection process begins by analyzing the individual's blink rate using a hybrid Convolutional Neural Network with Long Short-Term Memory (CNN-LSTM) architecture. This hybrid architecture is chosen for its proven ability to enhance detection accuracy by effectively capturing both spatial and temporal features. After completing the blink rate analysis, TrueSync transitions to analyzing the individual's lip-sync movements using the SyncNet model. Renowned for its reliability and accuracy, SyncNet delivers robust results in detecting inconsistencies between lip movements and audio. By integrating these two detection methods, TrueSync ensures trustworthy and precise outcomes, making it a reliable tool for deepfake video identification.

The TrueSync application focuses on two core detection features: lip-sync analysis and blink rate monitoring. These approaches are chosen to ensure high detection accuracy and reliable results.

The SyncNet module will be used for lip-syncing analysis. After extracting data from each video frame, features related to phonemes and visemes will be processed and trained to detect inconsistencies. For blink rate monitoring, the process will follow a similar pipeline but will concentrate on blink frequency. It will capture features such as the appearance of open or closed eyes, eyelid position, and pixel intensity variations across the eye region. This data will then be processed using the CNN-LSTM model. The application will analyze lip-sync and blink rate data separately. Both datasets will be trained using CNN-LSTM and SyncNet models. TrueSync will start with blink rate analysis, as facial features typically begin with the eyes, followed by lip-sync evaluation. This sequencing ensures both efficiency and accuracy in detection. To integrate these processes, a Fusion module will combine the outputs from the CNN-LSTM model and SyncNet.

By combining these two models, the application enhances detection capabilities, providing users with high-accuracy results. By proposing this innovative approach, TrueSync addresses the challenges posed by increasingly sophisticated deepfake manipulations. It offers a robust and scalable solution, grounded in this research, to improve the credibility and reliability of digital content. The implementation of this application is guided by Equation 2.4, as illustrated in Figure 4. The process more clearly in equation 1.

$$Lip - SyncAthenticatictyScore(LAS) = (1-\alpha)\cdot\left(1-\frac{AVSD}{T_{max}}\right)+\ \alpha\cdot(1-\frac{|BFD|}{R_{max}})\cdot BTC \quad (1)$$

The first variable, α, represents the weight assigned to the blink factors, typically expressed as either 0 or 1, summarizing the impact of blink behavior. Tmax denotes the maximum deviation (in milliseconds) of the audio-visual (AV) sync, which measures the difference between lip movements and corresponding audio. The Audio-Visual Sync Deviation (AVSD) quantifies how closely the lip movements align with the audio, indicating the precision of synchronization. For blink rate analysis, Rmax represents the maximum natural frequency of blinking in a typical individual, while the Blink

Frequency Deviation (BFD) reflects how much the individual's blink rate deviates from this natural frequency. If a person blinks faster or slower than usual, the BFD will capture this variance. This equation is crucial for the TrueSync application, as it helps accurately assess the degree of synchronization and provides a high detection score based on lip-sync and blink rate data. The visual lip-sync matching based on SyncNet is shown in equation 2.

$$SyncNet = \frac{1}{N}\sum_{i=1}^{N}\sin(V_i, A_I) \quad (2)$$

This equation computes the similarity between lip movements (visual features) and speech (audio features). The SyncNets variable represents the overall score, which is related to the video segmentation. The function sim (Vi, Ai) defines the similarity between visual features (Vi, lip embeddings) and audio features (Ai, audio embeddings) at each frame. N denotes the total number of frames in the video. To clarify further, audio and visual features (specifically from the lip region) are extracted for each frame. SyncNet then computes feature embeddings to measure the similarity between these audio and visual embeddings. Finally, the average of the similarity scores across all frames is used to generate the final score. The blink rate detection is expressed in equation 3.

$$R_{blink} = \frac{1}{T}\sum_{1}^{T}\sigma(f_{LSTM}(E_t)) \quad (3)$$

This equation is used for blink rate detection, where RBlink represents the average blink rate per second, as discussed earlier in the introduction and the blink rate section of this article. T stands for the total duration of blinks, while Et represents eye features, indicating whether the eye is open or closed, as determined by a Convolutional Neural Network (CNN). The Long Short-Term Memory (LSTM) network is responsible for identifying and detecting eye blinks within a specified time frame (σ). The value of σ indicates whether the LSTM output corresponds to a blink event. CNN is essential for detecting eye states (open or closed) and collaborates with the LSTM and temporal modules to identify anomalies that may suggest manipulation, such as deepfake videos or other deceptive content. The TrueSync-B is given by equation 4.

$$\text{TrueSync} - \text{B} = \beta * \text{S}_{\text{SyncNet}} + (1-\beta) * \left(1 - \left|\frac{\text{R}_{\text{blink}} - \text{R}_{\text{expected}}}{\text{R}_{\text{expected}}}\right|\right) \quad (4)$$

This equation integrates two key components: the visual lip-sync score (SsyncNet) and blink rate deviation, to assess both audio-visual synchronization and natural eye-blinking behavior, helping to determine whether the video is authentic or manipulated. The visual lip-sync score, calculated using the SyncNet method, ranges from 0 to 1, where higher scores indicate better alignment between speech and lip movement. For the blink rate, the component compares the observed blink rate (RBlink) to the expected rate (Rexpected), reflecting typical human behavior, which may vary depending on the individual's activity. As with the lip-sync score, a value close to 1 for the blink rate represents a realistic blink pattern. A weighting factor (β) allows the model to prioritize either lip-sync accuracy or blink naturalness, based on the application. By combining these two components, TrueSync-B generates a balanced, normalized score between 0 and 1. Higher scores reflect superior synchronization and realism, making the application suitable for deepfake detection and audiovisual content evaluation.

Instead of creating a single method, multi-modal methods are more resistant when it comes to detecting deepfake.

As there are plenty of models to be integrated, it achieves the concept of scalability.

The high robustness of multi-modal methods can foster the user's trust in detection systems.

3. Material and Methods

The TrueSync application processes user-input video data through advanced techniques to detect deepfake content. The process begins with video preprocessing, where the system utilizes SyncNet and CNN-LSTM to extract and analyze relevant features. The results are then generated as a percentage score, reflecting the possibility of the video being authentic or manipulated. If no abnormalities or deepfake indicators are detected, the system reprocesses the video through an iterative preprocessing stage to ensure accuracy and thorough detection. Figure 5 is visual Lip-sync Matching Flowchart and figure 6 is blink rate Flowchart. TrueSync Flowchart is shown in figure 7. For visual lip-sync detection in the TrueSync application, features related to the alignment of lip movements with corresponding sound are extracted from each video frame. These extracted features are trained and processed using the SyncNet model, which evaluates the results by comparing them to a dataset of normal lip movement behaviours. The evaluation results are then combined with blink rate data to calculate the final score percentage, determining the likelihood of the video being authentic or manipulated. For the blink rate model in the TrueSync application, a hybrid approach combines Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models. The CNN layer applies filters, introducing non-linearity, and reducing spatial dimensions.

Additionally, it converts two-dimensional (2D) image data into a one-dimensional (1D) format, making it suitable for sequential learning and feature classification. The processed output from the CNN serves as input to the LSTM layer, which then analyzes these features. The LSTM layer focuses on identifying and evaluating blink events and their rates, ensuring accurate and reliable detection for distinguishing authentic videos from deepfake content.

In the TrueSync application, visual lip-sync detection begins by extracting features from each video frame that correspond to lip movements and associated audio. This data is processed and trained using the SyncNet model, which evaluates the results by comparing them to a dataset of normal lip movement behaviors. The evaluation results from the SyncNet model are then combined with blink rate data to calculate the final score percentage. This score determines the likelihood of the video being authentic or manipulated, providing a comprehensive analysis by integrating both visual lip-sync and blink rate features.

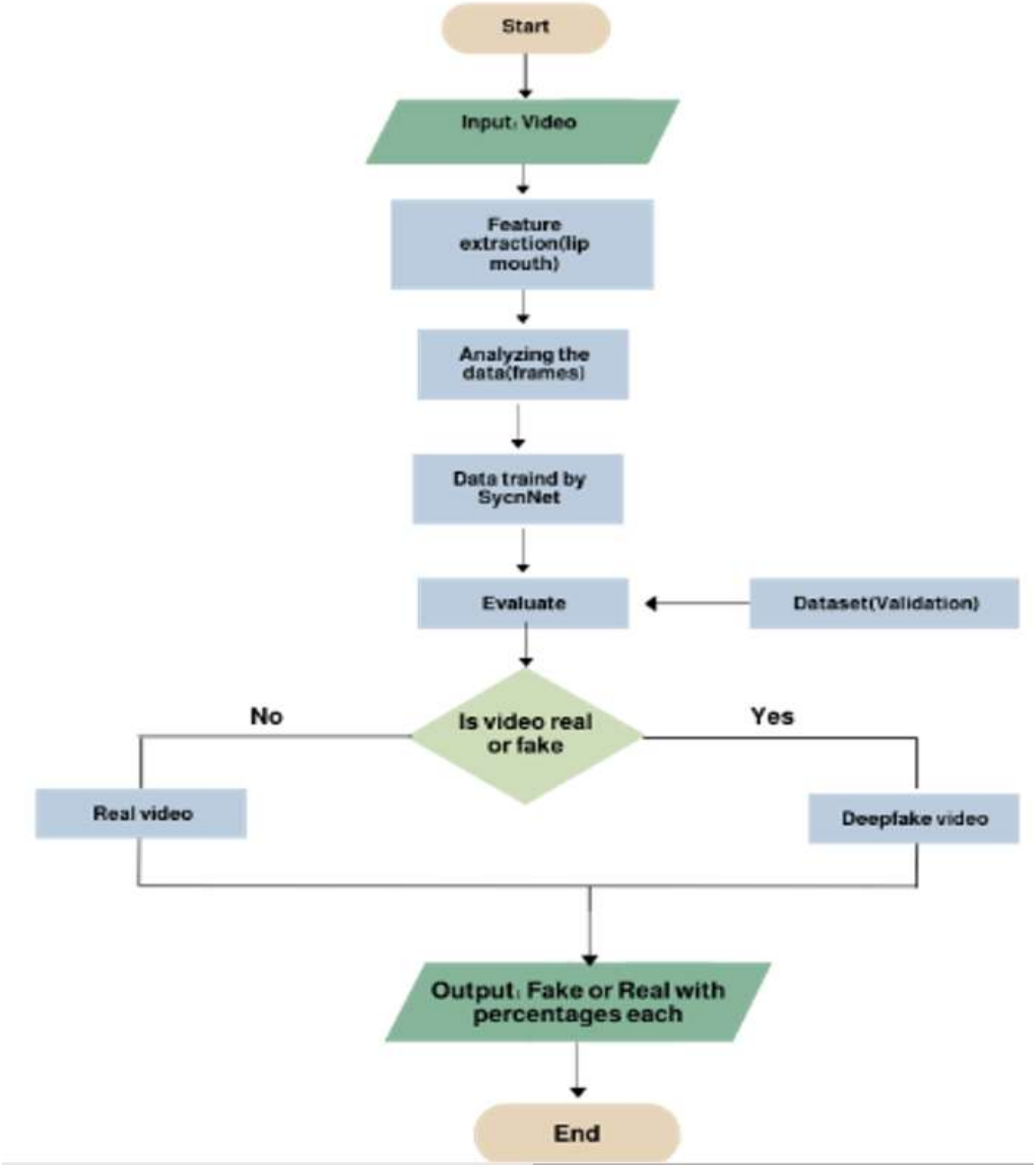

Figure 5. *Visual Lip-sync Matching Flowchart.*

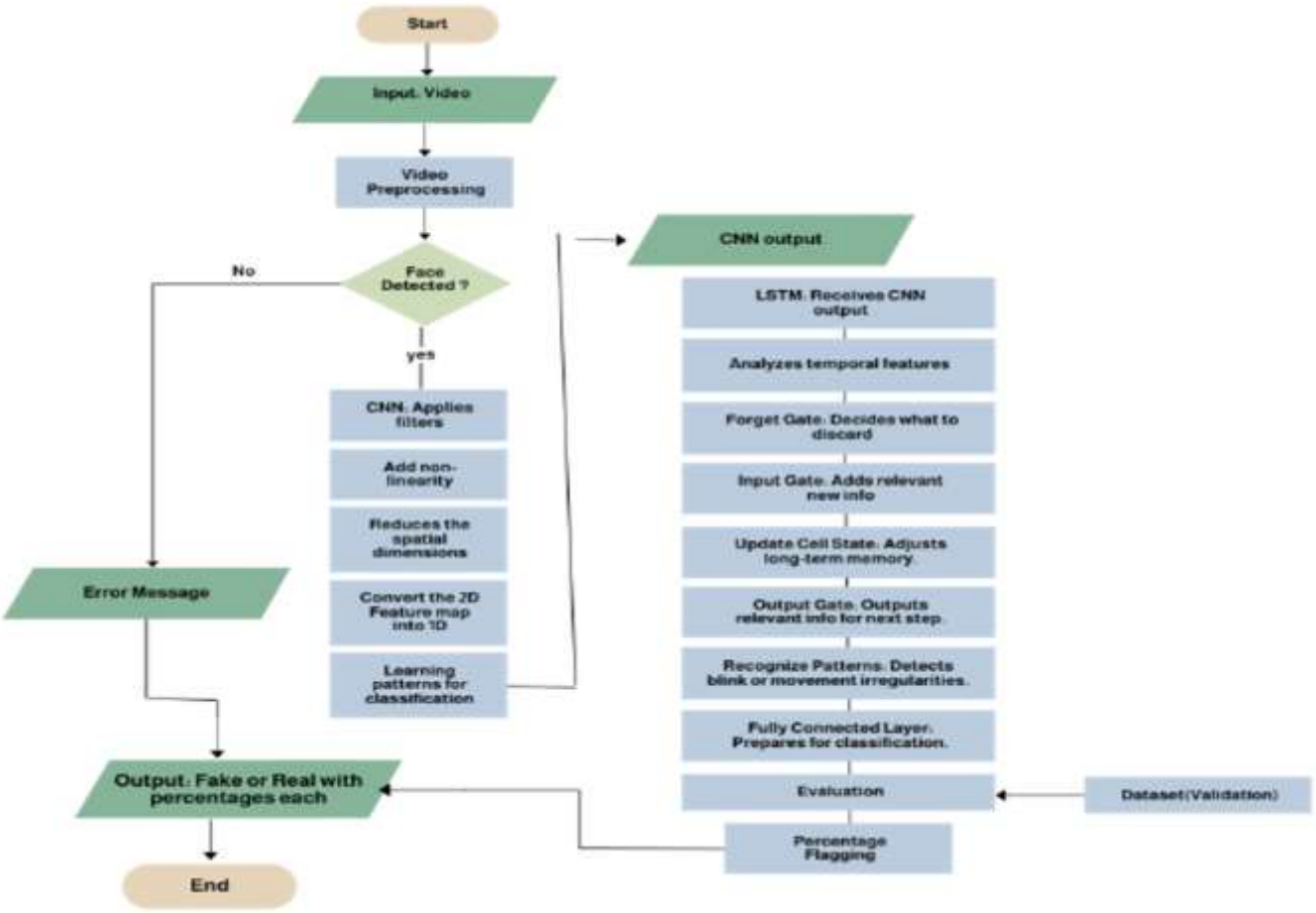

Figure 6. *Blink rate Flowchart.*

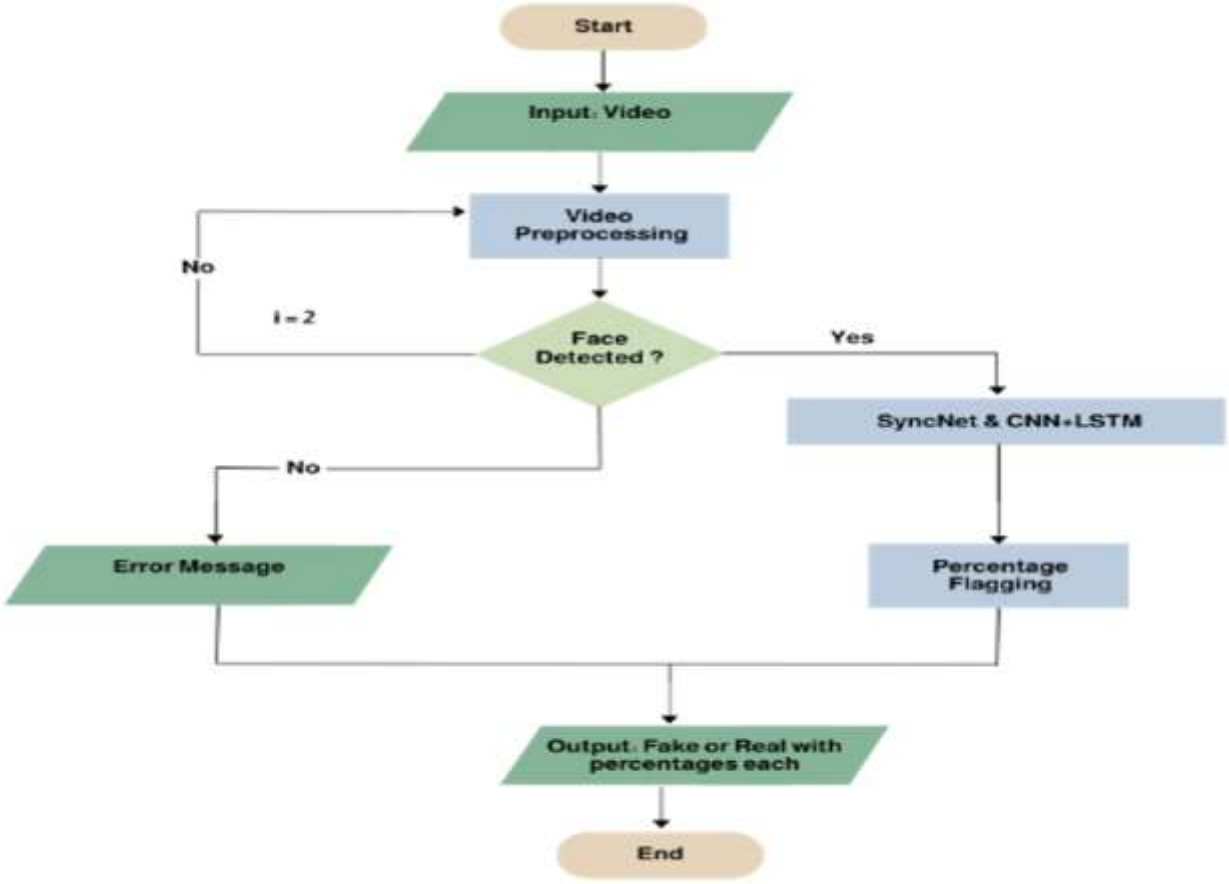

Figure 7. *TrueSync Flowchart.*

4. Results and Discussions

The TrueSync application demonstrates its effectiveness in detecting video manipulation by analyzing two critical features: lip-sync and blink rate.The application provides users with a score percentage, indicating the likelihood of a video being authentic or manipulated. This functionality serves as a valuable mitigation tool, enhancing the credibility of digital content.Given the increasing difficulty in distinguishing between authentic and manipulated videos, TrueSync addresses this challenge by offering a user-friendly platform that allows non-technical users to upload videos for deepfake detection. The application integrates two advanced detection models—visual lip-sync matching and blink rate analysis—within a single, accessible interface. By focusing on these two key features, TrueSync achieves not only high percentage score but also accurate and reliable results, ensuring robust performance and usability for a wide range of users.

5. Conclusions

This project introduces TrueSync, a robust and user-friendly application designed to help non-expert users detect deepfake videos. By integrating multi-modal detection techniques—visual lip-sync matching and blink rate monitoring—TrueSync enhances detection accuracy and resilience against adversarial attacks. This combination ensures scalability, adapting to future advancements in deepfake technology. Ultimately, TrueSync contributes to fostering trust and security in digital media, offering an accessible and reliable tool for combating the growing challenges posed by deepfake content.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1]Gupta, H. (2024). Perceptual synchronization scoring of dubbed content using phoneme-viseme agreement. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 392–402.

[2]Zhou, Y., & Lim, S. N. (2021). Joint audio-visual deepfake detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14800–14809.

[3]Halperin, T., Ephrat, A., & Peleg, S. (2019). Dynamic temporal alignment of speech to lips. ICASSP 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3980–3984. IEEE.

[4]Park, S. J., Kim, M., Choi, J., & Ro, Y. M. (2024, April). Exploring phonetic context-aware lip-sync for talking face generation. ICASSP 2024-2024 IEEE *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4325–4329. IEEE.

[5]Yoon, D., & Cho, H. (2024). Lip and voice synchronization using visual attention. *The Transactions of the Korea Information Processing Society*, 13(4), 166–173.

[6]Guera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. Proceedings of the *15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS*), 1–6.

[7]Li, Y., Chang, M. C., & Lyu, S. (2018). In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. *arXiv preprint arXiv:1806.02877*.

[8]Farid, H. (2020). Creating, identifying, and combating deep fakes. *IEEE International Workshop on Information Forensics and Security (WIFS)*. Retrieved from https://farid.berkeley.edu/downloads/publications/wifs20.pdf

[9]Gholipour, A., Taheri, A., & Mohammadzade, H. (2021). Automated lip-reading robotic system based on convolutional neural network and long short-term memory. In Social Robotics: *13th International Conference, ICSR 2021, Singapore, November 10–13, 2021*, Proceedings 13 (pp. 73–84). Springer International Publishing.

[10]Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms*, 15(5), 155.

[11]Al-Khazraji, S. H., Saleh, H. H., Khalid, A. I., & Mishkhal, I. A. (2023). Impact of deepfake technology on social media: Detection, misinformation, and societal implications. The

Eurasia *Proceedings of Science Technology Engineering and Mathematics*, 23, 429–441.

[12]Mallet, J., Krueger, N., Vanamala, M., & Dave, R. (2023). Hybrid deepfake detection utilizing MLP and LSTM. *arXiv*. Retrieved from https://arxiv.org/pdf/2304.14504

[13]Arshad, S., & Shah, S. C. A. (2024). Hybrid optimized deep feature fusion-based deepfake detection in videos using spotted hyena optimizer. *Computers & Security*, 134, 102848. https://doi.org/10.1016/j.cose.2023.102848

[14]GeeksforGeeks. (2023). Residual networks (ResNet) – *Deep learning*. Retrieved from https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/

[15]Cozzolino, D., Poggi, G., & Verdoliva, L. (2017). Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection. *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, 159–164.

[16]Yao, H., Huang, Z., & Tan, S. (2021). Detection of AI-synthesized video using temporal cues: Eye blinks and mouth movements. *Journal of Information Security Research*, 13(3), 155–167.

[17]Raina, A., & Arora, V. (2022). SyncNet: Using causal convolutions and correlating objective for time delay estimation in audio signals. *arXiv preprint arXiv:2203*.14639.

[18]Jung, T., Kim, S., & Kim, K. (2020). DeepVision: Deepfakes detection using human eye blinking patterns. *IEEE Access*, 8, 83144–83154.

[19]Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. *MIT Press.*

[20]Chollet, F. (2017). Deep learning with Python. *Manning Publications.*

[21]Wang, Y., & Zhang, H. (2024). Hybrid approaches for deepfake detection. *Journal of Multimedia Tools and Applications*, 2(4), 73–90.

[22]Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery*, 14(2), e1520.

[23]Robert, N. R., A. Cecil Donald, & K. Suresh. (2025). Artificial Intelligence Technique Based Effective Disaster Recovery Framework to Provide Longer Time Connectivity in Mobile Ad-hoc Networks. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.713

[24]ZHANG, J. (2025). Artificial intelligence contributes to the creative transformation and innovative development of traditional Chinese culture. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.860

[25]M.K. Sarjas, & G. Velmurugan. (2025). Bibliometric Insight into Artificial Intelligence Application in Investment. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.864

[26]S. Esakkiammal, & K. Kasturi. (2024). Advancing Educational Outcomes with Artificial Intelligence: Challenges, Opportunities, And Future Directions. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.799

[27]Bandla Raghuramaiah, & Suresh Chittineni. (2025). BreastHybridNet: A Hybrid Deep Learning Framework for Breast Cancer Diagnosis Using Mammogram Images. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.812

[28]Almutairi, Z., & Elgibreen, H. (2022). Advanced detection of audio-visual deepfake patterns. *Algorithms*, 15(6), 155–160.