A machine learning approach to detect lie using anyone's voice

Ahsan Rahman	**170104089**
Asif Sorowar	**170104097**
Mohana Rahman	**170104101**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Fall 2020

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2021

A machine learning approach to detect lie using anyone's voice

Submitted by

Ahsan Rahman	**170104089**
Asif Sorowar	**170104097**
Mohana Rahman	**170104101**

Submitted To

Faisal Muhammad Shah, Associate Professor

Farzad Ahmed, Lecturer

Md. Tanvir Rouf Shawon, Lecturer

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2021

ABSTRACT

Detecting deception is very effective in criminal investigations, finding fake news, jurisprudence, law enforcement, and national security. Still a reliable and Non-invasive deception detection technique is in progress. Numerous researches have been done using various modalities, such as video, audio, conversational dialog cues, and text. We have proposed a non-invasive deception detection technique using victim's voice or audio data. After thorough feature analysis of each modality, we have obtained 56.615% cross- validated accuracy using SVM only for audio.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Lie detection refers to the investigative practices used to determine a person's truthfulness and credibility. This is largely determined through the consideration of certain behavioral and physiological cues as well as larger contextual and situational information [1]. Lie detection is a common practice to the law enforcement agency for investigative purpose. The ability for a human to detect deception is very limited. It was reported that average accuracy of detecting lies without special aid is 54% [2]. To aid deception detection more accurately physiological signals are collected. The most widely used technology to detect deception is polygraphy which relies on physiological signals, particularly heart rate/blood pressure, respiration and skin conductivity. Though claims of 90% validity by polygraphy advocates, the UN National Research Council has reported concerns about effectiveness [3,4]. Besides, there are some limitations that also affect the effectiveness of polygraphy. A professional is required to collect the data. Also attaching various sensors to the subject may cause high levels of anxiety in subject in themselves. All these drawbacks provide enough reason to introduce substitute noninvasive technology.

Here we will try to understand, victim's voice is how much effective for lie detection.

Chapter 2

Literature Reviews

For better understanding and keeping up with the latest findings, we reviewed a lot of papers regarding this topic. Here some of the latest research paper on micro expression and deception detection.

Deep Learning Approach for Multimodal Deception Detection

In this paper [5] authors proposed a deep neural network model to detect multiple modal deceptions. They have combined features from different modalities such as video, audio, and text along with Micro-Expression fea- tures, and showed that detecting deception in real-life videos can be more accurate.

In their past research, they have shown that psycholinguistic features based on the Linguistic Inquiry and Word Count (LIWC) lexicon were helpful in detecting deceptive behavior. But polygraph tested measure physiological features such as heart rate, respiration rate, skin temperature of the sub- ject under investigation tests were not reliable and often misleading, since judgments made by humans were often biased. They found that facial expressions and hand gestures were to be very helpful in detecting deceptive nature. For the multimodal, they extracted textual, audio, and visual features sepa- rately. 3D-CNN not only extracts features from each image frame, but also extracts spatio-temporal features from the whole video which helps in iden- tifying facial expressions such as smile, fear, or stress. For extracting Textual Features they have used Convolutional Neural Networks (CNN). OpenSMILE is used for extracting Audio Feature. Before extracting the features, they have made sure that there were no unnecessary signals in the audio that af- fects the quality of the extracted features. To remove the background noise, they have used the audio processing tool SoX (Sound Xchange). Their model obtained 97% of accuracy for all modal features.

Deep Learning Driven Multimodal Fusion For Automated Deception Detection

In this paper [6] authors extended the work that presented in previously. They have fused

audio, visual, and textual cues and have introduced a deep learning driven multimodal fusion (Early and Late) for automated deception detection. Where, Early Fusion is Feature level and Late Fusion is Decision Level.

For extracting Visual Feature they have used 3D-CNN. 3D-CNN was used due to it's inherent ability to extract both spatial (intra frame) as well as temporal (inter frame/contextual) features from video.

Text-CNN have used to extract features from textual modality. They have used openSMILE for extracting audio features, such as pitch and voice inten- sity . The features were extracted at frame rate of 30Hz and 100ms sliding window. They have used a Real-life Trial corpus dataset.The dataset con- tains 121 real-life trial videos (61 deceptive and 60 truthful). The medium of conversation in all videos is English.

After the simulation results and critical performance analysis of the proposed

unimodal deception detection models they showed that: (1) the audio based deception detection model achieved the prediction accuracy of 87.5% (2) the automated extracted textual cues based deep CNN approach achieved the prediction accuracy of 83.78% as compared to the prediction accuracy of 60.33% presented in (3) the visual based 3D deep CNN achieved the ac- curacy of 78.57% as compared to the 76% accuracy in manual annotation based approach. The simulation results of their proposed multimodal early and late fusion approaches, incorporating audio, visual, and textual features achieved the highest accuracy of 92% and 96% as compared to the predic- tion accuracy of 82% achieved by the state-of-the-art approach, where only visual and textual cues were considered.

An Acoustic Automated Lie Detector

In this paper [7] authors focused on acoustical features such as MFCC, en- ergy, and pitch contour information to build machine learning models to automate lie detection. The dataset consists of speech data from an interac- tive lying game contains 498 truth recordings and 434 lie recordings.Three different features(MFCC, energy, and pitch) were extracted from each au- dio file.MFCCs are extracted by using Librosa (librosa.feature.mfcc), with 20 MFCCs generated per audio file, each MFCC padded with zeros to a fixed length of 1000.Energy envelopes were calculated as the sum of the square of the magnitude of the speech signals.Pitch contours also are generated with Librosa.

After extracting features from the dataset, various models are trained and tested on either MFCCs, energy, or pitch. The models were implemented using scikit-learn's library for logistic regression (LR), decision tree classifier (DTC), random forest (RF), gradient boosting classifier(GBC), linear ker- nel support vector machine (SVM), long short-term memory (LSTM) ,and stochastic gradient descent classifier(SGD).To improve the accuracy and F1 scores of the models in this research, majority voting ensemble learning are

applied for a combination of the best-performing classifiers trained on either MFCCs, energy and pitch contours.

The highest-performing model for MFCC was the LSTM-based sequential model, with an accuracy of 54.6% which was greater than even the ensemble classifier results. The worst model trained on MFCCs was the SGD classifiers, which had the lowest accuracy. Ensembling for MFCC did not increase model performance.For pitch, the highest-performing model in terms of accuracy was the RF classifier, at 53.9%. The worst model for pitch was again the SGD classifier, with an accuracy of 49.6%. Pitch-trained models were thus the worst-performing models overall, compared to MFCC and energy.For En- ergy ensembling, increases model performance with an accuracy of 55.8

Deception Detection From Speech Signals

In this paper [8] authors presented an effective method to detect deception by using only speech signals, which extracted from the video clips. Dataset consists of 121 videos including 61 deceptive and 60 truthful trial clips.The dataset called “Real Life Trial Data” because it is created from court records where deceptive and truth statements can be classify easily.

All speech signals were normalized between -1 and 1 interval. Then, the Hamming windowing method was applied to each normalized signals as a FIR filter.After, the Hamming Windowing method is applied to each signal. Discrete Wavelet Transform(DWT) is used to obtain time-frequency features of the speech signals. Each speech is labeled as truthful (1) or deceptive (0).To reduce the size of the feature set, some statistical parameters such as standard deviation, energy, entropy, kurtosis and skewness of the decom- posed signals are obtained.Then getting the feature set and Extreme Learn- ing Machine(ELM), a fast and effective classifier is employed to classify de- ceptive and truthful voices.

Obtain 91.66% performance ratio when 24 speech examples were tested. This results show that the performance of ELM classifier decreases to 54.16%.The final results obtained after a lot of trials show an accuracy above 95

A Multi-View Learning Approach To Deception Detection

In this paper [9] authors applied for the very first time a multi-view learn- ing (MVL) approach to detect deception where each view corresponds to a feature group and focused on analyzing nonverbal behavior likes face-based low level, not hand crafted features.A Multimodal data set consists of 121 recordings(61 deceptive and 60 truthful) in real-life court trials and TV in- terviews for depicting deception.

Facial features extracted by using pre-trained Deep Neural Networks (DNNs) using VGG-Face, Alex Net, Google Net and Res Net and extracted deep face features are combined with gaze, head movements, hand gestures based nonverbal features and also with the verbal features .The MUMIN coding scheme is intended as a tool for studying hand gestures and

facial displays in interpersonal interactions. The Open Face tool is used to extract AUs(Facial Action Units) and extraction of AUs is frame-wise and combined into a single vector.

Video and audio-based nonverbal features and verbal feature are extracted and concatenated in a single vector and build a model employ by different classifiers(SVM, LMKL,MVL).MVL performs better than the other classifiers(SVM, LMKL) in deep face features.Deep face features improves the deception detection results not only when they are combined with other verbal/non- verbal features and MVL applied, but also when they are used alone with a single kernel learning method. The best accuracy is 98% obtained by Face- AlexNet- FINE-TUNING Others.Lexical features and audio-based features leading to the lowest classification accuracy.

Multimodal Deception Detection using Real-Life Trial Data

In this paper [10] authors studied about multimodal system that detects de- ception in real-life trial data using verbal, acoustic and visual modalities. The dataset consists of video clips obtained from real court trials and collected from trials with different outcomes: guilty verdict, non-guilty verdict, and exoneration. For guilty verdicts, deceptive clips are collected from a defen- dant during a trial and truthful videos are collected from witnesses within the same trial.

Aim to detect deception at the topic -level for that only used the trial out- comes to point the subject as deceptive or not.A subject-level deception de- tection system are often evaluated fairly, by comparing its predictions to the subject-level ground-truth, which is that the trial outcome, with the thought that the trial outcome is correct. To avoid annotation bias show the modali- ties within the subsequent order: first show either Text or Silent video, then show Audio, followed by Full video. Annotators labeled all 121 video clips in our dataset, which has 59 different instances. Apply majority voting to the labels assigned by each annotator over all the clips belonging to an equiva- lent subject to calculate the agreement at the subject-level.

The transcriptions are obtained using Amazon Mechanical Turk within the original dataset in video clips where multiple speakers are portrayed. Vi- sual Behavior Annotations were conducted using the MUMIN multi-modal scheme, which incorporates several different facial expressions related to overall facial expressions, eyebrows, eyes and mouth movements, gaze direc- tion, also as head and hand movements. Used the Open Face library with the default multi-person detection model to get the facial action (AUs) units. To extract the linguistic features applied Automatic Speech Recognition (ASR) to the videos using the Google Cloud Speech API and obtained the corre- sponding transcriptions.

Among the various classifiers, the RF classifier is the best classifier for lin- guistic and acoustic features, while the NN performs best with the visual features. NN classifier outperformed with an accuracy of 84.18% on the in- dividual or combined sets of verbal and non-verbal

features and showed that a system using score level combination can detect deceptive subjects.

Robust Algorithm for Multimodal Deception Detection

In this paper [11] authors proposed a multimodal deception detection frame- work is based on combining the decision from the audio, text and nonverbal features using majority voting. The dataset consists of 121 videos, where 61 are deceptive and 60 are truthful video clips, reflecting a real-life question and answer scenario and also provides the manual labels of the 39 different verbal and nonverbal cues.

The 3D based visual features, text features and extracting saliency features from the face, and end-to-end features are extracted using DCNN (Deep Con- volutional Neural Network) and the audio features are extracted using Open Smile. The visual features supported the 2D appearance models are wont to capture the micro movements from blink , eyebrow motion, wrinkle occur- rence and mouth motion.

The audio modality evaluated by three different features (Mel Frequency Cepstral Coefficients (MFCC), Log-Energy of MFCC, and CC(Cepstral Coeffi- cients)) with three different classification schemes (linear SVM, SRKDA(Spectral Regression Kernel Discriminant Analysis), and Long Short Term Memory). Among three different features sets and classifiers the best performance is obtained with the CC-SRKDA where CCR(Correct Classification Rate) is 76%. The text modality evaluated by five different algorithms(LSTM, SRKDA, SVM, Bag-of-words - SVM , Bag-of-N-Grams - SVM) and the Bag-of-N Grams (BoNG) features together with the linear SVM classifier showed the best performance with CCR is 84%. The Micro-expressions is tested with three different clas-

sifiers (AdaBoost ,Random Forest ,SVM) and the AdaBoost classifier gave the best performance with CCR is 88%. The final result obtained by using majority voting applied the features combining by the three independent modalities and show the best performance with CCR is 97%.

2.1 Overall Review

Table 2.1: Literature Review of Research Papers

Title	Methodology of detection	Accuracy/Result
		Continued on next page

Table 2.1 – continued from previous page

Title	Methodology of detection	Accuracy/Result
A Deep Learning Approach for Multimodal Deception Detection	Combined features from different modalities such as video, audio, and text along with Micro-Expression features.	Obtained 97% of accuracy after extracting textual, audio, and visual features.
Deep Learning Driven Multimodal Fusion For Automated Deception Detection	Proposed early and late fusion model which combines audio, visual, and textual features. Between early and late fusion, late fusion performs better.	Obtained 92% (early) and 96% (late) of accuracy after combining audio, video and visual features(A + V + T).
An Acoustic Automated Lie Detector	Various models are trained and tested on either MFCCs, energy, or pitch. The models are implemented using scikit-learn's library for LR,DTC,RF, GBC, SVM, LSTM ,and SGD.	The highest-performing model for MFCC was the LSTM-based sequential model, with an accuracy of 54.6%.
Deception Detection From Speech Signals	Statistical parameters such as standard deviation, energy, entropy, kurtosis and skewness of the decomposed signals are obtained.Then Extreme Learning Machine(ELM) classifier is employed.	The final results obtained after a lot of trials show an accuracy above 95%.
A Multi-View Learning Approach To Deception Detection	MVL applied in Deep face features performs better than the other classifiers(SVM, LMKL)	The best accuracy is 98% obtained by Face-AlexNet-FINE-TUNING and Others.
Multimodal Deception Detection using Real-Life Trial Data	Applied RF and NN classifiers. RF classifier is best for linguistic and acoustic features and the NN performs best with the visual features	The individual or combined sets of verbal and non-verbal features show the best accuracy is 84.18%

Continued on next page

Table 2.1 – continued from previous page

Title	Methodology of detection	Accuracy/Result
Robust Algorithm for Multimodal Deception Detection	CC-SRKDA is the best for evaluating audio.Bag-of-N Grams features with the LSVM classifier shows the best performance for text.The AdaBoost classifier gives the best performance for evaluating non verbal features	Use majority voting for the features combining by the three independent modalities(audio, text and non verbal) show the best performance is 97%.

Most of the dataset used in deception detection research are of specific environments such as courtroom trial data, mock crime scenarios and game shows.So,obtained accuracy becomes higher than casual environment dataset.

Chapter 3

Data Collection & Processing

3.1 Dataset

To assess the performance of any detection approach requires experimenta- tion with data that heterogeneous enough to simulate real life scenario to an acceptable level. Generality, Realism and Representativeness are the key factor to select the datasets.

Bag of Lies Dataset

Bag-of-Lies [5] is a multi-modal dataset consisting of video, audio and eye gaze from 35 unique subjects collected using a carefully designed experi- ment for the task of automated deception detection (binary classification into truth/lie). It has a total of 325 manually annotated recordings consisting of 162 lies and 163 truths. For the experiment, each subject was shown 6-10 select images and were asked to describe them deceptively or other- wise based on their choice. Video and Audio were captured using a standard phone camera and microphone. We are using 80% of the dataset as training set and other 20% as test set. For every video authors provided a csv. We will convert the video to audio.wav for feature extraction. They also provided a finalized annotation file where they manually annotated 1(truth) or 0(lie) for each run of each subject. Finalized annotation dataset sample is shown in 3.1.

A	B	C	D	E	F	G	H	I	J
eeg	gaze	image	run	truth	usernum	video			
./Finalised/	./Finalised/	./Finalised/	0	1	0	./Finalised/User_0/run_0/video.mp4			
./Finalised/	./Finalised/	./Finalised/	1	0	0	./Finalised/User_0/run_1/video.mp4			
./Finalised/	./Finalised/	./Finalised/	2	0	0	./Finalised/User_0/run_2/video.mp4			
./Finalised/	./Finalised/	./Finalised/	3	0	0	./Finalised/User_0/run_3/video.mp4			
./Finalised/	./Finalised/	./Finalised/	4	0	0	./Finalised/User_0/run_4/video.mp4			
./Finalised/	./Finalised/	./Finalised/	5	1	0	./Finalised/User_0/run_5/video.mp4			
./Finalised/	./Finalised/	./Finalised/	0	1	1	./Finalised/User_1/run_0/video.mp4			
./Finalised/	./Finalised/	./Finalised/	1	0	1	./Finalised/User_1/run_1/video.mp4			
./Finalised/	./Finalised/	./Finalised/	2	1	1	./Finalised/User_1/run_2/video.mp4			
./Finalised/	./Finalised/	./Finalised/	3	1	1	./Finalised/User_1/run_3/video.mp4			
./Finalised/	./Finalised/	./Finalised/	4	0	1	./Finalised/User_1/run_4/video.mp4			
./Finalised/	./Finalised/	./Finalised/	5	1	1	./Finalised/User_1/run_5/video.mp4			
./Finalised/	./Finalised/	./Finalised/	0	1	2	./Finalised/User_2/run_0/video.mp4			
./Finalised/	./Finalised/	./Finalised/	1	0	2	./Finalised/User_2/run_1/video.mp4			
./Finalised/	./Finalised/	./Finalised/	2	0	2	./Finalised/User_2/run_2/video.mp4			
./Finalised/	./Finalised/	./Finalised/	0	0	3	./Finalised/User_3/run_0/video.mp4			
./Finalised/	./Finalised/	./Finalised/	1	0	3	./Finalised/User_3/run_1/video.mp4			
./Finalised/	./Finalised/	./Finalised/	2	1	3	./Finalised/User_3/run_2/video.mp4			
./Finalised/	./Finalised/	./Finalised/	3	1	3	./Finalised/User_3/run_3/video.mp4			
./Finalised/	./Finalised/	./Finalised/	4	0	3	./Finalised/User_3/run_4/video.mp4			
./Finalised/	./Finalised/	./Finalised/	5	1	3	./Finalised/User_3/run_5/video.mp4			
./Finalised/	./Finalised/	./Finalised/	6	0	3	./Finalised/User_3/run_6/video.mp4			
./Finalised/	./Finalised/	./Finalised/	7	1	3	./Finalised/User_3/run_7/video.mp4			

Figure 3.1: Finalised Annotation Dataset

3.2 OpenSMILE

OpenSMILE (open-source Speech and Music Interpretation by Large-space Extraction) [12] is an open-source toolkit for audio feature extraction and classification of speech and music signals. OpenSMILE is widely applied in automatic emotion recognition for affective computing.It has speech related features such as MFCC, Mel Frequency, Linear spectral coefficient, Loudness, Voicing, Fundamental frequency envelope, Pitch, Jitter, DDP jitter, Shimmer, Pitch onsets, duration.

3.3 Feature Extraction

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that re-

quire a lot of computing resources to process. Feature extraction is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set.

3.4 Feature Selection

Feature selection is the process of selecting relevant features based on various methods.It reduced those features which has much less influence on the classification problem. It is used to decrease the over-fitting problem and is also important to overcome the imbalance data-set problem.

3.5 Feature Scaling

Feature scaling is used to normalize the range of independent variables or features of data.We need to scale value of features and provide equal weight to all features in order to obtain this same scale for all data.

Chapter 4

Methodology

4.1 Feature Extraction

There are a variety of methods for extracting audio.openSMILE [12] is a free and open-source toolkit for extracting high-dimensional audio features. We have used openSMILE to extract characteristics from the input audio in this study. The following (audio-specific) low-level descriptors can be computed by openSMILE:

- Frame Energy
- Frame Intensity / Loudness (approximation)
- Critical Band spectra (Mel/Bark/Octave, triangular masking filters)
- Mel/Bark-Frequency-Cepstral Coefficients (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 (LSP, aka. LSF)
- Fundamental Frequency (via ACF/Cepstrum method and via Subharmonic- Summation (SHS))
- Probability of Voicing from ACF and SHS spectrum peak
- Voice-Quality: Jitter and Shimmer

• Formant frequencies and bandwidths

• Zero and Mean Crossing rate

• Spectral features (arbitrary band energies, roll-off points, centroid, entropy, maxpos, minpos, variance (= spread), skewness, kurtosis, slope)

• Psychoacoustic sharpness, spectral harmonicity

• CHROMA (octave-warped semitone spectra) and CENS features (energy- normalised and smoothed CHROMA)

• CHROMA-derived features for Chord and Key recognition

• F0 Harmonics ratios

There are six feature sets in openSMILE. They are:

• ComParE_2016

• GeMAPSv01a

• GeMAPSv01b

• eGeMAPSv01a

• eGeMAPSv01b

• eGeMAPSv02

Each feature set can be extracted on two levels:

• Low-level descriptors (LDD)

• Functionals

4.2 Feature Selection

Here we used ComParE_2016 and eGeMAPSv02 feature sets. And both Low- level descriptors and Functionals feature levels.

The discovered features when utilizing the ComParE_2016 feature set and the Low-level descriptors feature level were divided into four segments. (1) only MFCC (2) only Audio Spectrum (3) MFCC and Audio Spectrum features (4) All features.

When utilizing the eGeMAPSv02 feature set and Functionals feature level, we discovered too many extra features. For deception detection, we need to extract the features as quickly as possible. It will take longer if there are too many features that need to be extracted from audio. As a result, we'll need some features that can be extracted rapidly and give us better

accuracy. While researching many other papers [8,9], we found that MFCC features are the most significant features for deception detection. So we basically took MFCC features from the Functionals feature level. The features of the eGeMAPSv02 feature set and Low-level descriptors were divided into five categories. (1) Only MFCC, (2) Frequency, Bandwidth, and Amplitude, (3) Frequency and Bandwidth, (4) Amplitude and Bandwidth, and (5) All Features.

4.3 Classification

The mean, standard deviation, and median values from both ComParE_2016 and eGeMAPSv02 were used to train machine learning algorithms, where algorithms were Logistic Regression and SVM. In the SVM algorithm, we used different kernels like Linear, Polynomial, Gaussian Radial Basis Function (RBF) and Sigmoid.

Chapter 5

Experiments and Results

Using the mean, standard deviation, and median values from ComParE_2016 features separately in machine learning techniques, we discovered the following accuracy in 5.1 and 5.2

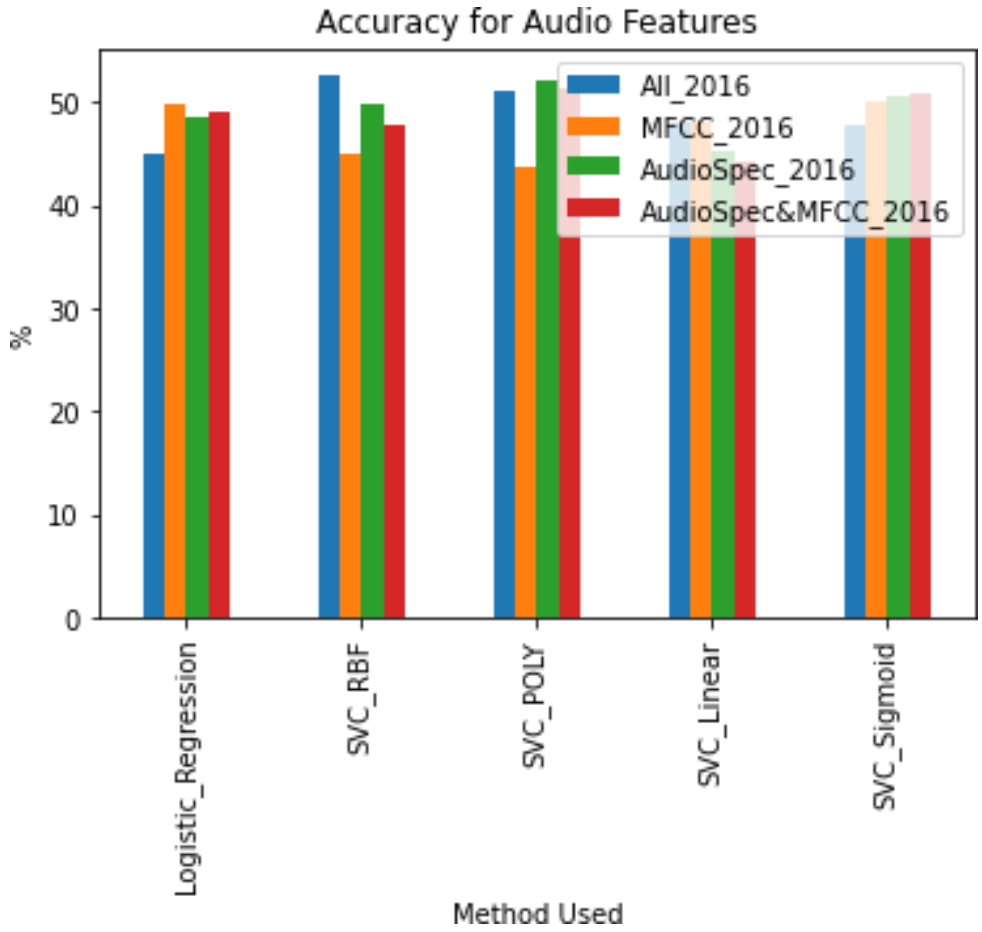

Figure 5.1: Bar plot when the ComParE_2016 feature set is used

index	All_2016	MFCC_2016	AudioSpec_2016	AudioSpec&MFCC_2016
Logistic_Regression	44.923077	49.846154	48.615385	48.923077
SVC_RBF	52.615385	44.923077	49.846154	47.692308
SVC_POLY	51.076923	43.692308	52.000000	51.384615
SVC_Linear	48.000000	48.000000	45.230769	44.307692
SVC_Sigmoid	47.692308	50.153846	50.461538	50.769231

Figure 5.2: Accuracy when the ComParE_2016 feature set is used

Here using just Audio Spectrum characteristics from feature set ComParE_2016, feature level Low-level descriptors and applying them to the SVC model, we were able to get a height accuracy of 52%, while the kernel was Polynomial.

Using the mean, standard deviation, and median values from eGeMAPSv02 features separately in machine learning techniques, we discovered the fol- lowing accuracy in

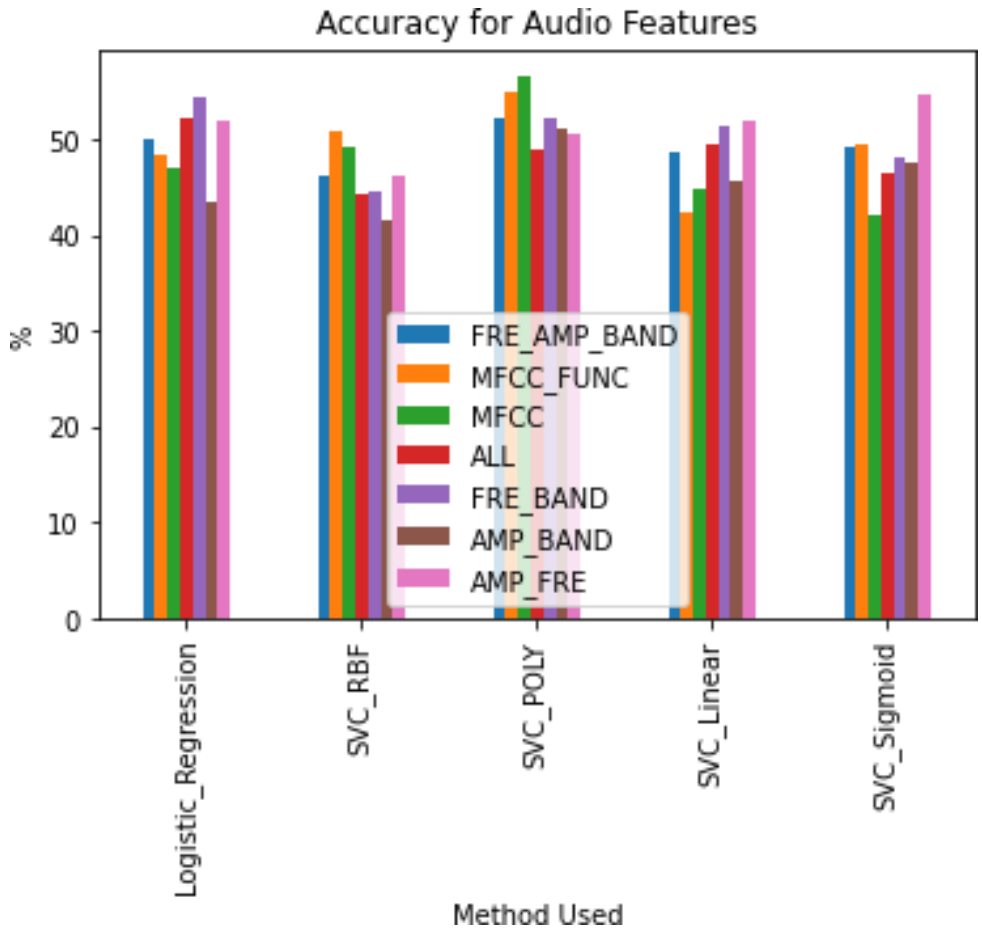

Figure 5.3: Bar plot when the eGeMAPSv02 feature set is used

index	FRE_AMP_BAND	MFCC_FUNC	MFCC	ALL	FRE_BAND	AMP_BAND	AMP_FRE
Logistic_Regression	50.153846	48.307692	47.076923	52.307692	54.461538	43.384615	52.000000
SVC_RBF	46.153846	50.769231	49.230769	44.307692	44.615385	41.538462	46.153846
SVC_POLY	52.307692	55.076923	56.615385	48.923077	52.307692	51.076923	50.461538
SVC_Linear	48.615385	42.461538	44.923077	49.538462	51.384615	45.538462	52.000000
SVC_Sigmoid	49.230769	49.538462	42.153846	46.461538	48.000000	47.692308	54.769231

Figure 5.4: Accuracy when the eGeMAPSv02 feature set is used

Here using just MFCC characteristics from feature set eGeMAPSv02, feature level Low-level descriptors and applying them to the SVC model, we were able to achieve a height accuracy of roughly 57%, while the kernel was Polynomial.

After comparing the height accuracy of the two feature sets, we can con- clude that using just audio for deception detection is not a good idea. For higher accuracy, we need to integrate other features. However, we can also state that audio's MFCC features can be used for deception detection.

Chapter 6

Future Work and Conclusion

Detecting Deception is a crucial part in investigation. Current Method re- quire a well setup run by a expert professional, which is not possible for most organization. Cost, Reliability, Complexity and lot of factor is preventing its widespread use among right people. The goal of this thesis is to introduce a easier and a non-invasive way to detect deception using video, audio, gaze and EEG through various machine learning algorithm. Improving detection rate and reducing false positive rate will be prioritized.

In future we will perform ablation study. We will also tune hyperparameters for different models including Deep learning model using different feature sets and then we will make comparison with our base machine learning model.

References

[1] "What is deception detection?." https://www.paulekman.com/deception/deception-detection/#:~:text=Deception%20detection%20refers%20to%20the,larger%20contextual%20and%20situational%20information.

[2] P. Ekman and M. O'Sullivan, "Who can catch a liar?," *American psychologist*, vol. 46, no. 9, p. 913, 1991.

[3] G. Jacobs, "The physiology of mind–body interactions: The stress response and the relaxation response," *Journal of alternative and complementary medicine (New York, N.Y.)*, vol. 7 Suppl 1, pp. S83–92, 02 2001.

[4] S. L. Happy and A. Routray, *Recognizing Subtle Micro-Facial Expressions using Fuzzy Histogram of Optical Flow Orientations and Feature Selection Methods*. 01 2018.

[5] H. Zhang, L. Feng, N. Li, Z. Jin, and L. Cao, "Video-based stress detection through deep learning," *Sensors*, vol. 20, no. 19, 2020.

[6] M. Gogate, A. Adeel, and A. Hussain, "Deep learning driven multimodal fusion for automated deception detection," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–6, 2017.

[7] S. Ekici, T. Kavas, Y. Akbulut, and A. Sengur, "Deception detection from speech signals," 05 2017.

[8] 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)*, pp. 599–606, 2018.

[9] "An Acoustic Automated Lie Detector." https://www.cs.princeton.edu/sites/default/files/alice_xue_spring_2019.pdf.

[10] A. C. Merzagora, S. Bunce, M. Izzetoglu, and B. Onaral, "Wavelet analysis for eeg feature extraction in deception detection," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2434–2437, 2006.

[11] A. Turnip, M. Amri, H. Fakhrurroja, A. Simbolon, M. A. Suhendra, and D. Kusumandari, "Deception detection of eeg-p300 component classified by svm method," pp. 299–303, 02 2017.

[12] U. M. Sen, V. Perez-Rosas, B. Yanikoglu, M. Abouelenien, M. Burzo, and R. Mihalcea, "Multimodal deception detection using real-life trial data," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Monday 20th September, 2021 at 7:41pm.