A Novel Audio-Video Multimodal Deep Learning Model for Improved Deepfake Detection To Combat Disinformation

Louis Huang[1] **& Emma Huang**[1]

Received July 01, 2025
Accepted September 21, 2025
Electronic access November 09, 2025

In 2025, approximately 8 million deepfakes are circulated online, doubling every six months. Deepfakes are computer-generated media that imitate a person's appearance or voice and are increasingly used for disinformation. Studies have found that humans struggle to detect deepfakes, while deep learning has shown greater promise. Most existing models use only a singular modality, limiting generalizability, and the few multimodal approaches rely on generic features that are not optimized for deepfake detection. This study develops and evaluates an audio-video multimodal deepfake detection model. A multimodal deepfake dataset of over 20,000 video files and additional emotion datasets were used to train three model variants: general, deepfake, and emotion. Audio and video embeddings from these models were input into 13 downstream machine learning algorithms, creating over 3,600 models and iterating over 77.7 million files. The top model, using deepfake embeddings, achieved an accuracy of 99.53% and an AUC of 99.96%, outperforming state-of-the-art unimodal methods, which reached 89.73% accuracy and 91.70% AUC, and multimodal methods, which achieved 80% AUC. Cost analysis indicated this approach could process the 95 million videos posted on Instagram daily for a one-time infrastructure cost of $7,200 and about $5,000 per year in electricity. Face feature importance analysis found that the eyes and nose were most critical for detection by both humans and computers, providing insight into future efficiency improvements.

Keywords: Audio Deepfake Detection, Video Deepfake Detection, Multimodal Deepfake Detection, Human Deepfake Detection, Transformer Models, dlib Landmark Detection, Feature Extraction

Introduction

In 2025, an estimated 8 million deepfakes are shared online, with this number doubling every six months[1]. Deepfakes are digitally altered videos, images, or audio of people to make them seem like they said or did something that they never did (Fig. 1). They are a rapidly improving tool for disinformation, a form of misinformation with a malicious intent. Video and image deepfakes often involve swapping the face of a preexisting video with that of the target. In contrast, audio deepfakes are generated either from scratch using text-to-speech (TTS) or through voice cloning of the target's speech using real speech samples. Deepfakes pose serious challenges with far-reaching consequences. They can convincingly fabricate evidence, potentially undermining the integrity of legal proceedings. In addition, deepfakes can impersonate celebrities, family members, or friends to commit financial fraud, with industry estimates (Deloitte) projecting economic losses in the United States exceeding $40 billion by 2027[2]. Moreover, they can also be weaponized to influence political elections and international affairs, with malicious deepfakes targeting world leaders like Ukrainian President Volodymyr Zelenskyy[3], former United States President Joe Biden[4], and United Kingdom Prime Minister Keir Starmer[5]. On a broader scale, deepfakes erode trust in digital information and challenge the authenticity of media across domains. Therefore, addressing this growing challenge is essential to preserving public trust, democratic processes, and societal stability.

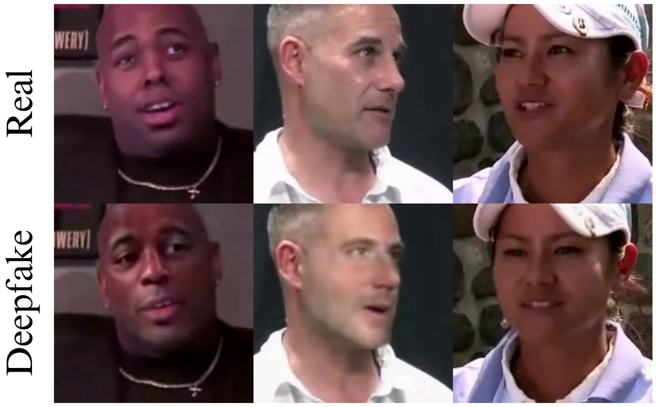

Fig. 1 Deepfakes are often challenging to distinguish with the naked eye.

[1] *H. H. Dow High School*

Deepfake Detection Strategies

Numerous studies have demonstrated that humans struggle to reliably detect deepfakes. Kobis et al. (2021) found that even with awareness, people could not consistently detect video deepfakes, with an average accuracy of only 57%[6]. Furthermore, Mai et al. (2023) made similar conclusions for audio deepfakes, with an average accuracy of only 73% and no significant improvement after a period of learning[7]. These results may be due to a seeing-is-believing heuristic, where people assume what they see is real until proven otherwise, making human deepfake detection difficult to improve[6]. To address these limitations, researchers have turned to machine learning and deep learning techniques for automated deepfake detection. Traditional machine learning models have shown moderate success in deepfake detection. For instance, Singh and Singh (2021) employed Random Forest (RF) and Support Vector Machine (SVM) to achieve an accuracy up to 79% for audio deepfake detection[8]. Moreover, Bonomi et al. (2021) used SVM to attain an accuracy of 86.19% for detecting video deepfakes generated using multiple different techniques[9]. However, such approaches often rely heavily on handcrafted features, which can limit generalization to unseen manipulation methods and hinder adaptability to the rapidly evolving landscape of deepfake generation[9].

By contrast, deep learning (DL) models based on neural networks can automatically learn high-level, manipulation-invariant representations directly from raw data by optimizing millions of parameters through end-to-end training. Unlike traditional machine learning, DL architectures such as convolutional layers (for spatial locality and weight sharing), recurrent units or self-attention mechanisms (for sequential and cross-modal dependencies), and deep nonlinear compositions can approximate highly complex decision boundaries in high-dimensional spaces. This capacity allows DL models to capture subtle, non-linear patterns and cross-feature interactions that conventional algorithms with shallow architectures and limited feature transformations cannot. Guera and Delp (2018) developed a hybrid architecture combining a convolutional neural network (CNN) and recurrent neural network (RNN) to classify video deepfakes with an accuracy of over 97%[10], while Wang et al. (2020) created a deep neural network (DNN) model for audio deepfake detection that had an accuracy of 98.1%[11]. Deep learning models perform exceptionally well, as Rana et al. (2022) found that deep-learning models outperformed both traditional machine-learning models and humans in image and video deepfake detection, achieving a mean accuracy of 89.73% across 50 studies and area under the ROC curve (AUC) of 0.917 across 37 studies versus 86.86% accuracy and 0.909 AUC for traditional machine learning[12].

However, most existing deep-learning approaches use only video or audio data, lacking generalizability and the ability to detect both audio deepfakes and video deepfakes within a single framework. The integration of multimodal data, specifically combining video and audio information to detect deepfakes, remains relatively unexplored. Yet, it is an essential step toward building more robust deepfake detection systems. Despite the strong performance of unimodal deep learning models, their effectiveness can be limited in real-world scenarios. Video-only models may fail to detect audio manipulations, such as synthesized or dubbed speech, while audio-only models cannot identify visual forgeries like face swaps or expression alterations. Deepfake detection methods can be hindered by degradations such as compression artifacts or intentionally added noise, which are sometimes introduced to evade detection[13]. Such effects can be particularly problematic for unimodal approaches, where performance depends heavily on the integrity of a single modality. By integrating both audio and video modalities, multimodal systems have the potential to leverage complementary information, improving robustness and enabling simultaneous detection of a wider range of manipulations[14]. This multimodal fusion can help address the shortcomings of unimodal detectors and enhance overall detection accuracy in diverse, realistic environments.

In terms of existing multimodal deepfake detection models, Mittal et al. (2020) used a deep learning network to compare the emotional similarity between the audio and visual content of a given video, achieving an AUC of 84.4%, but this method is less effective when both modalities are deepfaked[14]. Salvi et al. (2023) used pretrained deep learning models to extract audio and video features that were then sent into a classifier, reaching an AUC of 80% on the FakeAVCeleb dataset[15]. The downside to this approach is that generic features do not necessarily contain useful information for deepfake detection.

Transformers

Originally developed for machine translation, transformer-based deep learning models have recently gained traction in the field of deepfake detection. These models rely on a self-attention mechanism to capture long-range dependencies within data, an essential capability when analyzing complex patterns in both video and audio deepfakes[16]. Transformers can take advantage of large amounts of training data, a key requirement for achieving high accuracy in deepfake detection. One of the most popular transformers for image processing is the Vision Transformer, which converts an image into fixed-size patches before inputting them into the transformer encoder[17] (Fig. 2). For an example in deepfake detection, Khormali and Yuan (2022) used an end-to-end Vision Transformer for image deepfake detection to reach a maximum accuracy of 99.49%[18]. On the audio side, wav2vec 2.0 has emerged as a leading choice, using a transformer encoder to contextualize low-level audio features[19] (Fig. 3). wav2vec 2.0 has been shown to be effective for audio deepfake detection, with Tak et al. (2022) implementing wav2vec 2.0 to achieve a superior equal error rate (EER) of

Table 1 Summary of prior deepfake detection strategies[8–11,14,15].

Study	Modality	Primary Technique	Strengths	Weaknesses
Singh and Singh (2021)[8]	Audio	Machine learning	Uses classical ML (RF, SVM) with bispectral and cepstral features that target synthesis artifacts; reports strong test accuracy and clear feature-based interpretability	Relies on handcrafted features and smaller-scale ML which may underperform vs deep learning on large diverse datasets; scalability and generalization to unseen TTS/VC methods unclear, potential overfitting if datasets are small
Bonomi et al. (2021)[9]	Video	Machine learning	Exploits spatio-temporal texture dynamics via LDP-TOP; compact feature set with low-complexity SVM yields strong accuracy and can identify manipulation method	Performance may degrade under heavy compression or low-resolution video; limited adaptability to novel deepfake techniques without retraining and potential dataset bias
Güera and Delp (2018)[10]	Video	Deep learning	Temporal-aware CNN + RNN captures frame-level and sequential cues; high accuracy on very short clips (as little as 2 seconds)	Vulnerable to advanced face-swap tools that remove temporal inconsistencies; limited robustness to unseen generation methods and to low-resolution or heavily compressed video
Wang et al. (2020)[11]	Audio	Deep learning	High detection accuracy with low false alarm rate; robust to voice conversion and real-world noise and validated across commercial TTS systems and multiple languages	Depends on access to and monitoring of neuron behaviors from a specific speaker-recognition DNN, increasing deployment complexity and model-dependence; potentially vulnerable to adaptive adversarial attacks or dataset bias that could reduce generalization
Mittal et al. (2020)[14]	Audio + Video	Deep learning	Introduced a novel approach by comparing emotional congruence between audio and visual modalities	Performance may degrade when both audio and visual modalities are deepfaked; limited by the quality of emotional cues
Salvi et al. (2023)[15]	Audio + Video	Deep learning	Leveraged existing robust deep learning models for feature extraction and deepfake classification	Relying on generic features may not capture deepfake-specific artifacts, potentially reducing detection accuracy

2.85% in audio deepfake detection compared to a baseline[20]. Transformers are capable of converting raw, complex data into meaningful embedding vectors. The transformer models above can generate both image embeddings (Vision Transformers) and audio embeddings (wav2vec 2.0). These embeddings can then be used for downstream tasks, such as training another machine learning model.

Research Goals

The ability of deep learning transformer models to generate vector embeddings provides an effective approach for integrating multimodal data for deepfake detection. For example, an image transformer model such as Vision Transformer can be used to extract image embeddings, while an audio transformer model like wav2vec 2.0 generates audio embeddings. Training these transformers on deepfake samples will significantly improve the quality of the generated embeddings. In addition, training on emotion datasets allows for the creation of emotion embeddings, which, like Mittal et al. (2020) demonstrated, have the potential to improve detection accuracy[14]. Combining these embeddings and feeding them into a downstream model enables the development of a robust multimodal system that leverages complementary information from both visual and auditory modalities. This is the first goal of this research: to incorporate both audio and video data into a deep-learning pipeline for deepfake detection.

In addition, there is limited research identifying which specific facial features are most informative for both human and automated deepfake detection in videos and images. While the

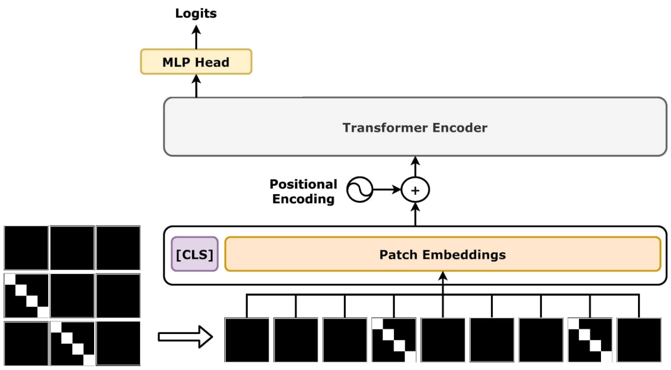

Fig. 2 The Vision Transformer uses patch embedding vectors and the special CLS token as input into a transformer encoder. Image licensed under the Creative Commons Attribution 4.0 International license from https://github.com/dvgodoy/dl-visuals.

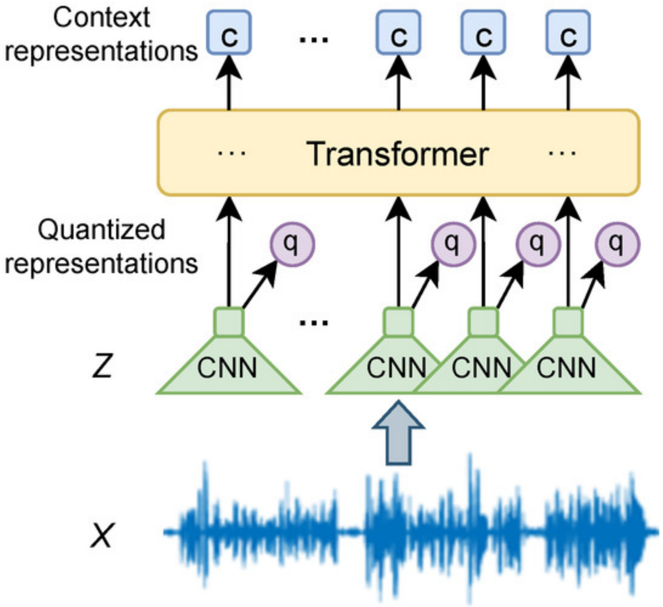

Fig. 3 In wav2vec 2.0, raw audio is encoded by a CNN into latent speech representations. These representations are passed through a transformer to produce contextualized embeddings. Image licensed under Creative Commons Attribution 4.0 International license from https://www.mdpi.com/2079-9292/13/6/1103.

MIT Media Lab offers general guidelines, such as focusing on facial texture and shadows, these broad recommendations lack actionable specificity for practical application[21]. Empirical studies have shown that certain facial regions, like the eyes, mouth corners, and nose, often exhibit artifacts or inconsistencies in deepfakes due to difficulties in accurately synthesizing fine-grained details or natural micro-expressions[22]. Leveraging this knowledge could allow both humans and computational models to prioritize these regions, improving detection efficiency by focusing on the most discriminative features rather than processing the entire face uniformly.

To support this, our approach uses dlib, an established facial landmark detection tool which has been extensively validated in the literature for accuracy and robustness in diverse conditions[23]. These tools enable precise localization of key facial regions, facilitating targeted feature extraction. By combining validated landmark detection with analysis of region-specific importance, the second goal of our work is to provide concrete insights into which facial areas contribute most to distinguishing real from fake, thereby enhancing both human feasibility and computational efficiency in deepfake detection.

Methods

Data Collection

Existing datasets were used to train and evaluate the deep-learning models. All datasets used were publicly available and open source.

The primary dataset used in this study was FakeAVCeleb, a multimodal audio-video deepfake dataset containing 21,544 video files, each with a resolution of 224×224 pixels[24]. There are four distinct classes in this dataset: RealVideo-RealAudio, RealVideo-FakeAudio, FakeVideo-RealAudio, and FakeVideo-FakeAudio (Fig. 4). Furthermore, this dataset contains 500 real videos of English-speaking celebrities, which were used as a base to create the deepfake variants. The dataset is also balanced for ethnicity and gender. Two state-of-the-art (SOTA) video deepfake generation methods were employed in the creation of FakeAVCeleb: FSGAN[25] and Faceswap[26]. One SOTA voice cloning method, SV2TTS[27], was used to generate the audio deepfakes. FakeAVCeleb is currently the only deepfake dataset that contains both labeled deepfake audio and video, making it a highly desirable option for this multimodal detection research. However, a notable limitation of the dataset is its class imbalance, with a higher number of FakeVideo-FakeAudio and FakeVideo-RealAudio samples compared to the other two classes. In this investigation, FakeAVCeleb was used to train the deepfake audio and video deep learning models in Part 1 of the model architecture and to evaluate the overall process in Part 2 of the model architecture.

On the other hand, two emotion datasets were used to train deep learning models for the extraction of emotion embeddings. First, the FER-2013 dataset was used to train the emotion video deep learning model. This dataset contains 32,298 images of faces, each measuring 48 by 48 pixels. There are seven different emotions in the dataset: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral[28]. Second, a combination of 4 audio emotion datasets, CREMA-D[29], SAVEE[30], RAVDESS[31], and TESS[32], were used to train the emotion audio deep learning model. This dataset contains 12,798 audio recordings with the same seven emotions as the earlier dataset[33]. Both datasets

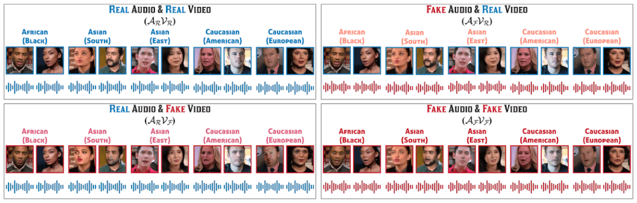
Real Audio & Real Video
($\mathcal{A}_\mathcal{R}\mathcal{V}_\mathcal{R}$)
Fake Audio & Real Video
($\mathcal{A}_\mathcal{F}\mathcal{V}_\mathcal{R}$)
Real Audio & Fake Video
($\mathcal{A}_\mathcal{R}\mathcal{V}_\mathcal{F}$)
Fake Audio & Fake Video
($\mathcal{A}_\mathcal{F}\mathcal{V}_\mathcal{F}$)
African (Black)
Asian (South)
Asian (East)
Caucasian (American)
Caucasian (European)

Fig. 4 The FakeAVCeleb dataset contains four distinct classes: RealVideo-RealAudio, RealVideo-FakeAudio, FakeVideo-RealAudio, and FakeVideo-FakeAudio. Within each class, there are equal numbers of African, Asian (South), Asian (East), Caucasian (American), and Caucasian (European) samples, as well as equal numbers of male and female samples. Image licensed under the Creative Commons Attribution 4.0 International License from https://sites.google.com/view/fakeavcelebdash-lab/home.

have relatively equal class distribution, with similar numbers of samples per class, except for Disgust, in the case of the video dataset, and Surprise, in the case of the audio dataset.

Model Architecture

The model architecture is split into two parts. The underlying idea of this architecture is to use deep learning models to extract valuable auditory and visual information in the form of vector embeddings, for easy combination to make a final prediction. This study investigated three different types of embeddings: general, deepfake, and emotion. The general embeddings were obtained using pretrained deep learning models, while the deepfake and emotion embeddings were derived from specialized models, which were trained in Part 1 (Fig. 5). Both models used for these embeddings were Transformers-based: Vision Transformers for video and wav2vec 2.0 for audio. The specific pretrained checkpoints used for the general models were 'google/vit-base-patch16-224' for Vision Transformer and 'jonatasgrosman/wav2vec2-large-xlsr-53-english' for wav2vec 2.0, both sourced from Hugging Face. Fine-tuning these general checkpoints enabled the training of the deepfake and emotion deep learning models. The multimodal deepfake dataset mentioned earlier was used to train the deepfake audio and video deep learning models, and vice-versa for the audio emotion and video emotion datasets. Specifically, for the deepfake video deep learning model, five images were extracted from each video using OpenCV at approximately uniform intervals across its duration, excluding the first and last frames. This was done to ensure temporal coverage while avoiding boundary frames, and to generate still-image inputs compatible with the Vision Transformer architecture.

The models in this study were trained using a common set of hyperparameters to ensure consistency across all variants. The training process was conducted with a batch size of 16 for both training and evaluation, balancing memory usage and computational efficiency. The models were evaluated and saved at the end of each epoch, facilitating continuous monitoring of performance. Training lasted four to five epochs, stopping when the validation loss ceased to decrease, which proved sufficient for model evaluation within a reasonable timeframe. Most hyperparameters were selected based on prior experience and commonly used values in related work. While no exhaustive grid or random search was performed, several learning rates were tested, and 5e-5 was chosen for its stable convergence and strong validation performance, consistent with transformer-based architectures in similar tasks. The training process also employed 16-bit floating-point precision (FP16) to accelerate computations on compatible GPUs, leveraging the efficiency of Tensor Cores. These training parameters were applied consistently across all model variants (general, deepfake, and emotion), ensuring that the results were directly comparable and that any observed performance differences could be attributed to the model architecture rather than hyperparameter choices.

For the face feature importance analysis in particular, before training the deepfake video deep learning model, dlib facial landmarks (which identified facial features) and OpenCV (which masked the identified regions) were used to block out one of five facial features: eyes (dlib points 37-48), nose (dlib points 28-36), mouth (dlib points 49-68), forehead (estimated as a rectangular region above the eyebrows using points 18-27 and extended upward), or jaw (constructed from jawline points 1-17, divided into left jaw, chin, and right jaw, and extended downward below the mouth). By blocking out each of these facial parts, the resulting deep learning model will be unable to use that facial feature to identify deepfakes, giving a way to determine the importance of each face feature (Fig. 6). Essentially, if blocking out a facial feature leads to a significant decrease in model performance, then that feature is very important for deepfake detection.

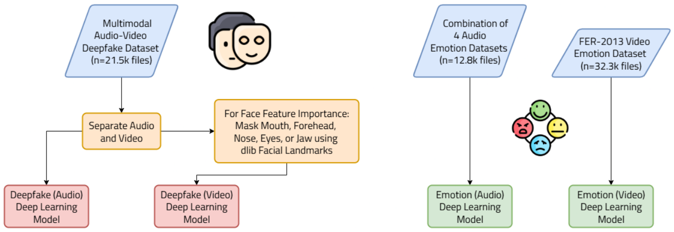
Multimodal Audio-Video Deepfake Dataset (n=21.5k files)
Separate Audio and Video
For Face Feature Importance: Mask Mouth, Forehead, Nose, Eyes, or Jaw using dlib Facial Landmarks
Deepfake (Audio) Deep Learning Model
Deepfake (Video) Deep Learning Model
Combination of 4 Audio Emotion Datasets (n=12.8k files)
FER-2013 Video Emotion Dataset (n=32.3k files)
Emotion (Audio) Deep Learning Model
Emotion (Video) Deep Learning Model

Fig. 5 In Part 1 of the model architecture, the deep learning models were trained. In addition, the face feature importance analysis was achieved by programmatically blocking out one of five facial features (mouth, forehead, nose, eyes, or jaw) before training the video deepfake deep learning model.

Following the training of the deepfake and emotion deep learning models, Part 2 of the model architecture now uses these models, along with the pretrained general models, to extract audio and video embedding vectors, with general, deepfake,

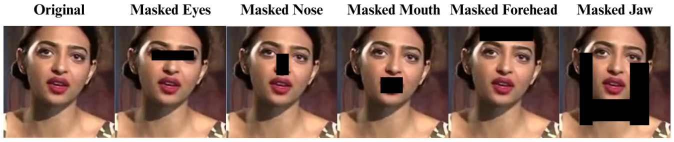

Fig. 6 After applying dlib facial landmarks and OpenCV to the original image, each of the five facial parts was successfully masked.

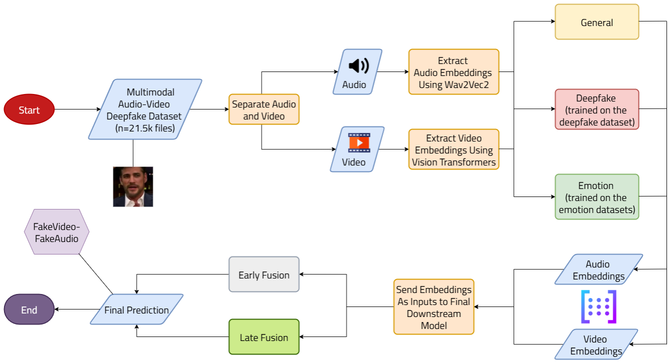
Start
Multimodal Audio-Video Deepfake Dataset (n=21.5k files)
Separate Audio and Video
Audio
Video
Extract Audio Embeddings Using Wav2Vec2
Extract Video Embeddings Using Vision Transformers
General
Deepfake (trained on the deepfake dataset)
Emotion (trained on the emotion datasets)
Audio Embeddings
Video Embeddings
Send Embeddings As Inputs to Final Downstream Model
Early Fusion
Late Fusion
FakeVideo-FakeAudio
Final Prediction
End

Fig. 7 In Part 2 of the model architecture, information was extracted in the form of vector embeddings and used to train the downstream machine learning models, of which there were 13 different architectures. To combine the audio and video embeddings, two different fusion methods were used: early fusion (merging before downstream model training) and late fusion (merging the outputs of two separate downstream models).

and emotion variants (Fig. 7). To ensure consistent audio embedding lengths, all audio embeddings were standardized to a shape of (389, 1024), corresponding to 7.8 seconds of audio, the average sample length in FakeAVCeleb. Shorter inputs were zero-padded, and longer ones were cropped to the target length. Subsequently, max temporal pooling was applied across the first dimension (time), resulting in a fixed-size 1024-dimensional audio embedding for computational efficiency. In contrast, the video embedding was obtained as the 512-dimensional CLS token from the output of the Transformers encoder. The five-images-per-video dataset used in Part 1 was also used here initially, getting embeddings for each image. Then, mean pooling was applied across each of the five-image groupings to get an averaged embedding for each video.

Afterward, the embeddings are then sent as inputs to a final downstream machine learning model, combined using two different methods: early fusion and late fusion. In early fusion, both audio and video embeddings are sent as inputs to a single downstream machine learning model. On the other hand, in late fusion, one downstream model is trained on the audio embeddings while another model is trained on the video embeddings, and the outputs of these two models are combined to make the final prediction. Thirteen different downstream traditional machine learning models were tested: logistic regression, linear discriminant analysis (LDA), decision tree, random forest, XGBoost, HistGradientBoost, LightGBM, k-nearest neighbors (KNN), support vector machine (SVM) with linear kernel, SVM with polynomial kernel, SVM with radial kernel, SVM with sigmoid kernel, and naive Bayes. Default model parameters were used for all models, as the primary objective was to evaluate which downstream model architectures, rather than hyperparameter configurations, perform best in this deepfake detection task.

In total, 3,612 models were created using this process, iterating over 77.7 million files: 84 for early fusion, and 3528 for late fusion. The disparity between these two numbers was due to computational and time constraints.

All code for dataset processing, machine-learning models, and statistical analysis was written in Python via a Jupyter Notebook hosted in the cloud on Kaggle. The Pytorch and Hugging Face Transformers Python libraries were used to implement the deep learning models, while scikit-learn and other model-specific libraries (XGBoost, LightGBM) were used for the downstream machine learning models. Deep learning model training was conducted with one NVIDIA P100 GPU available on Kaggle, with all other processes being run on CPU.

Model Analysis

The primary evaluation metric for this study was the area under the ROC curve (AUC), chosen for its threshold-independent nature and robustness to class imbalance. This makes AUC a reliable measure of the model's overall discriminative capability across all classes, which is particularly important given the inherent imbalance present in the FakeAVCeleb dataset.

t-SNE

To qualitatively evaluate the effectiveness of the extracted embeddings, t-distributed stochastic neighbor embedding (t-SNE) was employed to project high-dimensional embeddings into a two-dimensional space for visualization. t-SNE is a nonlinear dimensionality-reduction technique particularly well-suited for identifying and visualizing different clusters of complex high-dimensional data. In this study, t-SNE was used to qualitatively assess the quality of the three types of embeddings, and how much valuable information they contained for distinguishing between deepfake and real media.

Early Fusion Vs. Late Fusion

To compare the performance of the early fusion and late fusion models, AUC was used as an evaluation metric, with the type of fusion being the independent variable. Specifically, to assess whether the observed difference in AUC was statistically significant, a Mann-Whitney U test was applied.

Performance of Different Embeddings

The AUC metric was also used to evaluate and compare the performance of the three types of embeddings, using a Kruskal-Wallis test and a post-hoc Dunn's test for statistical significance.

Downstream Model Performance
The performance of the different downstream model architectures, measured by AUC, was compared using a Kruskal-Wallis test and a post-hoc Dunn's test with the Bonferroni correction.

Cost Analysis
To estimate the real-world costs of running the model, a cost analysis model was developed using Microsoft Excel to calculate the projected costs of processing the approximately 95 million videos posted on Instagram each day[34]. Average numbers were gathered to approximate the typical operating costs of running the proposed deepfake detection model at scale.

Face Feature Importance
After blocking out the different face parts, as explained earlier in Part 1 of the model architecture, McNemar's test with the Holm-Bonferroni correction was applied to determine whether there was a significant difference in accuracy between the full model and each of the blocked-out Vision Transformer models.

Results

Deep Learning Model Training Results in Part 1 of the Model Architecture

Both the deepfake and emotion models showed good training results (Table 2). For example, both deepfake audio and video models achieved a validation accuracy of around 100%, while the performance of the emotion models was slightly lower, due to the greater number of classes in the dataset (7 for emotion vs 2 for deepfake). The video emotion model performance was lower than that of the audio emotion model, which can be attributed to the greater difficulty of the FER-2013 image dataset, where the state-of-the-art performance is only 73.28%[35]; the achieved accuracy of 67.76% is therefore relatively close to the SOTA, indicating competitive performance given the dataset's challenging nature. Training time for all models ranged from three to five hours and may be further reduced with the usage of higher-performance computing resources.

Table 2 Validation accuracies of Part 1 deep learning model training (%).

		Dataset Type	
		Deepfake	Emotion
Model Type	Audio	100.00	88.13
	Video	99.97	67.76

t-SNE

Fig. 8 displays the t-SNE visualizations. Each dot represents one sample, with red representing the "Deepfake" class and blue representing the "Real" class. Clear clustering and separation of classes (the red dots separated from the blue dots) were observed only for the deepfake embeddings. In contrast, the general and emotion embeddings did not exhibit such separation. This provides some qualitative confirmation that the deepfake deep learning models are the most effective because they assign distinctive vector representations between the samples of the two classes, enabling accurate classification.

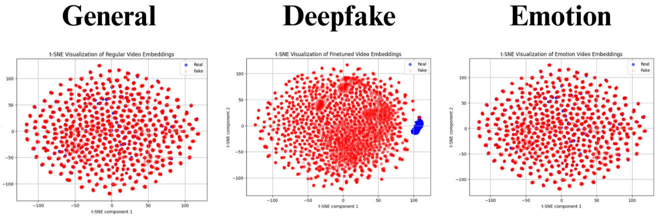

Fig. 8 In the t-SNE visualizations of general, deepfake, and emotion video embeddings, each blue dot represents one real sample, and each red dot represents one deepfake sample. Only the deepfake embeddings had clear separation between these two colors/classes.

Early Fusion Vs. Late Fusion

Fig. 9 shows the mean and one standard deviation of the AUC for the early fusion and late fusion models. A nonparametric Mann-Whitney U test was selected instead of a t-test because the data did not follow a normal distribution, supported by a significant Shapiro-Wilk test ($p < 0.001$). Using a significance level of $p = 0.05$, the Mann-Whitney U test was insignificant ($p = 0.18$), suggesting that the type of fusion method did not significantly affect model performance.

Performance of Different Embeddings

Fig. 10 displays the average AUC of the models created using each of the three types of embeddings. A Shapiro-Wilk test was significant ($p < 0.001$), indicating non-normality. As a result, the non-parametric Kruskal-Wallis test was conducted, showing a significant difference between at least two of the groups ($p < 0.001$). After this, a post-hoc Dunn's test with the Bonferroni correction was performed for pairwise comparisons. The models created using deepfake embeddings had a significant difference in AUC compared to those created using general or emotion embeddings (both cases $p < 0.001$), suggesting that deepfake embeddings have superior performance. In addition, there was no significant difference between the AUC of the emotion models and the general models, implying that emotional features are ineffective for deepfake detection.

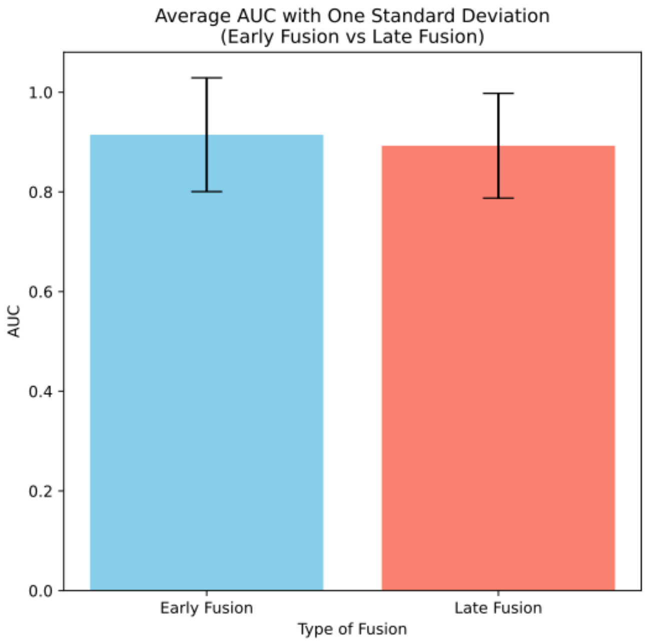

Fig. 9 There was no significant difference ($p = 0.18$) in AUC between the early fusion and late fusion models.

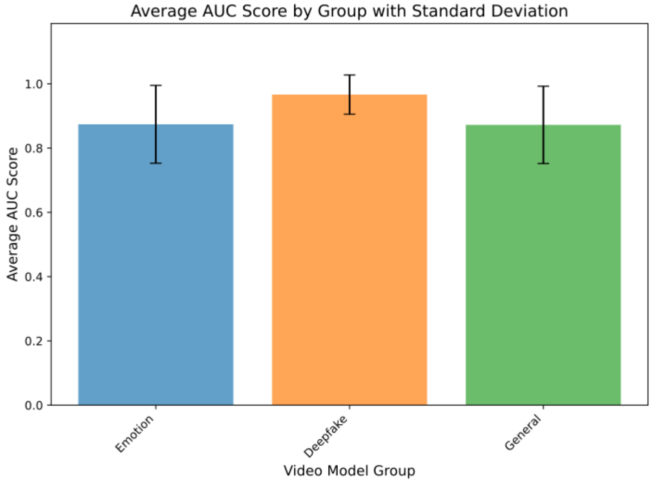

Fig. 10 Models that used deepfake embeddings had significantly higher AUCs ($p < 0.001$) compared to models created from the other two embedding types.

Downstream Model Performance

Several top downstream models performed at a similarly high level, as evidenced by a significant Kruskal-Wallis test ($p < 0.001$) and post-hoc Dunn's test, which showed no significant difference among the top six downstream models (Fig. 11). The best-performing model in this study achieved a test accuracy of 99.53%, an AUC of 99.96%, a precision of 98.74%, and a recall of 95.95%, substantially outperforming both human evaluators and the current state-of-the-art deepfake detection methods, which report 89.73% accuracy, 91.70% AUC, 88.89% precision, and 89.47% recall in the existing literature[12] (Table 3). In addition, the best model surpassed existing multimodal deepfake detection approaches. For example, Salvi et al. (2023)[15] achieved an AUC of only 80% on the FakeAVCeleb dataset.

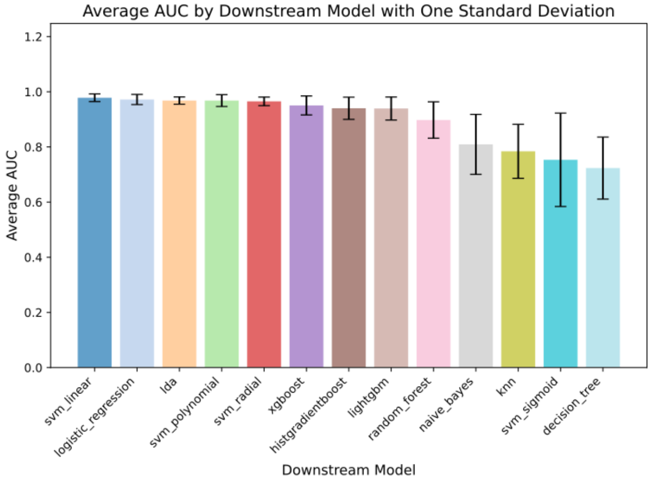

Fig. 11 Although the lowest-ranked models had a considerable decline in performance, the best downstream models all performed similarly ($p < 0.001$).

Table 3 Best model metrics versus current literature (%).

	Accuracy	AUC	Precision	Recall
Best model	**99.53**	**99.96**	**98.74**	**95.95**
Current literature[12]	89.73	91.70	88.89	89.47

Cost Analysis

Table 4 summarizes the results of the cost analysis model. Calculations were based on the processing rate (samples per second) of a single NVIDIA Tesla P100 GPU. Then, after determining the required samples per second to process the 95 million videos uploaded to Instagram daily, the number of necessary GPUs was determined, and the electricity cost was based off this. It was estimated that a one-time cost of $7,200, along with an annual electricity cost of $5,124.60, would be sufficient to support real-time deepfake detection for Instagram's daily inflow of digital content, an insignificant expense for a multi-billion-dollar company like Instagram. This calculation assumes a consistent posting rate throughout the day and typical electricity costs, though the actual costs are unlikely to deviate significantly from these estimates.

Table 4 Necessary cost in USD for the model to process the 95 million videos posted to Instagram daily.

# of NVIDIA Tesla P100 GPU:	# of samples	Time to run (s)	Processing rate (samples per second)		
1	107,720	1,680	64.12		
Number of videos posted on Instagram daily[34]	Required processing rate (samples per second)	Required # of GPUs	Cost per GPU ($)[36]	Total one-time cost ($)	
95,000,000	1099.54	18	400	7,200	
Power consumption of 1 GPU (W)[37]	Total power for 18 GPUs (kW)	Energy used in 24 hours (kWh)	Average US commercial electricity rate in 2024 ($/kWh)[38]	Daily electricity cost ($)	Yearly electricity cost ($)
250	4.5	108	0.13	14.04	5,124.60

Face Feature Importance

McNemar's test was applied to compare the test accuracies of the full model and models with specific facial regions masked. A significant difference ($p < 0.05$) was found between the full model and the models with masked eyes and nose (Fig. 12), suggesting that these regions are particularly important for deepfake detection performance. These results indicate that the eyes and nose provide critical cues for identifying deepfakes and should be prioritized by both human evaluators and automated systems.

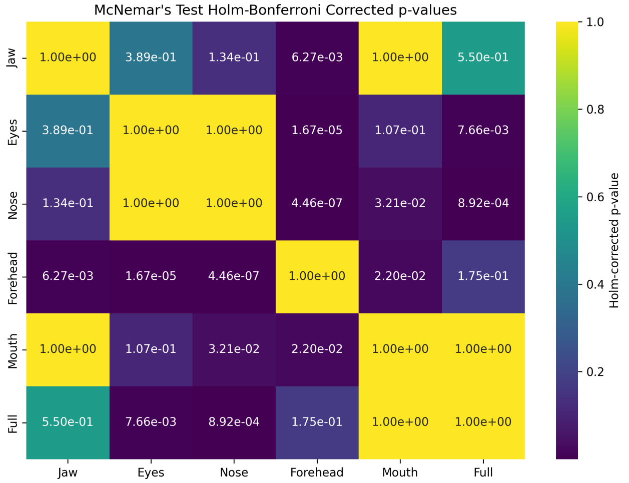

Fig. 12 Statistically significant differences in accuracy ($p < 0.05$) were found between the full model and the masked eyes and nose models, indicating that these regions contribute most to deepfake detection performance.

Discussion

This research presents a comprehensive evaluation of embedding-based deep learning deepfake detection models, uniquely designed to utilize both audio and video information to detect both audio and video deepfakes with one process. Fusion strategy had a minimal impact on overall performance, with no significant difference in the performances of the early fusion and late fusion models. Deepfake embeddings had the best performance and effectively distinguished between real and deepfake, while general and emotion embeddings did not. Many of the top downstream models were close to each other in performance, suggesting that the embeddings were information-rich enough that high performance could be achieved with simple downstream models. The best model overall achieved SOTA results with an accuracy of 99.53% and an AUC of 99.96%, substantially outperforming humans and current top models, including multimodal solutions. The model is cost-effective, requiring only a few thousand dollars for processing tens of millions of videos, which demonstrates the feasibility of implementing it at scale. One intended use case is for all videos posted online to be run through the deepfake detection model for flagging, which is achievable because of the low investment required.

In addition, this research identified the nose and eyes as the most critical facial regions for both human and computer-based deepfake detection, opening new avenues for more efficient and effective deepfake detection strategies. Therefore, both of the main goals of this research were accomplished.

Limitations and Future Directions

This study has several limitations that point to important avenues for future work aimed at improving the practical applicability and robustness of multimodal deepfake detection. One critical challenge is the performance drop often observed when models are applied to real-world or out-of-distribution content. This domain shift arises from differences between the training data distribution and the characteristics of in-the-wild media, including variations in compression, resolution, lighting, speech accents, and background noise. In multimodal systems, failures may occur if either the audio or video modality is degraded, such as low-quality audio disrupting lip-sync analysis or video artifacts obscuring facial features, or if the manipulations differ from the methods represented in training. Additionally, models can inadvertently be overfit to dataset-specific artifacts rather than learning manipulation-invariant features, making them brittle when encountering unseen manipulation styles or naturally occurring distortions.

Although AUC was chosen as the primary evaluation metric to mitigate the influence of class imbalance during performance assessment, further strategies such as loss reweighting or oversampling could enhance robustness under imbalanced conditions and should be investigated in subsequent studies.

Future work should prioritize adversarial robustness, as deepfake generation techniques are rapidly evolving and adversaries may deliberately attempt to evade detection through subtle perturbations or by exploiting weaknesses in learned representations. Developing models resilient to such attacks through adversarial training, certified defenses, or robust feature learning will be essential for real-world deployment. Beyond English audio, multilingual and cross-lingual generalization remains underexplored; collecting and training on diverse multilingual datasets will improve detection accuracy across languages, dialects, and accents, which is particularly relevant for global applications. Practical deployment scenarios also require attention to real-time inference constraints, including model efficiency, latency, and resource usage, especially for applications in live streaming or on mobile devices. Research into lightweight architectures, model compression, and optimized inference pipelines can help bridge this gap. Moreover, there exists potential for hyperparameter optimization to further improve performance. Finally, future work should continue to emphasize domain adaptation techniques to improve generalization to diverse, noisy, and out-of-distribution data, and investigate fusion strategies that dynamically weigh audio and video signals depending on their quality or reliability in a given sample. Together, these directions will contribute to more robust, scalable, and deployable deepfake detection systems.

Conclusion

As deepfakes continue to grow in sophistication and prevalence, developing effective detection methods is essential to mitigate the risks they pose to information integrity and public trust. This study demonstrates the effectiveness of a multimodal deep learning approach that combines both audio and video information to detect deepfakes with SOTA performance. Moreover, the results of the facial feature importance analysis provide a significant contribution in the area of efficient deepfake detection strategies that even humans can utilize. Together, these findings represent a significant step forward in deepfake detection, offering a robust, interpretable, and economically viable framework that can adapt to the evolving challenges of synthetic media.

Acknowledgments

This research has received no external funding.

References

1 N. Jacobson, *Deepfakes and their impact on society*, https://www.openfox.com/deepfakes-and-their-impact-on-society.

2 S. Lalchand, V. Srinivas, B. Maggiore and J. Henderson, *Generative AI is expected to magnify the risk of deepfakes and other fraud in banking*, https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2024/deepfake-banking-fraud-risk-on-the-rise.html.

3 B. Allyn, *Deepfake video of Zelenskyy could be tip of the iceberg*, https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia, in info war, experts warn.

4 S. Bond, *A political consultant faces charges and fines for Biden deepfake robocalls*, https://www.npr.org/2024/05/23/nx-s1-4977582/fcc-ai-deepfake-robocall-biden-new-hampshire-political-operative.

5 G. Cluley, *UK Prime Minister Keir Starmer and Prince William deepfaked in investment scam campaign*, https://www.bitdefender.com/en-us/blog/hotforsecurity/uk-prime-minister-keir-starmer-and-prince-william-deepfaked-in-investment-scam-campaign.

6 N. Kbis, B. Dolealov and I. Soraperra, *Fooled twice: People cannot detect deepfakes but think they can.*

7 K. Mai, S. Bray, T. Davies and L. Griffin, *Warning: Humans cannot reliably detect speech deepfakes*.

8 A. Singh and P. Singh, *Detection of AI-synthesized speech using cepstral bispectral statistics*.

9 M. Bonomi, C. Pasquini and G. Boato, *Dynamic texture analysis for detecting fake faces in video sequences*.

10 D. Guera and E. Delp, *Deepfake video detection using recurrent neural networks*.

11 R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma and Y. Liu, *DeepSonar: towards effective and robust detection of AI-synthesized fake voices*.

12 M. Rana, M. Nobi, B. Murali and A. Sung, *Deepfake detection: A systematic literature review*.

13 L. Verdoliva, *Media forensics and deepfakes: an overview*.

14 T. Mittal, U. Bhattacharya, R. Chandra, A. Bera and D. Manocha, *Emotions dont lie*.

15 D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang and S. Tubaro, *A robust approach to multimodal deepfake detection*.

16 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, . Kaiser and I. Polosukhin, *Attention is all you need*.

17 A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner and X. Zhai, *An image is worth 16x16 words: transformers for image recognition at scale*.

18 A. Khormali and J.-S. Yuan, *DFDT: an end-to-end deepfake detection framework using Vision Transformer*.

19 A. Baevski, H. Zhou, A. Mohamed and M. Auli, *wav2vec 2.0: a framework for self-supervised learning of speech representations*.

20 H. Tak, M. Todisco, X. Wang, J.-W. Jung, J. Yamagishi and N. Evans, *Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation*.

21 M. Groh, *Detect deepfakes: How to counteract misinformation created by AI*, `https://www.media.mit.edu/projects/detect-fakes/overview`.

22 T. Carvalho, C. Riess, E. Angelopoulou, H. Pedrini and A. Rezende Rocha, *Exposing digital image forgeries by illumination color classification*.

23 V. Kazemi and J. Sullivan, *One millisecond face alignment with an ensemble of regression trees*.

24 H. Khalid, S. Tariq, M. Kim and S. Woo, *FakeAVCeleb: A novel audio-video multimodal deepfake dataset*.

25 Y. Nirkin, Y. Keller and T. Hassner, *FSGAN: Subject agnostic face swapping and reenactment*.

26 I. Korshunova, W. Shi, J. Dambre and L. Theis, *Fast face-swap using convolutional neural networks*.

27 Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Moreno and Y. Wu, *Transfer learning from speaker verification to multispeaker text-to-speech synthesis*.

28 I. Goodfellow, D. Erhan, P. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, . Romaszko, B. Xu, Z. Chuang and Y. Bengio, *Challenges in representation learning: a report on three machine learning contests*.

29 H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova and R. Verma, *CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset*.

30 P. Jackson and S. Haq, *Surrey Audio-Visual Expressed Emotion (SAVEE) database*, `https://www.researchgate.net/publication/260311132_Surrey_Audio-Visual_Expressed_Emotion_SAVEE_database`.

31 S. Livingstone and F. Russo, *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English*.

32 K. Dupuis and M. Pichora-Fuller, *Toronto Emotional Speech Set*.

33 U., *Valainis*, `https://www.kaggle.com/datasets/uldisvalainis/audio-emotions`.

34 C. J, *Instagram doubles monthly users to 500M in 2 years, sees 300M daily*, `https://techcrunch.com/2016/06/21/instagram-500-million`.

35 Y. Khaireddin and Z. Chen, *Facial emotion recognition: state of the art performance on FER2013*.