Non-Invasive Real-Time Multimodal Deception Detection Using Machine Learning and Parallel Computing Techniques

Fahad Abdulridha

fahad.m.abdulridha@aliraqia.edu.iq

Al-Iraqia University

Baraa M. Albaker

Al-Iraqia University

Research Article

Keywords: Deception Detection, Multimodal, Real-time systems, Visual features, Acoustic features, Linguistic features, Machine learning, Parallel computing

Posted Date: February 1st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3910179/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Social Network Analysis and Mining on May 7th, 2024. See the published version at https://doi.org/10.1007/s13278-024-01255-4.

Abstract

Detecting deception reliably has been a sought-after goal for researchers since the early 20th century. This is due to its high stakes nature, especially when considering the setting under which a deception detection system would be utilized, such as law enforcement sectors, judicial bodies, and criminal investigation bodies. Therefore, recent literature has greatly focused on developing systems with ever increasing accuracies to minimize false positives. Attempts to diversify the sources of data being analyzed and classified as well as the reliance on artificial intelligence technologies have shown great success. But little attention was paid to the applicability of these systems in real-life scenarios. an idle deception detection system needs to exhibit accuracy but also perform in real-time, a feature that is lacking in the current state-of-the-art. In this work, a non-invasive real-time multimodal deception detection system is developed using advanced machine learning techniques. It combines data sources from video and audio streams to extract visual, acoustic, and linguistic features. It also utilizes parallel computing techniques to ensure high performance adequate for real-time usage. Furthermore, a user-friendly graphical user interface was built to facilitate the use of the system. Multiple experiments were conducted to determine its accuracy under various circumstances and combinations of features, the system had a detection accuracy of 99.83% under real-life, high-stakes scenarios using visual and acoustic features, and 91.03% accuracy under controlled environments, while an 89.54% accuracy was achieved using mixed environments and using visual, acoustic, and linguistic features.

1. Introduction

One of the most important challenges humans have struggled with is the ability to discern lies in communication, whether it's in the judicial, law enforcement, or day-to-day contexts. This is especially true when it comes to high stakes situations where criminal behavior is suspected. This led to many attempts to develop a method or a system which efficiently, effectively, and repeatedly makes that distinction. Attempts such as taking the wrist pulse or observing facial expressions were commonplace. However, it has been shown that human's ability to detect deception manually is no more accurate than pure chance[1], [2].

Advancements in technology have facilitated the emergence of more refined systems, bolstering confidence in the ability to detect deception. Among the pioneering implementations is the polygraph and its various iterations—a device designed to gauge physiological changes in the human body, including temperature, heart rate, and respiration rate, among others. These changes are believed to be correlated with deception. Despite its initial adoption in popular media, the polygraph faced substantial criticism for its susceptibility to producing false positives in readings [3]. This criticism ultimately prompted its abandonment by judiciary bodies, acknowledging its demonstrated unreliability [4]. However, the polygraph, while deemed unreliable, maintained its popularity among popular media. This popularity was not unfounded, it was due to its ability to work in real-time situations and therefore provide immediate feedback, which is a significantly valuable feature as evident by its popularity despite

the emergence of far more reliable and well researched systems in the literature that lacked real-time capabilities.

With the advent of artificial intelligence and machine learning, innovative methods began to develop, relying less on physiological changes induced by stress—a major contributor to false positive results. Instead, these approaches shifted focus towards features directly associated with deception, often involuntarily manifested while maintaining minimum correlation with stress. Addressing a significant challenge, these new technologies successfully reduced dependence on invasive techniques that involve encroaching upon the subject's personal space with sensors and monitoring devices, either attached to the body or within their immediate surroundings. Invasive systems not only exacerbate the subject's stress levels, but also perform sub optimally when the data analyst or judge, overseeing the system and interpreting output data, is physically present and visible to the subject. Buller and Burgoon's work, “Interpersonal Deception Theory” [5], illustrates how various interview contexts, including the environment, incentives, and relationships, can impact the subject's behavior, compelling them to exert greater effort in concealing the truth.

One of the initial solutions proposed to address the challenges of stress and invasiveness was the utilization of facial expressions. By capturing subtle changes in facial expressions, it becomes possible to identify deceit. The advantage of employing facial expressions for deception detection lies in its independence from sensors, judges, or the presence of an interviewer or investigator. A basic camera, positioned at a distance or discreetly concealed, along with a teleprompter for posing questions, proves sufficient for this method. This efficiency stems from the absence of a need for special sensors or human analysts to interpret data. The sole input required is a straightforward video feed of the subject's face, and the analysis is executed by an algorithm with minimal to no human intervention.

Another promising and equally non-invasive modality for detecting deception, involves the analysis of audio features generated by the subject under investigation. These features encompass physical changes in acoustic expressions, including pitch, intensity, spectral, cepstral (MFCC), duration, spectral harmonicity, psychoacoustic spectral, and sharpness, among others. Another set of features in audio-based deception detection pertains to emotion-related attributes within the audio domain. In contrast to the low-level attributes of the acoustic domain, these emotion-driven features are considered high-level, as they are derived from a combination of low-level features. Examples include arousal levels, valence, and various emotion categories [6], [7], [8].

Finally, text-based deception detection has also risen in popularity due to its varied use cases, which includes written text in internet messaging services, emails, social media, news articles, as well as transcriptions of audio conversation recordings. These relied on detecting patterns in the text to indicate false information, spam, emotion, or deception attempts [9]. Text based deception detection was especially useful because it did not suffer from the same weak points that audio or visual modals had such as the quality of the recordings, the setting at which the recording was done, lighting, noise, among others.

While these modalities and the use of artificial intelligence were a major step in the field of deception detection, due to the high-stakes nature of the challenge, improving classification performance further was imperative. Furthermore, it needed to be more adaptable to real world applications and reduce bias while increasing confidence in the results produced. This led researchers to experiment with a multimodal approach. A technique that combines multiple sources of data and making predictions based on their combined patterns. This new approach further improved accuracy and reliability of deception detection systems. However, it didn't improve on the adaptability front. This is due to the multimodal approach being implemented in similar fashion to the individual modal, where classification was done as a separate process from the investigation and cannot be directly integrated into it due to the lack of real-time capability of these systems.

This work sets out to develop the first ever non-invasive real-time multimodal deception detection system that combines the high performance of multimodal techniques with real-time classification capabilities. In order to ensure reliability, a machine learning model was trained on two of the largest databases in the field, together they cover real-life environments such as court trials as well as controlled environments for low stakes situations. Visual, acoustic, and linguistic modals were utilized in various combinations to analyze the performance of the system under various situations. The raw data from each modal was aggregated separately in a fashion which produces samples that match among all modals in terms of sample size and sampling rate, which then undergo feature level fusion before being fed to the classifier. A deep analysis was conducted to determine the appropriate factors that govern the sample generation and matching for each modal, this allowed the generated samples to be meaningful in terms of discriminating patterns while maintaining interoperability with other modals.

2. Related work

The field of deception detection in recent years has been moving towards diversifying the techniques and modalities employed. In fact, a simple keyword search in the Scopus database in published articles relating to deception detection from the last 13 years, reveals there has been as much published papers with the keyword "multimodal" in in the title in the last 3 years as there was in the 10 years prior. This is due to the limitations that one modal will inevitably face, this is further exacerbated by the limited databases available for research. There is a total of 8 databases publicly available [10], only 2 of which support audio and acoustic features out of the box, the rest include raw video clips which require additional preprocessing and feature extraction. Furthermore, these datasets are relatively low in the number of participants with an average of 51 participants per database and 101 at the high end and 26 at the low end. Therefore, researchers quickly shifted towards a multimodal approach to address the problem of deception detection from as many angles as possible simultaneously.

Karnati et al. [11] developed a multimodal deception system which can discern deceptive sentiment using visual, auditory and EEG signals features. They used multiple databases in their experiments and a combination of deep learning techniques for the classification job. They achieved an accuracy of up to 98%. Kamboj et al. [12] attempted to create their own dataset by collecting videos from the internet of

political figures and labeling them using PolitiFact. The target modals they were aiming for are linguistic, acoustic, and visual features. A decision tree classifier was used for the classification job. They achieved accuracy up to 69%. Şen et al. [13] used multiple classifiers, namely Random Forest, Support Vector Machine, and Neural Network. They were experimented with against the authors own database which they developed and published in 2015 [2] which is made up of a collection of real life court trails videos. By utilizing visual and acoustic features, they achieved classification accuracy of 84% and 83% with the addition of the linguistic features. Chebbi and Jebara [14] were able to achieve a 93% accuracy using feature level fusion approach that involved video, audio and text features with KNN (K Nearest Neighbor) as the classifier. Farahani and Moradi [15] designed their own experiment with 52 participants, 40 male and 12 female. The participants' EEG data was collected while they ran through a series of predetermined scenarios where participants were either deceptive or truthful. Genetic Support Vector Machine (GSVM) to classify the EEG data along with a few other machine learning algorithms. They achieved 95.45% accuracy for autonomic response features and 93% for autonomic as well as event-related potential features. Another EEG based attempt was done by Lakshan et al. [16] where they developed a real-time deception detection, emotion detection, and attention detection system. Their best performing deception detection attempt had an F1 score of 87% using random forest classifier on 30 live participants.

Table 1 illustrates the multimodal attempts along with the data, features, and classifier type used in recent literature. These works were curated for their high performance or unique approach in features or classifier choice. It can be seen that multimodal systems are highly flexible with many opportunities for innovation, given the number of variables at the researcher's disposal to work with albeit at the expense of increased complexity.

Table 1. Top results from recent literature on multimodal deception detection

Paper	Year	Data	Feature	Classifier	ACC
[11]	2022	MU3D [17]	Video, Audio	Softmax Classfier	ACC=0.9814
[12]	2020	Politifact.Com	Linguistic, Acoustic, Visual	Decision Tree	ACC=0.69
[13]	2022	Real life court trials [2]	Visual and Acoustic	NN	ACC=0.8418
[14]	2021	Real life court trials [2]	Audio, Video	KNN	ACC=0.85; F1=0.85
[15]	2017	Experimental	EEG	GSVM	ACC=0.9545
[18]	2018	Real life court trials [2]	Visual and Verbal	SVM	ACC=0.99
[19]	2019	Real life court trials [2]	Audio, Text, Micro Expressions	AdaBoost, SRKDA, and Linear SVM	ACC=0.97
[20]	2021	Real life court trials [2]	Visual, Acoustic, and Verbal	FFCSN (An Adversial Learning Module)	ACC=0.97; AUC=0.9978
[21]	2019	Experimental (Females, Abortion Topic)	Thermal and Visual	Decision Tree	ACC=0.736

3. Databases used

The first dataset utilized in this study is "The Miami University Deception Detection Database (MU3D)" [17]. It consists of 320 videos, involving 80 participants (40 males and 40 females) with ages ranging from 18 to 26 (refer to Fig. 1). Each participant recorded four videos, recounting four different stories representing positive truth, negative truth, positive lie, and negative lie scenarios. The duration of the videos' ranges from 24 to 57 seconds, with an average duration of 35 seconds (as shown in Fig. 2). Audio sampling rate is either 44.1 kHz or 48 kHz.

The second database utilized in this study is called "Real Life Deception Detection Database" [2]. It comprises 121 publicly available videos recorded during trials and courtrooms, featuring witnesses or defendants. The dataset includes 61 deceptive clips and 60 truthful clips, involving 56 subjects (21 females and 35 males) aged between 16 and 60 years. The duration of the video clips ranges from 4 seconds to 1 minute and 11 seconds, with an average duration of 27.7 seconds (as depicted in Fig. 3). Audio sampling rate is either 44.1 kHz or 48 kHz.

4. Feature extraction and processing

Feature extraction is a major cornerstone in the field of deception detection in terms of quality as well as quantity. This is because data in its raw form are typically video or audio recordings which have no

discernable patterns relating to deception. Therefore, for a given deception detection system to be effective, information needs to be extracted from the raw data and presented in a fashion that is meaningful in accordance with the design of said system. For the system proposed in this work, three modals are introduced. Visual modal, where the facial features are extracted in real time fashion instead of treating facial queues as discrete events (also known as action units), they are captured as a continuous stream of data points representing position, orientation, and direction as well as the aforementioned action units for every frame in the video feed. The audio modal is also utilized where two feature sets are extracted, acoustic features representing the real time changes in physical attributes of the vocal domain, and linguistic features which are transcribed in real time for the target's speech.

4.1 Visual features

Visual features were extracted using the OpenFace 2.0 toolkit [22]. This toolkit extracts the eye gaze, head pose, facial landmarks and AUs for each frame in each video, this is achieved using a combination of computer vision and machine learning algorithms. The toolkit is capable of working in offline batch mode or real-time input using a video feed. The output dataset contains 8 eye gaze features (position and angle), 280 eye position features, 6 head pose features, 380 facial landmark features, 35 AU features for a total of 709 features (see fig. 4).

4.2 Acoustic features

Acoustic features were extracted using the OpenSMILE toolkit [23], A cross platform open-source tool for extracting audio features based on a wide array of feature sets. The tool aggregates datapoints extracted from the source audio feed into samples by calculating the average of all datapoints in a preset time frame which will control the sampling rate. The time frame can be manipulated to match the sampling rate of the acoustic features with the sampling rate of the visual features which is determined by the frame rate. This will be important for the real-time classification method employed in the system.

The ComParE_2016 feature set [8] used in this work has three feature levels, LLDs, low level descriptors calculated over a sliding window with 65 features, a secondary level of the LLD calculating the delta regression of LLDs with 6373 features, and finally, the functionals feature level which is a statistical mapping of the variables from the LLDs into static values with 65 features.

The acoustic features extracted using LLDs include energy, intensity, critical band spectra, MFCC, auditory spectra, loudness approximated from auditory spectra, perceptual linear predictive (PLP) coefficients, perceptual linear predictive cepstral coefficients (PLP-CC), linear predictive coefficients (LPC), line spectral pairs, fundamental frequency, probability of voicing from ACF and SHS spectrum peak, jitter and shimmer, formant frequencies and bandwidths, zero and mean crossing rate, spectral features, psychoacoustic sharpness, spectral harmonicity, CHROMA features, F0 harmonics ratios. LLD

feature level was chosen for this work for representing the physical changes of the audio signal, unlike the other two levels which are mathematical and statistical derivatives of LLD [24].

4.3 Linguistic features

The transcriptions for each of the two databases were produced using Whisper “large” speech to text model [25]. Each raw audio recording was transcribed individually and conversations where two or more people were involved were separated, only the target subject speech was kept for processing and model training. The produced transcriptions were timestamped on an individual word level for mapping to their corresponding audio and visual datapoints at the same time frame as illustrated in table 2. MU3D database produced 32,718 words with approximately 171 average words per minute. RL database produced 8,055 words with approximately 66 average words per minute.

Table 2. A sample of transcribed speech with timestamps

“My best friend is a really nice person….”			
Word	**Start time (Seconds)**	**End time (Seconds)**	**Confidence**
"my"	0.16	0.48	0.9409
"best"	0.48	0.719	1.0
"friend"	0.719	1.22	1.0
"is"	1.28	1.52	0.998
"a"	1.52	1.599	0.999
"really"	1.599	1.92	0.99
"nice"	1.92	2.24	0.996
"person"	2.24	2.74	0.95

In the preprocessing phase, two expanded datasets were derived from each original database, encompassing all the transcriptions and their corresponding labels. Subsequently, each dataset underwent lemmatization, a process that reduced the occurrence of similar words by transforming each word to its base form and eliminating duplicates. Additionally, stop words were eliminated from the dataset, and all words were converted to lowercase to ensure a consistent and informative final dataset.

The lemmatized dataset was then vectorized using Scikit-Learn’s implementation of MurmurHash3 hashing algorithm, a 32-bit value non-cryptographic hashing function that assigns a token occurrence to each word. The resulting vectorized transcriptions, now organized as a matrix, had their values scaled down using a term-frequency times inverse document-frequency (TF-IDF) algorithm. This was done to mitigate the influence of tokens with high frequencies, which tend to be less informative (refer to Fig. 7 for details).

5. Classification

Random forest (RF) is used for the model training and classification task due to its out of the box resilience to overfitting, high dimensional data and noise[26], [27]. This is especially useful for the datasets produced in this work. RF is also one of the most successful machine learning algorithms for real world problem solving [28].

The cuML python library by RAPIDS implementation of RF is used due to its utilization of the GPU to accelerate computation speed when working with such large datasets. Hyperparameter optimization was also performed using DASK, another python library that allows the utilization of the GPU for accelerated and distributed computation, the Randomized Search Cross Validation implementation is used with a 5-fold cross validation for optimizing the number of classifiers/decision trees in the ensemble (*n_estimators*), max number/percentage of features to be considered by each spawned decision tree (*mtry*), the max number of splits that each decision tree is allowed to make (*max_depth*). Table 3 highlights the values found:

Table 3. Estimation of optimized values for hyperparameters of the model

Hyperparameter	Best value
n_estimators	4000
mtry	0.8
max_depth	15

The hyperparameter values shown above seem to work best with all datasets with the remaining hyperparameters left as default. A great attention was given to the mtry hyperparameter due to the high dimensional data being worked with, despite RFs natural resilience to this factor, it was necessary to make sure that minimal overfitting was occurring since no feature selection or dimensionality reduction took place in producing any of the datasets. By keeping the mtry range of values to be searched high (0.6 – 0.9), it was ensured that the best mtry value found be sufficiently high to make sure little to no overfitting took place, albeit at the expense of the increase in computational requirements.

6. Pre-fusion single modal method and results

In order to set up a baseline for the real-time deception detection system, each modality will be experimented with individually in offline mode, to determine the effectiveness of the system compared to state-of-the-art.

6.1 Visual features

The feature extraction process generated time series samples for each video in the form of frames with 709 associated features. This resulted in the creation of 320 datasets for the MU3D database and 121 datasets for the "Real Life Deception Detection Database" (RL) where each video corresponds to one

dataset. All datasets under each database were then concatenated, creating two final datasets. Each of the two underwent normalization to scale their features within the range of 0 to 1.

In the realm of deception detection, a novel approach was adopted for training the machine learning classifier and label prediction. In it, every frame was treated as an individual data sample. However, since detecting deception from facial expressions relies on Macro and Micro expressions[29], [30], [31], [32], [33], which typically last between 0.2 and 0.5 seconds [6], a simple moving average calculation (as described in equation 1) was applied to each feature. This calculation employed a moving window (k) with values of 5, 10, 15, and 30. These values were chosen based on the average frame rate of the videos in the databases, which is 29 frames per second. This ensured that the minimum expression captured spanned no less than 5 frames and no more than 15 frames with the addition of 30 frame range or 1 second worth of data, to match the videos' frame rate. This approach yielded a new set of samples (p), each representing a unique value for a group of samples (n-k) equal to the window size, thereby allowing the classifier to consider multiple frames simultaneously.

The datasets resulting from the normalization and window averaging processes were concatenated within each dataset, leading to the creation of two larger datasets for each window size. Lastly, a third larger dataset was formed for each window size by combining the two previous datasets, as illustrated in figure 8. In total, these parameters produced nine datasets, with three datasets for each of the three window sizes.

$$SMA_k = \frac{1}{k}\sum_{i=n-k+1}^{n} p_i \qquad \text{Eq. 1}$$

Table 4. Classification results for each dataset and window size K for all facial features

Dataset	Accuracy
K=5	
MU3D	**78.79%**
RL	59.16%
MU3D+RL	74.38%
K=10	
MU3D	79.95%
RL	**94.68%**
MU3D+RL	81.97%
K=15	
MU3D	80.77%
RL	**94.58%**
MU3D+RL	82.68%
K=30	
MU3D	85.68%
RL	**97.75%**
MU3D+RL	88.75%

Two trends can be observed from the results in table 4. First, varying the increasing the window size K had a positive impact on the performance of all datasets, especially RL where a 38.06% improvement was achieved by increasing the window size from 5 to 10. This can be rationalized by the fact that RL dataset is comprised of video clips of widely different sources, which meant largely varying video angles, lighting, video quality, distance from the camera, all affecting micro expression detection especially on small k values, increasing k seemed to be very impactful since it allowed the classifier to capture more information since each sample was the product of a bigger data pool.

Second, MU3D seems to perform very well among all window sizes on average, this can be attributed to the fact that the dataset is generally far more uniform when it comes to filming angles, lighting, video quality and even the subjects are uniform in their seating and body positions. This effect carried to the MU3D-RL dataset bringing its average up even when k=5 where RL performed poorly.

6.2 Acoustic features

In order to produce comparable results to the visual features and later the real time results, the sampling rate was set to match the frame rate of the video feed. This corresponds to 30 samples per second. Similar to the visual modal, both databases were experimented with as well a third dataset produced from concatenating MU3D and RL acoustic datasets. The classification results are illustrated in table 5.

Table 5. Acoustic features classification results for each dataset

Dataset	Accuracy
Sampling rate = 1 sample/second	
MU3D	81.80%
RL	**91.75%**
MU3D+RL	62.33%

From the results above, it can be seen that RL dataset outperforms MU3D and the combined dataset. This can be justified by the fact that RL database is comprised of real-life interviews and court trails where the speakers are under distress and expressing real emotions in their speech patterns. Unlike visual features, this made their acoustic features more informative and less uniform where patterns can be detected. MU3D on the other hand includes subjects in low stakes environment, allowing them to be more in control of their emotions, making their acoustic features more uniform with less patterns to be detected [34].

6.3 Linguistic features

Text classification in offline mode was done in batches similar to the state-of-the-art systems approach. Disregarding the timestamps, a video's transcription is classified as a whole. Prior to the classification process, a feature selection algorithm was executed on the datasets to identify the most influential

keywords for training the Random Forest classifier. This algorithm employed a Chi-Square scoring method to select the top 100 features from each dataset. Subsequently, the classification process was carried out using various n-gram variations, and the results are depicted in table 6.

Table 6. Classification accuracy for each n-gram variation.

n-gram (top 100 features)	MU3D	RL
Unigram	%71.87	%75.67
Bigram	%81.25	**%94.59**
Trigram	%75.0	%67.56
Unigram + Bigram	%73.95	%75.67
Unigram + Bigram + Trigram	**%84.37**	%81.08

The results above show consistent patterns across both datasets. When all n-gram values are combined in the vectorizer, it yields robust results, with MU3D achieving its highest accuracy using this approach at 84.37%. In the case of RL, while it also achieves a high accuracy with this combination, its peak accuracy is obtained from bigrams at 94.59%, marking a state-of-the-art result in the field of deception detection [35].

7. Real-time feature level fusion and model training

7.1 Preprocessing

Most multimodal systems rely on voting mechanisms to make the final classification. This approach has shown a lot of merit in offline batch classification. However, a real-time system needs to classify the data on the fly. Furthermore, data size needs to be matched among all modalities, where in a given time frame, all the events that took place in it are aggregated and fed to the classifier as a single sample. In order to determine the appropriate time frame that all modals will adhere to, it needs to be large enough to encompass enough data among all modals to be meaningful and discriminating, as well as leave minimum empty data points where a modal experienced no events at a certain point in time within the time frame. It also needs to be small enough to minimize lag and maintain the real-time nature of classification.

To choose the right time frame, two factors were considered for each modal. The minimum and maximum data rate and the minimum and maximum sample size that amounts to a meaningful expression. For the visual modal, the sample rate is equal to the frame rate of the video feed, while the meaningful expression range is 0.2 to 0.5 seconds as mentioned previously. From the visual modal classification results above, 0.5 second frame appears to yield better results overall.

The linguistic modal sample rate range to choose from is equal to the visual modal data rate as the minimum, and shortest word utterance duration within both databases for the maximum, this range will

encompass the longest word utterance duration in both databases. The minimum meaningful expression the linguistic modal has to offer is a single word, and the maximum is an entire sentence.

The sample rate for the acoustic modal, unlike the visual modal, isn't fixed and can effectively be as high as the audio feed sample rate of up to 48 kHz. There isn't an obvious precise meaningful expression range for the acoustic modal, however, it can be narrowed down to equal the visual modal sample rate for the minimum and equal to the linguistic meaningful expression for the maximum. This ensures that acoustic samples integrate with the visual and linguistic modal which have a more fixed and defined sample size definitions. Table 7 illustrates the summary of factors and their ranges as well as the chosen value for the system.

Table 7. Summary of the time frame size factors and chosen values

Modal	Factors Considered	Range	Chosen Value
Visual	Sample Rate	Equal to video frame rate	30 samples/second
	Meaningful Expression Range	0.2 to 0.5 seconds	0.5 seconds
Linguistic	Sample Rate	Equal to visual modal data rate	30 words/second
	Meaningful Expression Range	Single word to an entire sentence	However many words that fit within 1 second
Acoustic	Sample Rate	Up to 48 kHz (audio feed sample rate)	30 samples/second
	Meaningful Expression Range	0.5 to 1 seconds	1 second worth of aggregated acoustic feature samples

In order to achieve feature level fusion among all modals, sample size needs to be matched. Each sample of data will be equivalent to 1 second of data from each modal which represents the time frame value. This is implemented by taking the average of 30 datapoints from the visual features and 1 second worth of data from the acoustic features, and however many consecutive words that fit within 1 second. This time frame value satisfies the criteria chosen for the acoustic and linguistic modalities. However, it doesn't satisfy the visual modal criteria of 0.5 seconds sample size that was chosen. To mitigate this, two sets of overlapping samples will be produced and interleaved with a 0.5 second delay. This creates samples that have an intersection time of 0.5 seconds. This allows the system to detect patterns that occur on the edges of any given sample since the next sample contains data that's captured from 0.5 to 1.5 seconds of the previous sample instead of 1.0 to 2.0 seconds. Figure 9 illustrates the sample generation and interleaving process.

$$Interleaved\ samples = C_1[1], C_2[1.5], C_1[2], C_2[2.5], \ldots\ C_1\left[\frac{i}{2}\right], C_2\left[\frac{i-1}{2} + 0.5\right] \quad for\ i \in \mathbb{N}_0 \qquad \text{Eq. 2}$$

$$Interleaved\ samples = \begin{cases} C_1\left[\frac{i}{2}\right], & if\ i\ is\ even \\ C_2\left[\frac{i-1}{2} + 0.5\right], & if\ i\ is\ odd \end{cases} \qquad \text{Eq. 3}$$

7.2 Feature level fusion

After matching all three modalities in terms of sample size, features were concatenated horizontally. This produced the final batch of datasets that will be used for model training and real-time classification. Three datasets were produced similar to the individual modal experiments, MU3D dataset, RL dataset and a combined dataset produced by vertically concatenating the two datasets. Finally, the values of the visual and acoustic features of each dataset were scaled down between 0 and 1. Figure 10 illustrates the overall feature extraction and preparation steps.

7.3 Classification results and discussion

The classification process discussed in section 5 was carried out for the three generated datasets. Each dataset was split into an 80% training set and a 20% testing set. Each dataset was run through five experiments, one for each modal, visual and acoustic modal, and all modals. All experiments were conducted under the same conditions such as hyperparameters and classification model. Classification results for all experiments are illustrated in table 8. Confusion matrix and learning curves for the combined datasets and modals are illustrated in figure 11 and 12 respectively.

Table 8. Real-time multimodal classification accuracy results

Modal	MU3D dataset	RL dataset	MU3D+RL dataset
Visual	90.55%	98.83%	88.75%
Acoustic	57.47%	85.57%	62.33%
Linguistic	55.79%	99.58%	62.47%
Visual + Acoustic	90.40%	**99.83%**	88.69%
Visual + Acoustic + Linguistic	**91.03%**	98.58%	**89.54%**

7.3.1 Modalities' Unique Contributions

The study demonstrates that each modality offers distinct strengths in the deception detection process. Visual cues consistently exhibit high accuracy across both the MU3D and RL dataset and their combination. This emphasizes the significance of non-verbal communication in discerning deceptive behavior. Acoustic cues under real-life data as shown in the RL results also shown respectable performance, this suggests that auditory signals in real-life applications have an impactful effect on detecting deception in real-time. The same effect is seen in the linguistic modal, where speakers under uncontrolled and relatively high stress environments have more discernable patterns in their communication.

7.3.2 Role of each dataset

The inclusion of the Real-Life Deception Detection Database (RL) provides a valuable perspective on the generalizability of the deception detection system to real-world scenarios. The dataset's influence on the system's performance underscores the importance of evaluating deception detection algorithms in authentic, diverse settings. The findings suggest that the multimodal approach, when applied to real-life data, is effective in capturing deceptive patterns that may differ from controlled environments. MU3D, while performing poorly on the acoustic and linguistic front, showed promising results in the visual and visual + acoustic modal. This highlights the versatile nature of the deception detection system in this work for its viability in controlled environments where speakers are more aware and in control of their non-verbal cues. The combination of the two datasets also provided similar insight into the versatility of the system as well as the importance of diversifying the training data to expose the system to more varied scenarios that can be encountered.

7.3.3 Synergistic Effects of Modal Combinations

Combining modalities continues to be a crucial strategy for achieving a comprehensive understanding of deceptive behavior. The Visual + Acoustic combination, and particularly when incorporating linguistic cues, results in a highly accurate deception detection system. The integration of modalities, especially when considering the nuances of real-life scenarios from the RL dataset, emphasizes the practicality and adaptability of the multimodal approach. The results achieved when combining all datasets and modals outperform all other modal combinations for the same dataset and outperform many of the single dataset experiments.

8. Comparing results to state-of-the-art

Currently, there are no works in literature which implement similar real-time capabilities to compare against. However, raw classification performance can be compared with the state-of-the-art to demonstrate the reliability of this system in terms of accuracy without considering real-time capabilities.

The classification model in this work was trained based on the MU3D dataset, RL dataset and their combination. Karnati et al. [11] utilized the same datasets in their work minus the combination of them, this makes it the ideal candidate for comparison. Their work employed the visual and audio modal in addition to electroencephalogram (EEG) modal. The classification is done using a softmax classifier on each modal and the final decision is made via a score level fusion.

The common experiments between the two works are the visual, audio and visual+audio modals on both RL and MU3D. for the RL dataset, this work outperforms theirs on the visual and visual+audio modals with 98.83% versus 97.35% and 99.83% versus 97.33% respectively while underperforming in the audio modal with 85.57% versus 94.45%. MU3D on the other hand underperformed overall with 90.55% versus 98.22% for the visual modal, 57.47% versus 95.80% for the audio modal, and 90.40% versus 98.14% for the visual+audio modal (see figure 13 for summary of result comparison).

9. System development and analysis

A graphical user interface was developed using python employing the real-time deception detection classifier based on the combined modals and datasets model. It captures video and audio stream input from a webcam, using the OpenCV library, the system is able to detect the existence of a face in the video stream and draws a green squire around the detected face, this is used to control classification job where the system only captures and classifies received data when a human face is detected. Another safety guard is implemented to halt classification is when no audio is detected, this is done in two stages. The first stage is when acoustic data falls below a given threshold which indicates no vocal signal is detected, and the second stage is to check whether a valid transcription is obtained from the audio stream, indicating whether or not actual speech exists. These safety guards are actively analyzing the data stream in real-time and can halt or resume the classification job accordingly.

The system captures data over a 1.5 second period. This is then split into two time frames, 0 – 1 second time frame and 0.5 – 1.5 second time frame. These two time frames are independently classified as containing deception or not and the results are printed. The system also indicates the results of both classifications by changing the color of the square drawn around the face to red if both results are deceptive, and orange if one of them is deceptive, as an additional visual cue for a more user-friendly interface as demonstrated by figure 14.

The performance of the system was also analyzed and optimized for real-time operation. Throughout the testing phase, the system was analyzed to find performance bottlenecks, and the largest bottleneck found was the feature extraction task followed by the data processing task, especially with the video stream. This was alleviated by employing parallel computing on these two tasks, where instead of processing these tasks sequentially, the feature extraction task for the video stream was split into 3 time frames, each 0.5 seconds long, and features were extracted from each of them in parallel, the first and second frame and second and third frames are then concatenated into their expected form. This provided a 7.86% speed up in performance in the video feature extraction and processing task. The same

technique was also used for the audio stream where acoustic features and transcriptions were extracted and processed for both time frames in parallel, this provided a 8.22% and 5.50% speed up in performance respectively.

Figure 15 illustrates the pre and post optimized system workflow. These optimizations provided some speedup in performance; however, the feature extraction task remains the largest bottleneck in the system due to technical limitations within OpenFace and OpenSmile that cannot be mitigated. For OpenFace, these limitations are lack of native compatibility with python and parallelism support that would allow feature extraction of multiple frames at once. For OpenSmile, it has native python compatibility, but parallelism is not supported, however, it provides an option to increase the number of workers dedicated to the extraction task, but the performance speed up plateau's at a certain point and no additional speed up is gained by increasing the number of workers, suggesting that there are hard limits in terms of speed within the tool. Table 9 highlights the worst recorded delays throughout the testing period for each task as well as the overall system delay in both sequential and parallel computing modes. The hardware specifications of the test machine are the following: the CPU is an AMD Ryzen 7 7700X, the GPU is an Nvidia GeForce RTX 4080 with 16 GB of VRAM, the RAM is a 32 GB DDR5 memory with 5200MHz speed.

Table 9. System performance in sequential computing vs. parallel computing (worst recorded case)

Task	Time delay in sequential computing (seconds)	Time delay in parallel computing (seconds)	Time difference (Seconds)	Speed up percentage
Visual	2.3722	2.1992	0.1730	7.86%
Acoustic	7.1937	6.6474	0.5463	8.22%
Transcription	7.0124	6.6467	0.3657	5.50%
Classification	0.3321	0.3320	~ 0	0.02%
Total time per system cycle	7.6900	7.1119	0.5782	8.13%

Conclusion

In this work, a real-time multimodal deception detection system with graphical user interface was developed to explore new avenues in terms of applicability in real-world scenarios. the system was able to outperform many of the-state-of-the-art systems in the field that operate in offline or batch mode. When employing all modalities, the system was able to achieve a 91% accuracy under controlled environments, 98.58% accuracy under real-life environments, and 89.54% accuracy under mixed environments. Its best accuracy however was achieved under the combination of the visual and acoustic modalities at 99.83% under real-life environments. These results, when paired with the real-time classification capability, sets up the system as a valuable tool for detecting deception in high stakes situations such as courtrooms and police investigations where immediate feedback is necessary for decision-making. It also highlights the viability of the real-time approach in the field which hasn't been

explored by the literature. This opens up new paths in the field of deception detection for researchers to explore and expand to develop versatile systems that can be highly adaptable in real-life applications.

Declarations

Authors contribution statement

Fahad Abdulridha: Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation. **Baraa M. Albaker**: Supervision, Writing - Review \& Editing, Investigation, Validation.

Competing interests

The authors of this work declare that to their knowledge, there are no competing financial interests nor personal relationships of any nature that would influence or cause bias in this work and the results reported in it.

References

1. P. Ekman and M. O'Sullivan, "Who can catch a liar?," *Am. Psychol.*, vol. 46, pp. 913–920, 1991, doi: 10.1037/0003-066X.46.9.913.
2. V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception Detection using Real-life Trial Data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, in ICMI '15. New York, NY, USA: Association for Computing Machinery, Nov. 2015, pp. 59–66. doi: 10.1145/2818346.2820758.
3. L. Saxe, D. Dougherty, and T. Cross, "The validity of polygraph testing: Scientific analysis and public controversy," *Am. Psychol.*, vol. 40, pp. 355–366, 1985, doi: 10.1037/0003-066X.40.3.355.
4. R. Adelson, "Psychological sleuths--The polygraph in doubt. Monitor on Psychology," https://www.apa.org/monitor/julaug04/polygraph.
5. D. B. Buller and J. K. Burgoon, "Interpersonal Deception Theory," *Commun. Theory*, vol. 6, no. 3, pp. 203–242, 1996, doi: 10.1111/j.1468-2885.1996.tb00127.x.
6. B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Interspeech 2009*, ISCA, Sep. 2009, pp. 312–315. doi: 10.21437/Interspeech.2009-103.
7. B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism." Accessed: May 28, 2023. [Online]. Available: http://www.interspeech2013.org/
8. B. Schuller *et al.*, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Interspeech 2016*, ISCA, Sep. 2016, pp. 2001–2005. doi: 10.21437/Interspeech.2016-129.
9. M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: a survey and new perspectives," *Soc. Netw. Anal. Min.*, vol. 10, no. 1, p. 82, Sep. 2020, doi:

10.1007/s13278-020-00696-x.

10. A. Omirali, A. Shoiynbek, K. Kozhakhmet, and N. Sultanova, “A Review of Deception Detection Databases,” 2022, doi: DOI : 06.2016-67962946/2022.7658.
11. M. Karnati, A. Seal, A. Yazidi, and O. Krejcar, “LieNet: A Deep Convolution Neural Network Framework for Detecting Deception,” *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 3, pp. 971–984, Sep. 2022, doi: 10.1109/TCDS.2021.3086011.
12. M. Kamboj, C. Hessler, P. Asnani, K. Riani, and M. Abouelenien, “Multimodal Political Deception Detection,” *IEEE Multimed.*, vol. 28, no. 1, pp. 94–102, Jan. 2021, doi: 10.1109/MMUL.2020.3048044.
13. M. U. Şen, V. Pérez-Rosas, B. Yanikoglu, M. Abouelenien, M. Burzo, and R. Mihalcea, “Multimodal Deception Detection Using Real-Life Trial Data,” *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 306–319, Jan. 2022, doi: 10.1109/TAFFC.2020.3015684.
14. S. Chebbi and S. B. Jebara, “Deception detection using multimodal fusion approaches,” *Multimed. Tools Appl.*, Jun. 2021, doi: 10.1007/s11042-021-11148-9.
15. E. D. Farahani and M. H. Moradi, “Multimodal detection of concealed information using Genetic-SVM classifier with strict validation structure,” *Inform. Med. Unlocked*, vol. 9, pp. 58–67, 2017, doi: 10.1016/j.imu.2017.05.004.
16. I. Lakshan, L. Wickramasinghe, S. Disala, S. Chandrasegar, and P. S. Haddela, “Real Time Deception Detection for Criminal Investigation,” in *2019 National Information Technology Conference (NITC)*, Oct. 2019, pp. 90–96. doi: 10.1109/NITC48475.2019.9114422.
17. K. Hugenberg, A. R. McConnell, J. W. Kunstman, E. P. Lloyd, J. C. Deska, and B. Humphrey, “Miami University Deception Detection Database,” Mar. 2017.
18. N. Carissimi, C. Beyan, and V. Murino, “A Multi-View Learning Approach to Deception Detection,” in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi’an: IEEE, May 2018, pp. 599–606. doi: 10.1109/FG.2018.00095.
19. S. Venkatesh, R. Ramachandra, and P. Bours, “Robust Algorithm for Multimodal Deception Detection,” in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, San Jose, CA, USA: IEEE, Mar. 2019, pp. 534–537. doi: 10.1109/MIPR.2019.00108.
20. L. Mathur and M. J. Mataric, “Unsupervised Audio-Visual Subspace Alignment for High-Stakes Deception Detection,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada: IEEE, Jun. 2021, pp. 2255–2259. doi: 10.1109/ICASSP39728.2021.9413550.
21. M. Abouelenien, M. Burzo, V. Perez-Rosas, R. Mihalcea, H. Sun, and B. Zhao, “Gender Differences in Multimodal Contact-Free Deception Detection,” *IEEE Multimed.*, vol. 26, no. 3, pp. 19–30, Jul. 2019, doi: 10.1109/MMUL.2018.2883128.
22. T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “OpenFace 2.0: Facial Behavior Analysis Toolkit,” in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, May 2018, pp. 59–66. doi: 10.1109/FG.2018.00019.

23. F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*, in MM ’10. New York, NY, USA: Association for Computing Machinery, Oct. 2010, pp. 1459–1462. doi: 10.1145/1873951.1874246.
24. F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, “On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common,” *Front. Psychol.*, vol. 4, p. 292, May 2013, doi: 10.3389/fpsyg.2013.00292.
25. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision”.
26. T. G. Dietterich, “Ensemble Methods in Machine Learning,” in *Multiple Classifier Systems*, in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.
27. R. Caruana, N. Karampatziakis, and A. Yessenalina, “An empirical evaluation of supervised learning in high dimensions,” in *Proceedings of the 25th international conference on Machine learning*, in ICML ’08. New York, NY, USA: Association for Computing Machinery, Jul. 2008, pp. 96–103. doi: 10.1145/1390156.1390169.
28. M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, “Do we need hundreds of classifiers to solve real world classification problems?,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, Jan. 2014.
29. M. G. Frank and E. Svetieva, “Microexpressions and Deception,” in *Understanding Facial Expressions in Communication: Cross-cultural and Multidisciplinary Perspectives*, M. K. Mandal and A. Awasthi, Eds., New Delhi: Springer India, 2015, pp. 227–242. doi: 10.1007/978-81-322-1934-7_11.
30. F. D. Farizi, S. Bangay, and S. Mckenzie, “Facial Cues for Deception Detection in Virtual Reality Based Communication,” in *Proceedings of the 3rd International Conference on Big Data and Internet of Things*, in BDIOT ’19. New York, NY, USA: Association for Computing Machinery, Aug. 2019, pp. 65–69. doi: 10.1145/3361758.3361782.
31. L. Su and M. Levine, “Does ‘lie to me’ lie to you? An evaluation of facial clues to high-stakes deception,” *Comput. Vis. Image Underst.*, vol. 147, pp. 52–68, Jun. 2016, doi: 10.1016/j.cviu.2016.01.009.
32. Z. Dong, G. Wang, S. Lu, L. Dai, S. Huang, and Y. Liu, “Intentional-Deception Detection Based on Facial Muscle Movements in an Interactive Social Context,” *Pattern Recognit. Lett.*, vol. 164, pp. 30–39, Dec. 2022, doi: 10.1016/j.patrec.2022.10.008.
33. M. Ding, A. Zhao, Z. Lu, T. Xiang, and J.-R. Wen, “Face-Focused Cross-Stream Network for Deception Detection in Videos,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 7794–7803. doi: 10.1109/CVPR.2019.00799.
34. M. K. Camara, A. Postal, T. H. Maul, and G. Paetzold, “Can lies be faked? Comparing low-stakes and high-stakes deception video datasets from a Machine Learning perspective.” arXiv, Aug. 18, 2023. doi: 10.48550/arXiv.2211.13035.

35. A. S. Constâncio, D. F. Tsunoda, H. de F. N. Silva, J. M. da Silveira, and D. R. Carvalho, “Deception detection with machine learning: A systematic review and statistical analysis,” *PLOS ONE*, vol. 18, no. 2, p. e0281323, Feb. 2023, doi: 10.1371/journal.pone.0281323.

Figures

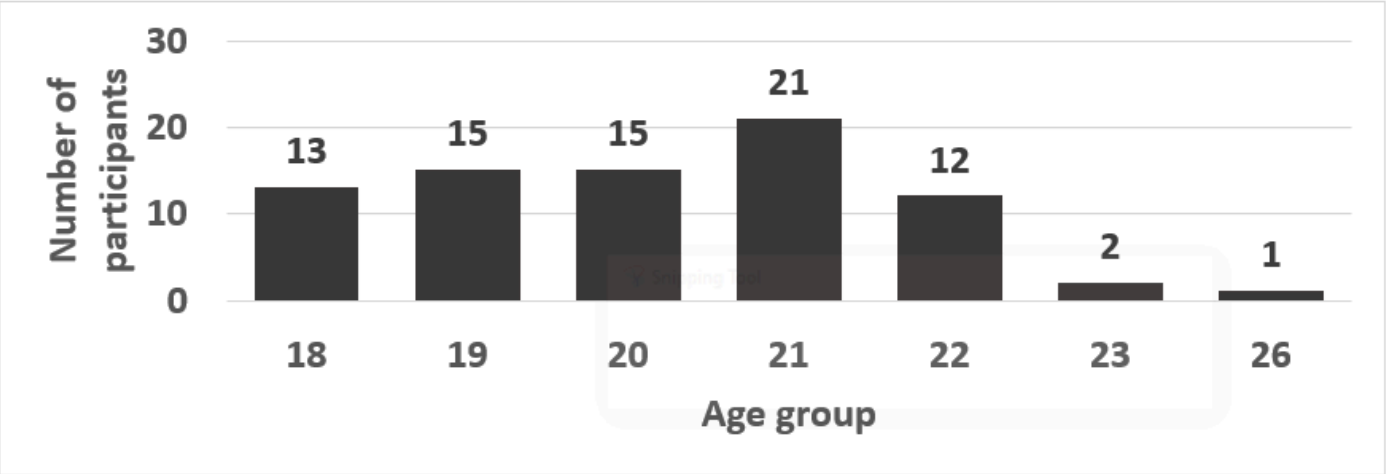

Figure 1

Number of participants for each age group

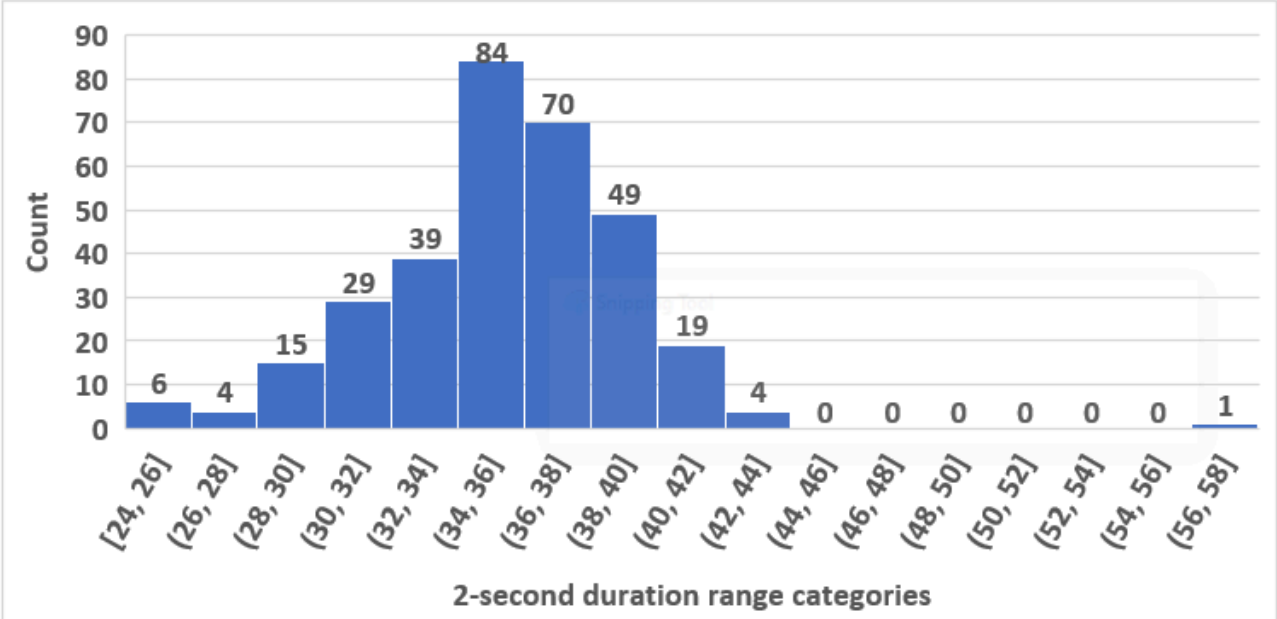

Figure 2

Audio clips duration distribution for MU3D dataset.

Figure 3

Various frames from the MU3D database

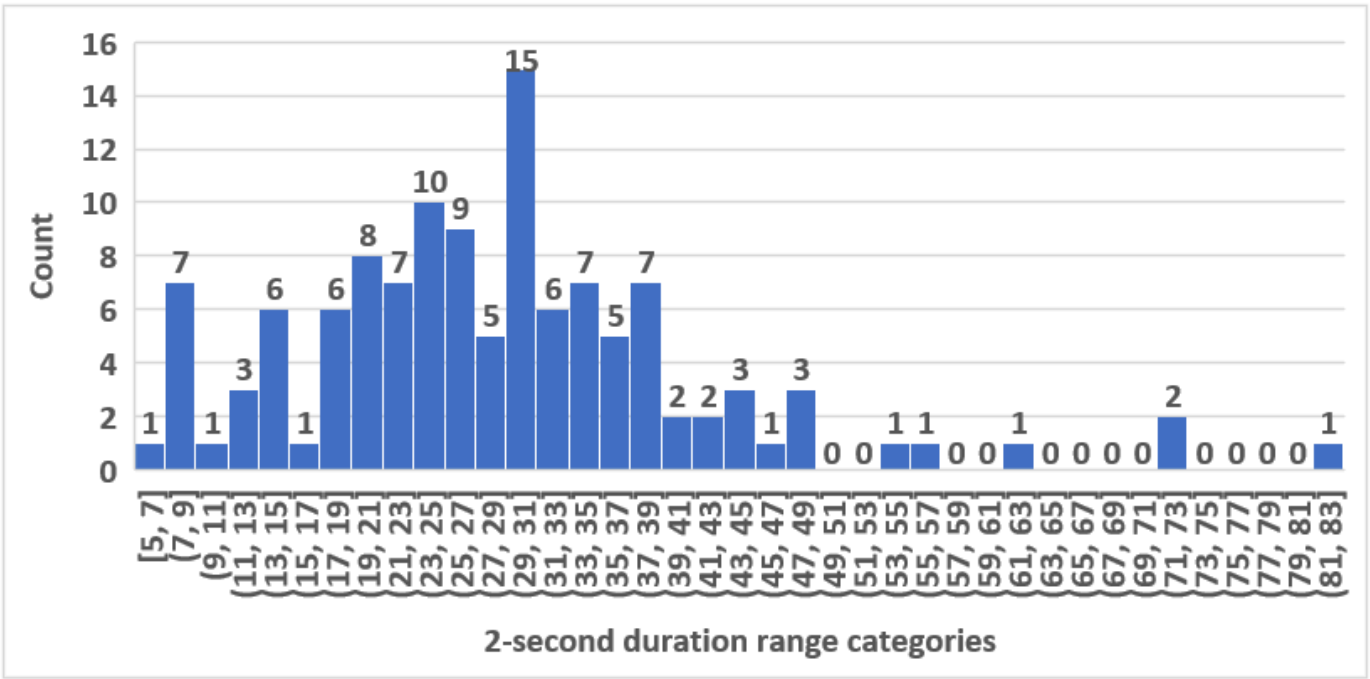

Figure 4

Audio clips duration distribution for RL dataset.

Figure 5

Various frames from the RL database

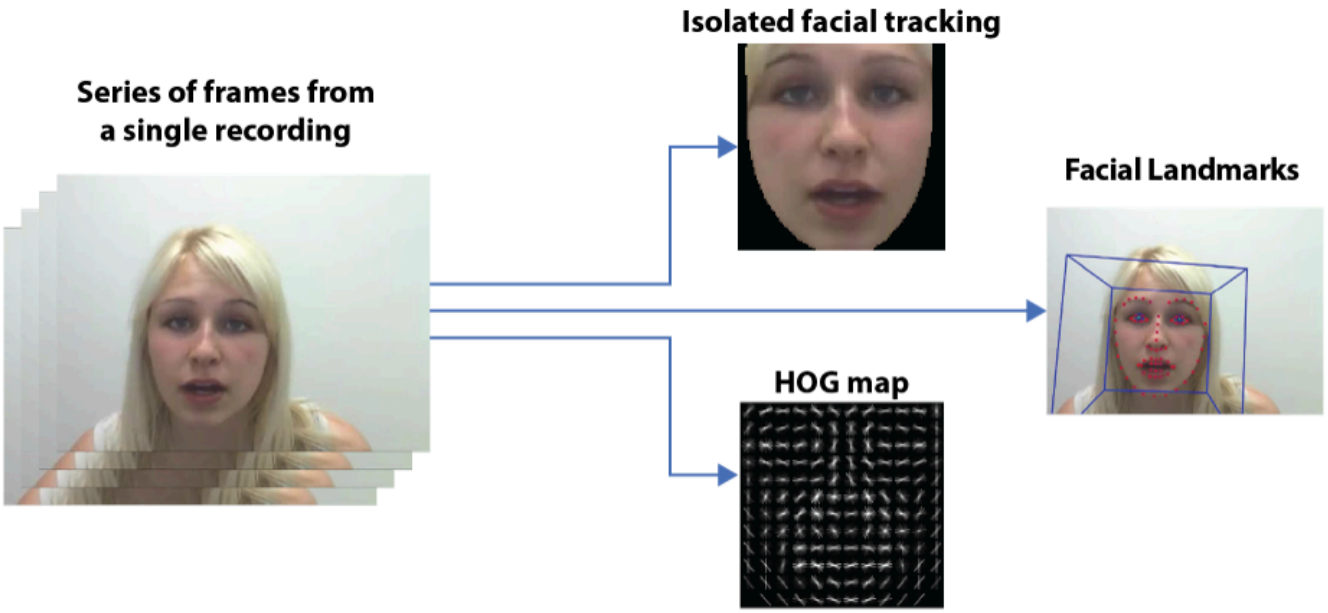

Figure 6

Feature extraction illustration of facial features

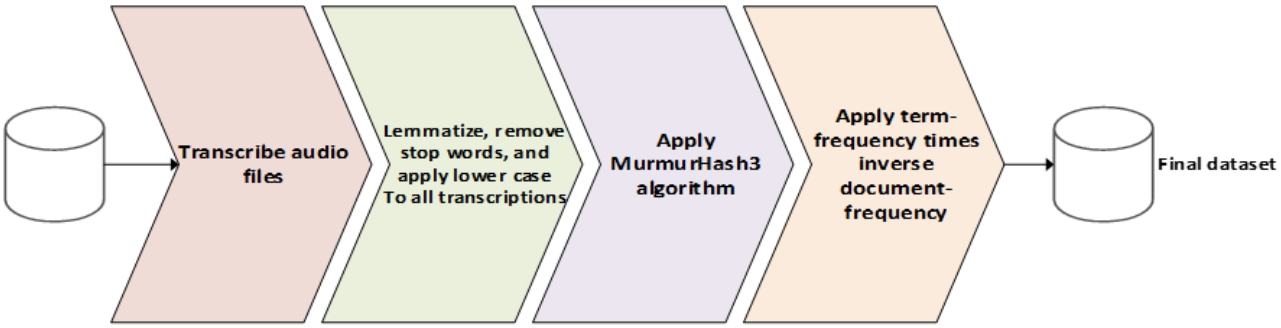

Figure 7

Audio transcription and text processing steps

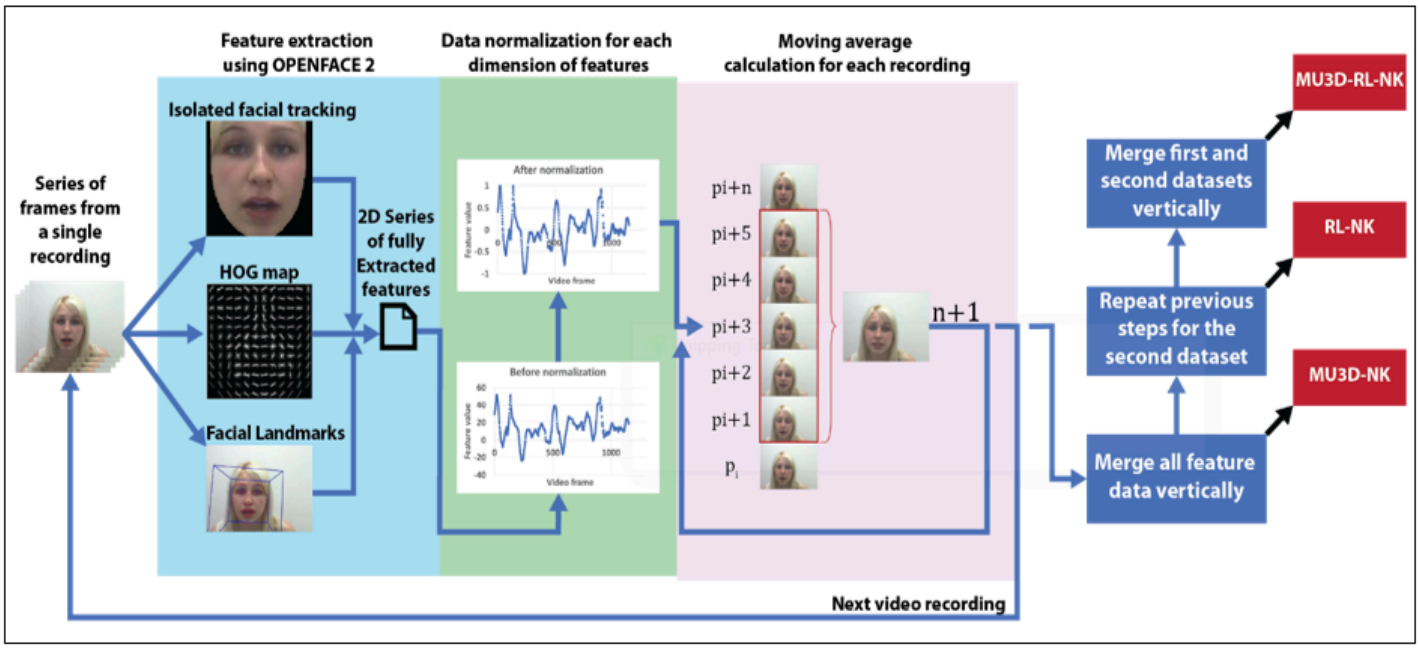

Figure 8

Data processing steps of visual features for a window size of K= 5

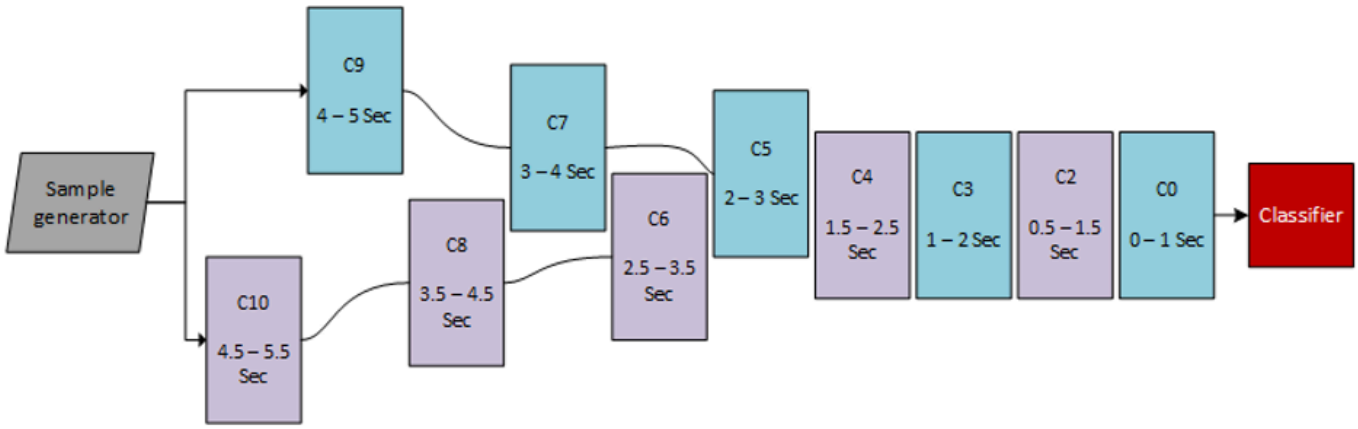

Figure 9

Sample generation and interleaving process

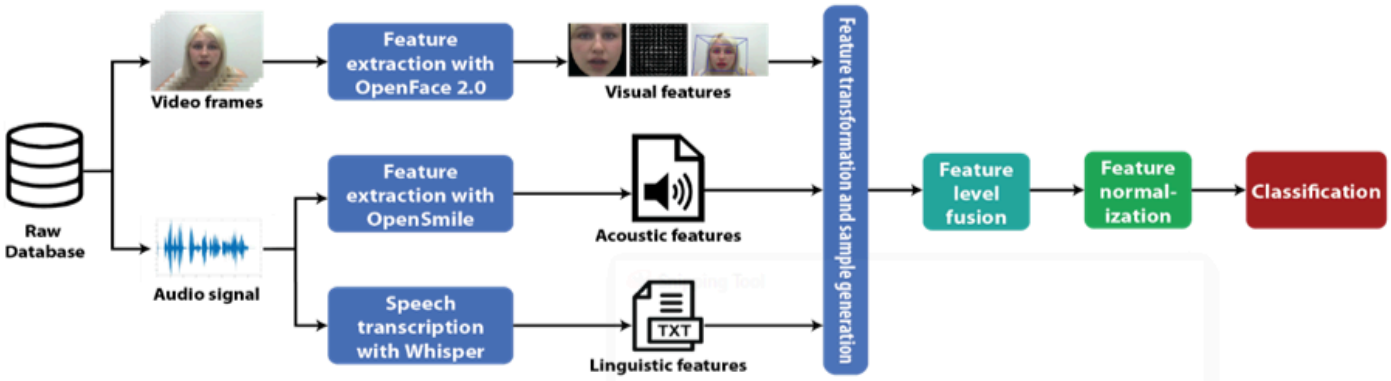

Figure 10

Overall sample generation, processing, and classification steps

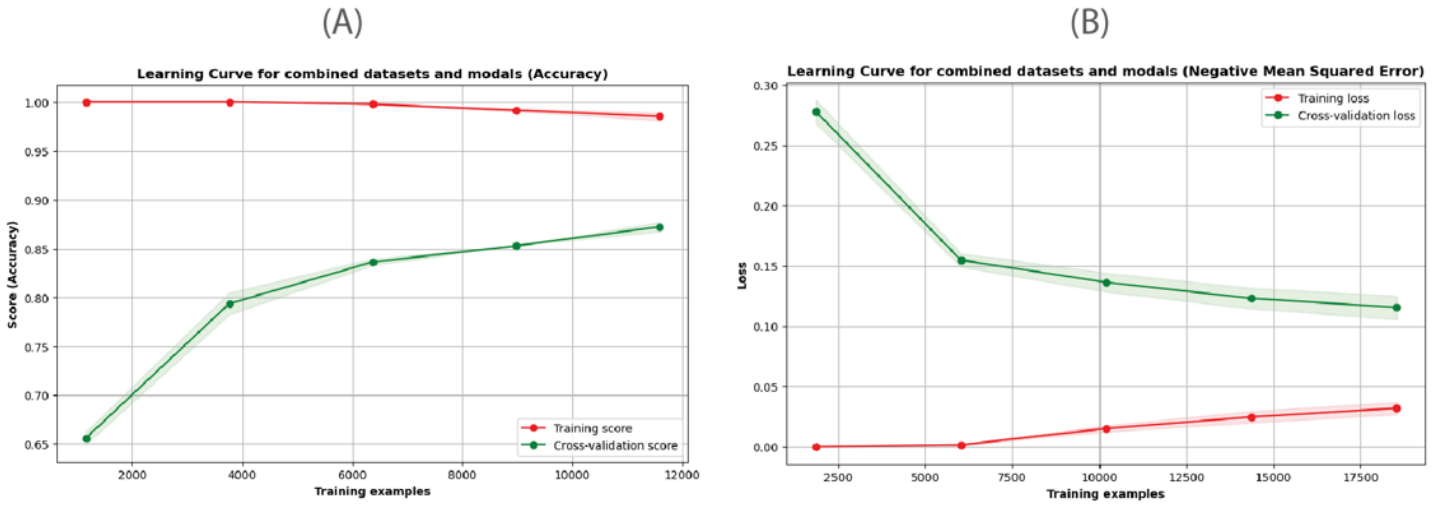

Figure 11

(A): Accuracy learning curve. (B): Negative Mean Squared Error learning curve.

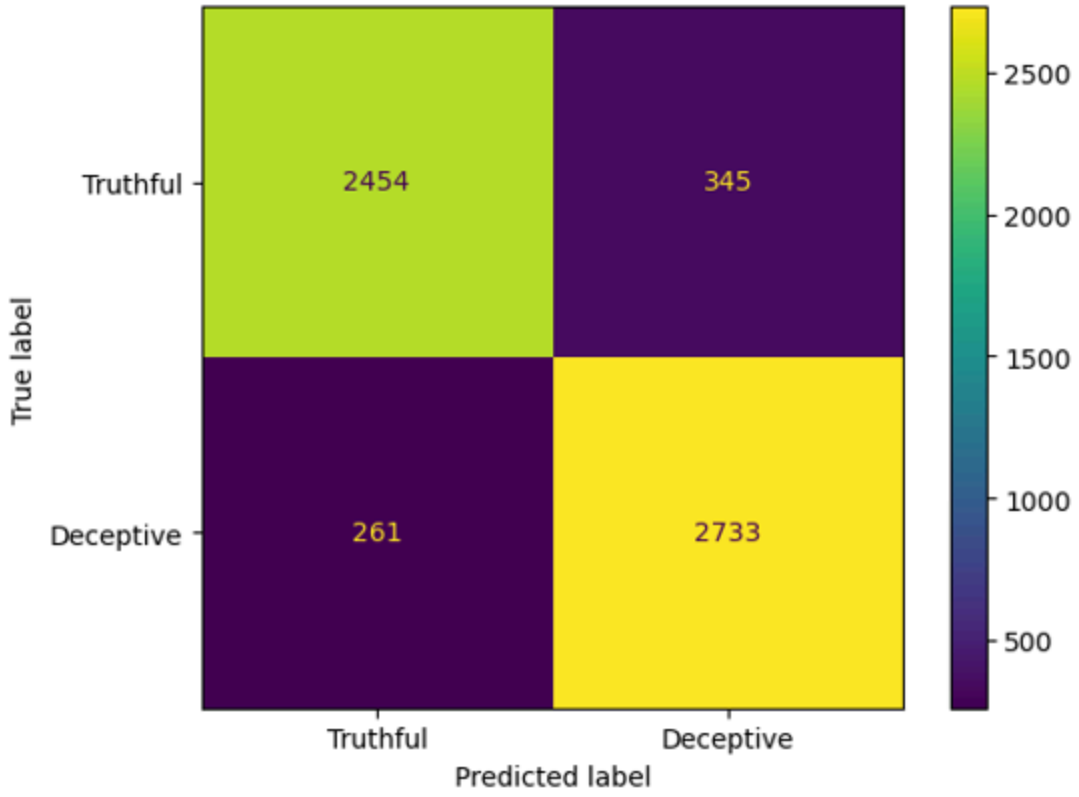

Figure 12

Confusion matrix for combined modals and datasets model

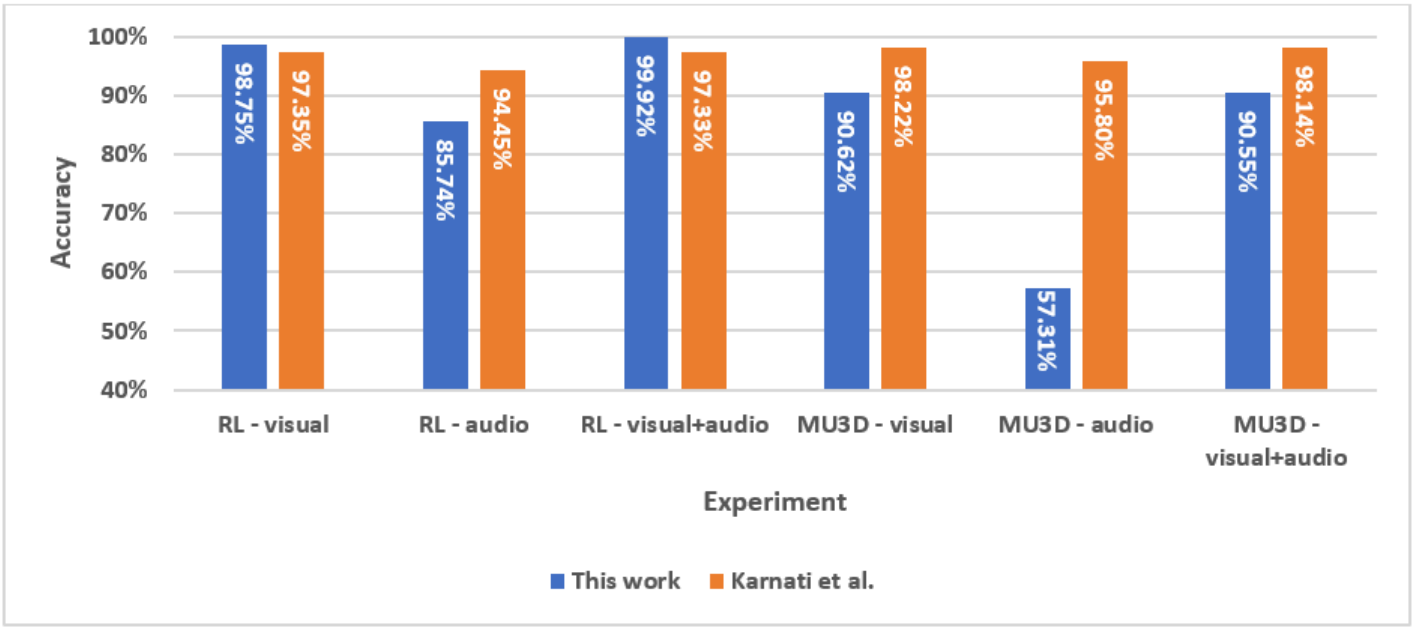

Figure 13

Performance comparison between this work and the work of Karnati et al. in terms of accuracy for matching experiments.

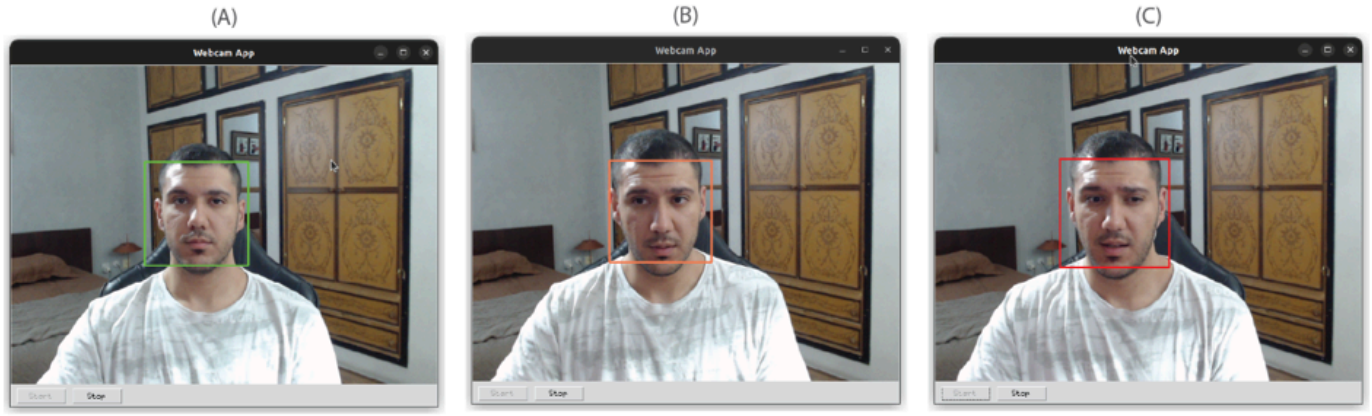

Figure 14

(A): GUI interface when no deception is detected. (B): GUI interface when deception is uncertain. (C): GUI interface when deception is detected

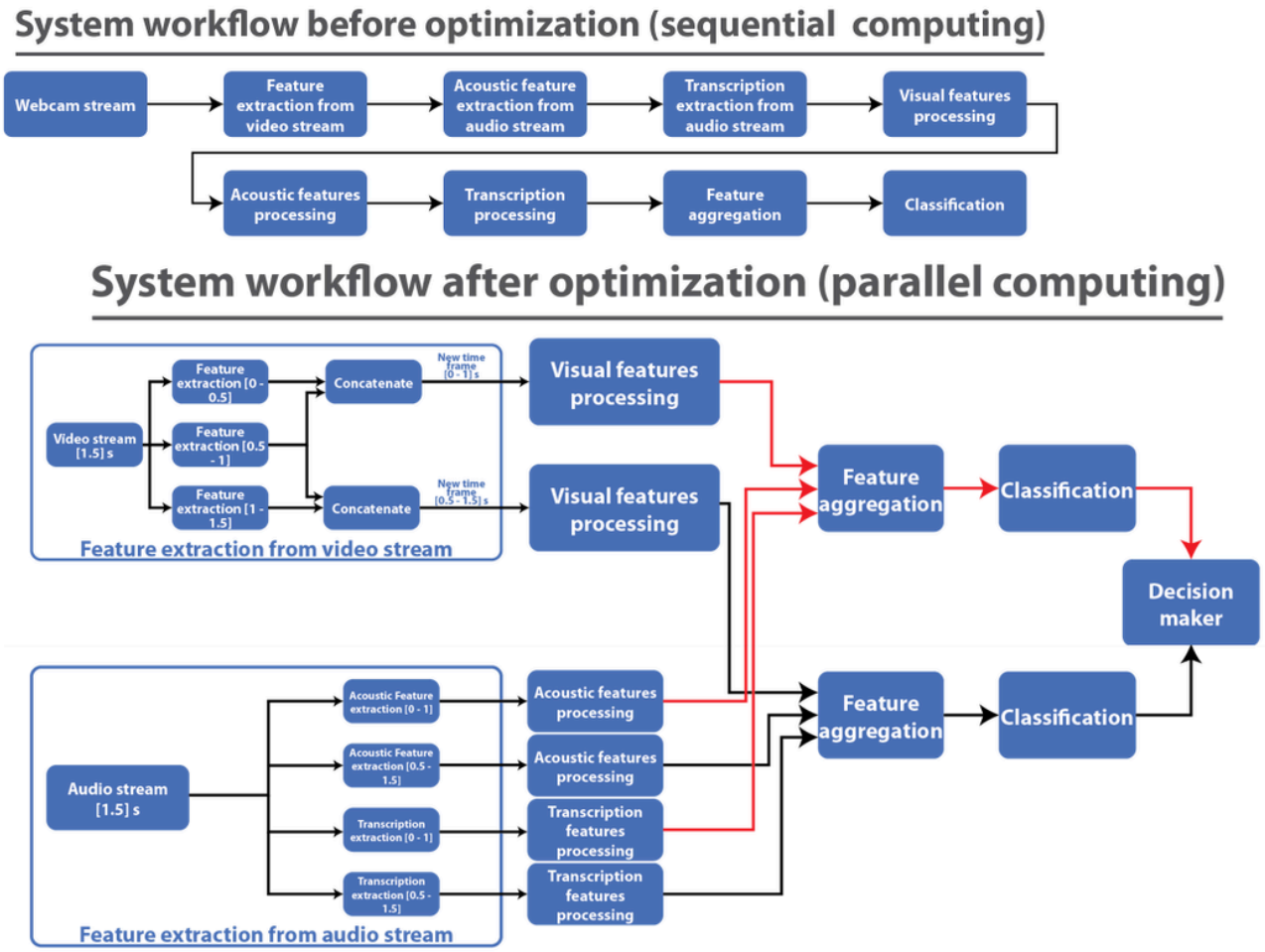

Figure 15

System workflow pre and post optimization