Unmasking the Truth: A Deep Learning Approach to Detecting Deepfake Audio through MFCC Features

Islam Altalahin
Department of Computer Science
Alzaytoonah University of Jordan
Amman, Jordan
202117006@std-zuj.edu.jo

Shadi AlZu'bi
Department of Computer Science
Alzaytoonah University of Jordan
Amman, Jordan
smalzubi@zuj.edu.jo

Assal Alqudah
Department of Computer Science
Alzaytoonah University of Jordan
Amman, Jordan
a.alqudah@zuj.edu.jo

Ala Mughaid
Dept. of Information Technology
Faculty of prince Al-Hussien bin Abdullah for IT.
The Hashemite University
P.O. Box 330127, Zarqa (13133), Jordan.
ala.mughaid@hu.edu.jo

2023 International Conference on Information Technology (ICIT) | 979-8-3503-2006-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIT58056.2023.10226172

***Abstract*—Deepfake content is artificially created or altered using artificial intelligence (AI) methods to appear real. Synthesis can include audio, video, images, and text. Deepfakes may now produce content that looks normal, making it more difficult to identify. Significant progress has been made in identifying video deep fakes in recent years; However, most of the investigations into voice deep fake detection have used the ASVSpoof-2019 dataset and several machine learning and deep learning algorithms. This research uses machine-based and deep-learning approaches to identify fake audio. Melted frequency cepstral coefficients (MFCCs) are used to extract the most useful information from the sound. We choose the 2019 ASVSpoof dataset, which is the latest reference dataset. Experimental results show that Convolutional Neural Networks (CNN): (CNN-LSTM) outperformed other machine learning (ML) models in terms of accuracy, achieving an accuracy of up to 88%.**

***Index Terms*—Deepfake Audio, CNN-LSTM, Melted frequency cepstral coefficients**

I. Introduction

Deepfakes are synthetic media created using artificial intelligence and machine learning algorithms that can create realistic fake images, videos, and audio. Deepfake technology can be used to deceive people, spread misinformation, manipulate public opinion, and impersonate individuals [1]. The potential risks of deepfakes are significant, as they can be used to create false evidence, fake news, or commit financial fraud or other crimes. While deepfake technology is relatively new, efforts are underway to develop ways to detect and prevent its creation and distribution. Researchers and organizations are developing new AI and machine learning techniques to identify deepfakes, creating new algorithms to detect manipulation, and designing new devices that can detect the presence of deepfakes in real time. Deepfake voice, in particular, is a major concern due to its potential to deceive and manipulate individuals, organizations, and entire communities. It is important to remain vigilant and develop new tools and techniques to detect and prevent the spread of deep fakes [1–3]. Where the focus in this research paper was on one of the methods for detecting deep falsification of voices using deep learning techniques to contribute to limiting its spread and limiting the many risks it causes. Several studies have explored the development of deep fakes and the use of deep learning algorithms to detect fake voices.

979-8-3503-2006-0/23/$31.00 ©2023 IEEE

II. ARTIFICIAL INTELLIGENCE AND DEEP FAKE

Artificial Intelligence (AI) has been rapidly advancing in recent years, with applications in various fields ranging from healthcare to finance. However, with the rise of AI technology comes concern about its misuse, particularly in the creation of deep fakes. Deepfakes are synthetic media, such as images, videos, and audio, that use AI algorithms to create realistic fake content. While deep fake technology has positive applications, such as in the entertainment industry, its ability to deceive and manipulate individuals and communities has raised significant concerns. As such, the development of detection and prevention methods is critical to mitigating the potential risks associated with deep fakes Do not number text heads-the template will do that for you [4–6].

A. Techniques used in deep fakes

Deepfakes: Audio deep fakes are created using artificial intelligence to generate manipulated audio content that sounds like it was spoken by a real person. This process involves training a machine learning model on a large dataset of audio recordings of the target individual. The concern with audio deepfakes is that they can be used to spread misinformation or to discredit individuals or organizations, and can be difficult to detect. Researchers are developing new methods to identify audio deepfakes by analyzing the audio signal for signs of manipulation, but as the technology used to create deep fakes continues to evolve, new detection methods will also need to be developed [7–9].

Voice synthesis: Voice synthesis is a deep fake technique that involves using deep learning to analyze a large dataset of recordings to generate synthetic audio that mimics the voice of a specific individual. The process involves training a deep learning model to identify patterns in the speaker's voice and generate new audio content that sounds like it was spoken by them, even if the words or phrases were never actually spoken. The primary application of voice synthesis in deep fake technology is the creation of fake audio recordings for spreading misinformation or discrediting individuals or organizations. Researchers are developing methods for detecting and identifying synthetic audio by analyzing the audio signal for signs of manipulation, but as the technology continues to evolve, new detection methods will also need to be developed [10–13].

Audio splicing: Audio splicing in deepfake technology involves taking audio samples of a target individual and manipulating them using digital audio editing software to create a new audio track that appears to be authentic. This technique is commonly used to create synthetic audio that is intended to deceive listeners into thinking it was produced by a specific individual. One of the primary applications of audio splicing in deep fake technology is in creating fake audio recordings to spread misinformation or discredit individuals or organizations. Researchers are developing new methods for detecting and identifying manipulated audio to address the potential harms of audio splicing deep fakes. These methods typically involve analyzing the audio signal for signs of splicing or other forms of manipulation, such as changes in background noise or inconsistencies in the timing or frequency of the speech. However, as deepfake technology evolves, new detection methods will likely need to be developed [14–17].

Video manipulation: Video manipulation in deep fake involves using AI and machine learning algorithms to create realistic but fake videos of individuals or events that did not actually occur. This can include techniques such as face swapping, lip-syncing, background replacement, and object removal. The use of video manipulation in Deepfake has raised concerns about its potential to spread false information and damage reputation. To address these concerns, researchers are developing methods to detect and identify deepfake videos by analyzing various features of the video, such as facial expressions, eye movements, and speech patterns. It is important to develop ways to prevent the spread of deepfakes, such as improving digital forensics techniques and increasing public awareness of the dangers of fake media [18–20].

B. Fake voice detection methods

Spectrogram Analysis: Spectrum analysis is a technique commonly used to detect fake sounds in audio recordings by analyzing the frequency and time information of an audio signal. Several studies have shown that spectrum analysis, combined with deep learning algorithms, is effective in detecting deepfake audio with high accuracy. Commercial tools such as Google's Project DDetect also use spectrogram analysis and machine learning algorithms to detect deepfake audio in real-time. While spectrum analysis is a powerful tool, it is important to note that deepfake technology is evolving rapidly, and new techniques are being developed to create more realistic fake sounds. Therefore, researchers and organizations must continue to develop and improve detection methods to keep up with these developments [4, 21, 22].

Check the speaker: Speaker verification is a process used to authenticate the identity of the person speaking in an audio recording, aiming to prevent the use of fake or manipulated recordings. Methods used for speaker verification rely on phonetic and linguistic features, with machine learning algorithms and deep neural networks being commonly used. Studies have shown that a deep learning model that uses phonetic and linguistic features can achieve high accuracy in identifying fake recordings. Additionally, a combination of Mel-Frequency Cybstral Coefficients (MFCCs) and Supporting Vector Machines (SVMs) has been proven effective in distinguishing between real and fake speech. Biometric features such as voiceprints and facial recognition have also been explored, with researchers proposing a multimodal approach that combines voice and facial recognition to verify the identity of speakers in video recordings [23–25].

Statistical analysis: Gaussian mixture model (GMM) is a statistical analysis technique used to detect fake audio by modeling the distribution of speech features for multiple speakers. The GMM is trained on a dataset of speech samples from known speakers to estimate the mixture parameters of Gaussian distributions that best represent the features of each speaker. GMM is effective in detecting artificial and deep speech generated by neural networks and text-to-speech systems. However, the accuracy of GMM-based speaker verification depends on the quality of the training data and the complexity of the speech features [3, 26–29].

Synthetic speech detection: The synthetic speech detection method is used to identify artificially generated voices from text-to-speech systems and other speech synthesis software. It involves phonemic and linguistic analysis, machine learning, and deep learning to identify patterns that characterize artificial speech. Machine and deep learning models are trained on a dataset of speech samples containing both real human and synthetic speech to identify features that discriminate between the two types of speech. Phonetic and linguistic analysis involves analyzing acoustic and linguistic features of speech to detect patterns. The synthetic speech detection method is important for detecting fake voice and ensuring audio recording authenticity. By combining different techniques, high-accuracy methods can be developed to distinguish artificial speech from real human speech [30, 31].

In general, the artificial speech detection method is an important tool for detecting fake voices and ensuring the authenticity of audio recordings. Using a combination of phonetic and linguistic analysis, machine learning, and deep learning techniques, it is possible to develop high-accuracy methods for detecting artificial speech and distinguishing it from real human speech [32, 33]. These technologies can be used together for fake voice detection and deep voice recognition. However, it is important to note that deep voice detection can be challenging, and new techniques are constantly being developed to create more realistic deepfakes. As such, researchers and organizations continue to work on developing new and improved methods for detecting fake votes

III. METHODOLOGY

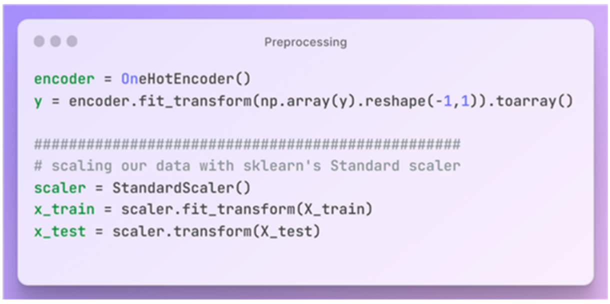
A. Dataset

ASVspoof Challenge 2019 is the third edition of the Autospeaker Verification Challenge and Counter-Challenge, which aims to advance progress in Autospeaker Verification through standardized dataset distribution and competitive evaluations, The ASVspoof 2019 database is divided into two logical access (LA) scenarios and physical access (PA) scenarios, both of which are derived from the VCTK ruleset. The sections are also divided into data sets for Training, Development and Evaluation, which are separate in terms of speakers and enrollment requirements. The training and development group contains known impersonation attacks, while the evaluation group also includes unknown attacks.

B. Preprocessing

Two pre-processing methods have been applied to the data: OneHotEncoding and Standard Scaling. After applying these pre-processing techniques, the data is now in a format suitable for training and evaluating machine learning models.

Shown in Figure 1



```
encoder = OneHotEncoder()
y = encoder.fit_transform(np.array(y).reshape(-1,1)).toarray()

#################################################
# scaling our data with sklearn's Standard scaler
scaler = StandardScaler()
x_train = scaler.fit_transform(X_train)
x_test = scaler.transform(X_test)
```

Fig. 1. Data preprocessing methods: OneHotEncoding and Standard Scaling

C. Features Extraction & Selection

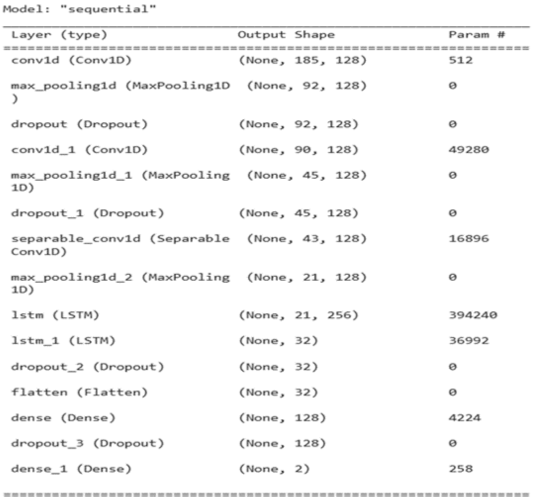
Melt Frequency Cybstral coefficients (MFCC) were used: MFCC features are widely used in speech and audio processing, as they provide a compact representation of the spectral shape of the audio signal. It is derived from the miliscale, a perceptually motivated frequency measure that approximates the response of the human auditory system to sound. MFCCs can capture the temporal properties of sound, which can be useful in detecting fake sounds, as different sound generation methods may have distinct sound signatures. The mean value of the MFCC feature (with 45 coefficients) is added to the result array. Feature selection can help generate a representative and dense subset of the raw data by selecting an appropriate subset of the features. Advantages of feature selection include reducing the amount of data required, requiring less storage, increasing the accuracy (PREC) of predictions, preventing overfitting, and speeding up training and implementation by simplifying variables. Selection of features, in general, is a critical tool to increase effectiveness so many features can be used such as: Zero Crossing Rate (ZCR), Root Mean Square (RMS).

D. Proposed techniques (CNN-LSTM)

The CNN-LSTM architecture is a combination of convolutional neural networks (CNNs) and long-term memory networks (LSTMs), which is effective for processing and learning patterns from sequential data such as speech and audio. CNN layers extract local patterns while LSTM layers learn about temporal dependencies enabling the model to capture both local patterns and long-term dependencies. This hybrid model has been successful in applications such as speech recognition, audio event detection, and video classification due to its ability to capture spatial and temporal information from the data.

IV. EXPERIMENTAL RESULTS

The combination of CNNs and LSTM is a popular architecture for various sequence learning tasks, including speech and audio processing. This hybrid model leverages the strengths of both CNNs and LSTMs to process and learn patterns from sequential data. The CNN layers extract local patterns from the data, while the LSTM layers learn the temporal dependencies in the extracted features. This combination enables the model to capture both local patterns and long-term dependencies in the sequential data, making it suitable for various sequence learning tasks, including speech and audio processing. Figure 2 illustrates a general structure for CNN. Performance metrics are essential for evaluating the



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 185, 128)          512

 max_pooling1d (MaxPooling1D  (None, 92, 128)          0
 )

 dropout (Dropout)           (None, 92, 128)           0

 conv1d_1 (Conv1D)           (None, 90, 128)           49280

 max_pooling1d_1 (MaxPooling  (None, 45, 128)          0
 1D)

 dropout_1 (Dropout)         (None, 45, 128)           0

 separable_conv1d (Separable  (None, 43, 128)          16896
 Conv1D)

 max_pooling1d_2 (MaxPooling  (None, 21, 128)          0
 1D)

 lstm (LSTM)                 (None, 21, 256)           394240

 lstm_1 (LSTM)               (None, 32)                36992

 dropout_2 (Dropout)         (None, 32)                0

 flatten (Flatten)           (None, 32)                0

 dense (Dense)               (None, 128)               4224

 dropout_3 (Dropout)         (None, 128)               0

 dense_1 (Dense)             (None, 2)                 258

=================================================================
```

Fig. 2. Sequential CNN-LSTM

performance of a classification model. They help in understanding how well is the model and identify improvement areas. Here's a detailed introduction to the performance metrics you've mentioned.

A. Confusion Matrix

A confusion matrix is a table that describes the performance of a classification model by comparing its predicted outcomes with the actual outcomes [21]. For a binary classification problem, the confusion matrix has four elements:

- True Positives (TP): The number of positive instances correctly classified as positive.
- True Negatives (TN): The number of negative instances correctly classified as negative.
- False Positives (FP): The number of negative instances incorrectly classified as positive.
- False Negatives (FN): The number of positive instances incorrectly classified as negative.

B. Accuracy

Accuracy is the most straightforward metric for classification. It is the ratio of correctly classified instances to the total number of instances [21]. Accuracy = (TP + TN) / (TP + TN + FP + FN)

C. Precision

Precision, also known as positive predictive value, is the ratio of true positive instances to the total number of instances predicted as positive [21]. Precision = TP / (TP + FP) A high precision indicates that the model has a low false positive rate

D. Recall (Sensitivity or True Positive Rate)

Recall is the ratio of true positive instances to the total number of actual positive instances [21]. Recall = TP / (TP + FN) A high recall indicates that the model has a low false negative rate

E. F1-Score

F1-score is the harmonic mean of precision and recall. It balances both precision and recall, providing a single value that represents the model's performance [21]. F1-score = 2 * (Precision * Recall) / (Precision + Recall) F1-score ranges from 0 to 1, where 1 indicates a perfect classifier, and 0 indicates the worst possible classifier Performance measures.: Where Figure 3 shows a number of performance measures that have been used, including Accuracy, Confusion Matrix, Accuracy, Precision, Recall, and F1-Score. This paper introduced the CNN-LSTM model. As shown in Table I, the figure represents the analysis of the results of the experiment to determine the accuracy of the data and evaluate the strengths and weaknesses of this algorithm.

TABLE I
SUMMARY OF SURVEYED DEEP FAKE DETECTION METHODS STUDIES

	Precision	Recall	F1-score	Support
0	0.83	0.96	0.89	313
1	0.95	0.79	0.86	287
Accuracy	-	-	0.88	600
Macro avg	0.89	0.87	0.87	600
Weighted avg	0.89	0.88	0.88	600

This study differed from prior efforts by combining three feature selection techniques with three classification systems. Promising results were reached, as these results are presented in Table I

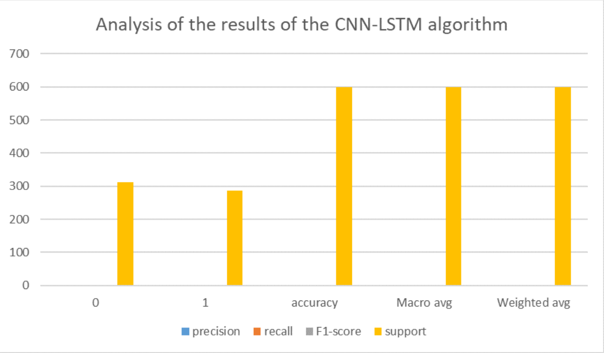


Fig. 3. The curve represents the analysis of the results of the experiment

In summary, the results showed that the methodology used in the study was effective in detecting fakes using the CNN-LSTM deep learning algorithm with an accuracy of 88%. Therefore, the CNN-LSTM model will be one of the good choices for deepfake detection

V. CONCLUSION AND FUTURE WORK

Deep fake account detection is crucial in cybersecurity to combat evolving threats. Detecting AI-generated personas helps prevent misinformation, social engineering attacks, and identity fraud. By identifying and countering these deceptive accounts, organizations can protect their reputation, user data, and maintain the trust of their online communities [8, 9, 12, 13]. We proved that AI is effective in detecting deepfakes voices. We used CNN-LSTM classification to classify sounds as real or fake based on the features of the sound extracted by the MFCC feature. Our results showed that the CNN-LSTM algorithm performed well, achieving an accuracy of approximately 88%. This indicates that the CNN-LSTM classifier is a suitable method for this task and maybe a better choice for detecting fake votes. Overall, our study highlights the potential of deep learning algorithms for deepfakes detection, which could have important implications for social media platforms, individual privacy, and online fraud prevention. More research is needed to explore techniques and approaches to help reduce forms of deepfakes and to further improve the detection of deepfakes.

REFERENCES

[1] S. AlZu'bi, R. Abu Zitar, B. Hawashin, S. Abu Shanab, A. Zraiqat, A. Mughaid, K. H. Almotairi, and L. Abualigah. A novel deep learning technique for detecting emotional impact in online education. *Electronics*, 11(18):2964, 2022.

[2] A. Mughaid, S. Al-Zu'bi, A. Al Arjan, R. Al-Amrat, R. Alajmi, R. Abu Zitar, and L. Abualigah. An intelligent cybersecurity system for detecting fake news in social media websites. *Soft Computing*, 26(12):5577–5591, 2022.

[3] A. Mughaid, S. AlZu'bi, A. Hnaif, S. Taamneh, A. Alnajjar, and E. AbuElsoud. An intelligent cyber security phishing detection system using deep learning techniques. *Cluster Computing*, 25(6):3819–3828, 2022.

[4] A. Mughaid, I. Obaidat, A. Aljammal, S. AlZu'bi, F. Quiam, D. Laila, A. Al-zou'bi, and L. Abualigah. Simulation and analysis performance of ad-hoc routing protocols under ddos attack and proposed solution. *International Journal of Data and Network Science*, 7(2):757–764, 2023.

[5] A. Mughaid, I. Obeidat, S. AlZu'bi, E. AbuElsoud, A. Alnajjar, A. R. Alsoud, and L. Abualigah. A novel machine learning and face recognition technique for fake accounts detection system on cyber social networks. *Multimedia Tools and Applications*, pages 1–26, 2023.

[6] A. Mughaid, S. AlZu'bi, A. Alnajjar, E. AbuElsoud, S. El Salhi, B. Igried, and L. Abualigah. Improved dropping attacks detecting system in 5g networks using machine learning and deep learning approaches. *Multimedia Tools and Applications*, 82(9):13973–13995, 2023.

[7] S. AlZu'Bi, S. A. Abushanap, I. AlTalahin, A. M. Abdalla, and A. A. Tamimi. Secure transmission of noisy images over fiber optic communication. In *2022 9th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, pages 1–5, 2022.

[8] Ala Mughaid, Ibrahim Obeidat, Shadi AlZu'bi, Esraa Abu Elsoud, Asma Alnajjar, Anas Ratib Alsoud, and Laith Abualigah. A novel machine learning and face recognition technique for fake accounts detection system on cyber social networks. *Multimedia Tools and Applications*, pages 1–26, 2023.

[9] Samar Hendawi, Shadi AlZu'bi, Ala Mughaid, and Nayef Alqahtani. Ensuring cybersecurity while leveraging social media as a data source for internet of things applications. In *International Conference on Advances in Computing Research*, pages 587–604. Springer Nature Switzerland Cham, 2023.

[10] J. Kong, J. Kim, and J. Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033, 2020.

[11] Z. Mu, X. Yang, and Y. Dong. Review of end-to-end speech synthesis technology based on deep learning. *arXiv preprint arXiv:2104.09995*, 2021.

[12] Ala Mughaid, Ali Alqahtani, Shadi AlZu'bi, Ibrahim Obaidat, Rabee Alqura'n, Mahmoud AlJamal, and Raid AL-Marayah. Utilizing machine learning algorithms for effectively detection iot ddos attacks. In *International Conference on Advances in Computing Research*, pages 617–629. Springer Nature Switzerland Cham, 2023.

[13] Ala Mughaid, Shadi AlZu'bi, Asma Alnajjar,

Esraa AbuElsoud, Subhieh El Salhi, Bashar Igried, and Laith Abualigah. Correction to: Improved dropping attacks detecting system in 5g networks using machine learning and deep learning approaches. *Multimedia Tools and Applications*, 82(9):13997–13998, 2023.

[14] T. W. Jing and R. K. Murugesan. Protecting data privacy and prevent fake news and deepfakes in social media via blockchain technology. In *Advances in Cyber Security: Second International Conference, ACeS 2020*, volume 2, pages 674–684. Springer Singapore, 2021.

[15] M. Tulio Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv e-prints*, 2020.

[16] A. A. Khan, O. Chaudhari, and R. Chandra. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. *arXiv preprint arXiv:2304.02858*, 2023.

[17] M. S. Rana and A. H. Sung. Deepfakestack: A deep ensemble-based learning technique for deepfake detection. In *2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom)*, pages 70–75. IEEE, 2020.

[18] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol. Deepfake audio detection via mfcc features using machine learning. *IEEE Access*, 10:134018–134028, 2022.

[19] M. Sharma and M. Kaur. A review of deepfake technology: an emerging ai threat. In *Soft Computing for Security Applications: Proceedings of ICSCS 2021*, pages 605–619, 2022.

[20] D. A. Sultan and L. M. Ibrahim. A comprehensive survey on deepfake detection techniques. *International Journal of Intelligent Systems and Applications in Engineering*, 10(3s):189–202, 2022.

[21] K. W. Cheng, H. M. Chow, S. Y. Li, T. W. Tsang, H. L. B. Ng, C. H. Hui, and S. W. Tsang. Spectrogram-based classification on vehicles with modified loud exhausts via convolutional neural networks. *Applied Acoustics*, 205:109254, 2023.

[22] C. Amadeus, I. Syafalni, N. Sutisna, and T. Adiono. Digit-number speech-recognition using spectrogram-based convolutional neural network. In *2022 International Symposium on Electronics and Smart Devices (ISESD)*, pages 1–6. IEEE, November 2022.

[23] C. B. Tan, M. H. A. Hijazi, and P. N. E. Nohuddin. A comparison of different support vector machine kernels for artificial speech detection. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 21(1):97–103, 2023.

[24] R. Tao, R. K. Das, and H. Li. Audio-visual speaker recognition with a cross-modal discriminative network. *arXiv preprint arXiv:2008.03894*, 2020.

[25] S. M. Shah, M. Moinuddin, and R. A. Khan. A robust approach for speaker identification using dialect information. *Applied Computational Intelligence and Soft Computing*, 2022.

[26] K. M. A. Alheeti, S. S. Al-Rawi, H. A. Khalaf, and D. Al Dosary. Image feature detectors for deepfake image detection using transfer learning. In *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*, pages 499–502. IEEE, December 2021.

[27] M. Musaev, M. Abdullaeva, and M. Ochilov. Advanced feature extraction method for speaker identification using a classification algorithm. In *AIP Conference Proceedings*, volume 2656, page 020022. AIP Publishing LLC, 2022.

[28] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, and K. A. Lee. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *arXiv preprint arXiv:2210.02437*, 2022.

[29] R. M. Hanifa, K. Isa, and S. Mohamad. A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, 90:107005, 2021.

[30] R. A. M. Reimao. Synthetic speech detection using deep neural networks. *arXiv preprint arXiv:1903.08750*, 2019.

[31] Z. Khanjani, G. Watson, and V. P. Janeja. How deep are the fakes? focusing on au-

dio deepfake: A survey. *arXiv preprint arXiv:2111.14203*, 2021.

[32] J. Li, X. Chen, Y. Taigman, and L. Wolf. Deep voice: Real-time neural text-to-speech. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3102–3111, 2018.

[33] S. AlZu'bi, A. Mughaid, F. Quiam, and S. Hendawi. Exploring the capabilities and limitations of chatgpt and alternative big language models. *Artificial Intelligence and Applications*, 2023.