# Multimodal Deception Detection using Deep Learning Techniques

Adrika Mukherjee
University of Stuttgart
Stuttgart, Germany
St164609@stud.uni-stuttgart.de

Jad Maalouly
University of Stuttgart
Stuttgart, Germany
st160254@stud.uni-stuttgart.de

## ABSTRACT

Deception Detection is not only a persistent problem but also extremely unattended. We encounter deception every day in our lives. The human perception for lying is no better than a toss of a coin. This topic holds a great momentous considering the magnitude and severity of its consequences. It is a specialized domain under Emotion Analysis but is avoided by the Scientific community primarily due to scarcity of unbiased data. In this paper we explore deception detection using three different modalities i.e. gaze, micro-expressions, and audio. We check how these three modalities will perform on data accumulated from different sources like CSC Deceptive Speech[7],youtube,BagOfLies[8],and Real-life Trial[1] data. We build a deep learning network which handles the data shortage for individual modalities effeciently. Then we build our model based on all possible combination of the three modalities. Finally, we compare how our model hold against traditional Machine Learning techniques. Our project can be found at https://github.com/AdrikaMukherjee/Deception-Detection/

## CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Computer vision**; **Machine learning approaches**.

## KEYWORDS

Deception detection, multi-modal, neural networks

**ACM Reference Format:**
Adrika Mukherjee and Jad Maalouly. 2020. Multimodal Deception Detection using Deep Learning Techniques. In *Proceedings of Fachpraktikum Interaktive Systeme (FIS'20).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Deception has a wide spectrum ranging from fun light games to high-stake trials. Humans are not well equipped for such detections since this is more of a psychological game. In most cases the most popular method to detect a lie is using a polygraph test, but this kind of test has been proven to be more and more unreliable[5]. Moreover, using such equipment can be quite expensive and unnecessary in trivial cases. Therefore, it is important to explore other solutions to approach this problem and make it more accessible for users like a police or a judge who needs to deal with this issue on a daily basis. There is some scientific evidence that states that human expressions do actually change when lying. During deception the body will involuntarily send physical cues. Those cues can be anything from physical gestures to the frequencies in our voices.
While a polygraph test enforces its own environment, demanding skin contact and human expertise, automated methods depend solely on the data collected in normal circumstances. Data acquired in an artificial setting might cause awareness as users will not take the context seriously and the observation will fail to acquire authentic emotional response.
In this paper we implement an automated learning system which takes a video as an input and classifies it as either truthful or deceptive. Different cues are extracted from the input video, like the gaze data (eye gaze vector, eye gaze vector-angle etc), micro-expression data (movement of eyebrows, head pose etc), audio data (MFCC, pitch etc) which is used to train the model.
In the first section of the paper, an idea about the different datasets used and the new data created from different Youtube content is examined. Then we explore the steps to extract the Audio, Gaze and Micro-expression features from the videos using different open source toolkits. We describe a Deep Learning approach to tackle the problem using individual modality and also using all possible combinations of modalities. Finally, we compare our work with existing Classical Machine Learning techniques.

## 2 RELATED WORKS

**In (Zhe Wu et. al., 2018)**[9], Improved dense trajectories(IDT) was used to compute local feature correspondences in consecutive frames which proved to be beneficial for predicting deception. They fused the score of classifiers trained on IDT features and high-level micro-expressions to improve performance. However, there approach was only tested on real-life trial dataset, also they used Classical Machine Learning models for training.
**In (Viresh Gupta et. al.)**[8], the appropriateness of gaze and EEG modalities is studied. A brand new multimodal dataset called Bag-of-Lies[8] is presented in their work. A user study was conducted in a realistic scenario that had 35 unique subjects providing 325 annotated data points with an even distribution of truth (163) and lie (162). They used Machine Learning approach to train their model by the aid of four different modalities, video, audio, EEG and gaze data using on Bag-of-Lies Dataset.
**In (Mingyu Ding et. al., 2018)**[3], cues generated by both face and the body of the subject from real-life trial dataset was taken


*FIS'20, Febraury 2020, Stuttgart, Germany*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

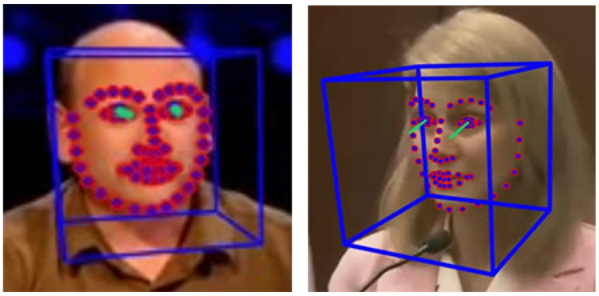

**Figure 1: Screeshots taken after running OpenFace[2] over our datasets.**

into consideration, using correlation learning across the spatial and temporal streams. The novel face-focused cross-stream network addresses the scarcity of data using meta-learning and adversarial learning.

## 3 DATASETS

One of the main challenges was to gather data that was curated in unbiased circumstances. The search finally ended in the following three datasets collected under optimal settings:

**Bag-Of-Lies[8]**: Consisting of 34 participants, where each user is shown around 7 up to 10 images where the user had to describe that image. The user has the freedom to decide whether to tell the truth or to a lie about the image. The dataset in total consists of 325 videos. 165 of these videos are annotated as truth while 161 are annotated with a lie. This user study was conducted in a controlled environment where the users were seated in front of a camera. Allthough, the quality of such a setup might be questionable, as there is no real consequence, if the participant lied or said the truth, nevertheless we used it as we could extract all three modalaties from it.
**Real-Life Trial[1]**: Consisting of 121 short videos, along with their transcriptions and gesture annotations. This dataset includes 50 truthful videos and 54 deceptive videos. These are real trial videos where people giving a testimony in front of a judge or jury. While this is not a controlled environment, this poses an issue for the camera position and angle. However, the people in this dataset give an emotionally aroused response owing the high-stake setting.
**CSC Deceptive Speech dataset[7]**: It consists of 32 hours of audio interviews from 32 native speakers of Standard American English (16 male,16 female) recruited from the Columbia University student population and the community. The subjects were supposed to convince the interviewer that they achieved a high score in a certain exam. For each question from the interviewer, subjects were asked to indicate whether the reply was true or contained any false information by pressing one of two pedals hidden from the interviewer under a table.
Both Bag of lies, Real life Trial Data consists of videos hence both Audio and facial expression, gaze features could be extracted. CSC Deceptive Speech[7] contains only speech data. Again, combining all datapoints did not solve the data shortage issue. This led to the creation of more data from Youtube-dl.

| | Accuracy | ROC AUC |
|---|---|---|
| DL Approach | 0.61 | 0.53 |
| Meta-Learning | 0.652174 | 0.634615 |
| Krishnamurthy et. al.[6] | 0.5231 | - |

**Table 1: Comparison for our uni-modal approaches and the approach of Krishnamurthy et. al.[6]**

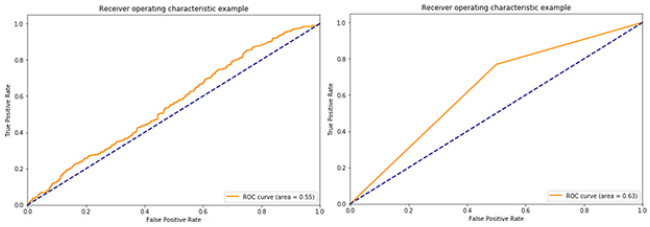


**Figure 2: Left: The ROC curve given our simple DL approach. Right: The ROC curve given by the Siamese network**

**Youtube**: A number of videos from the Fallon show, Split or Steal, 6-1 People where the contestant has to lie or say the truth to win the game was used to curate more data. Bieng of a game show nature, most of the videos is unusable, except a few seconds were the camera is face focused on the participant. These videos have been manually edited and annotated depending on the ultimate revelation of the liar in the game. The Final dataset collected from Youtube consists of 211 videos.

## 4 FEATURE EXTRACTION

We used OpenFace[2] for the extraction of both gaze and micro-expressions features. OpenFace is a tool for facial recognition. It is capable of tracking the head pose in addition to eye and facial landmarks. Openface takes a video as an input and outputs a .csv file with the extracted features from video for each frame. The Audio features are extracted using OpenSmile[4], which is another open source tool widely used for feature extraction. Unlike Openface, Opensmile takes an .wav file as an input and outputs a .arff file with the extracted Audio features.

We will be focusing on three main modalities:

**Gaze**: We select gaze_0_x, gaze_0_y, gaze_0_z which gives eye gaze direction vector in world coordinates for eye 0 (normalized), where eye 0 is the leftmost eye; gaze_1_x, gaze_1_y, gaze_1_z which gives eye gaze direction vector in world coordinates for eye 1 (normalized), where eye 1 is the rightmost eye; gaze_angle_x, gaze_angle_y which is eye gaze direction in radians in world coordinates averaged for both eyes; location of 2Dand 3D eye region landmarks are also extracted.
**Micro-Expression**: Micro-Expression cues accounts to the subtle motion done involuntarily. Facial Action Units (AUs) features like inner brow raiser, cheek raiser, nose wrinkler etc are extracted which indicates human facial expression. Features related to Head

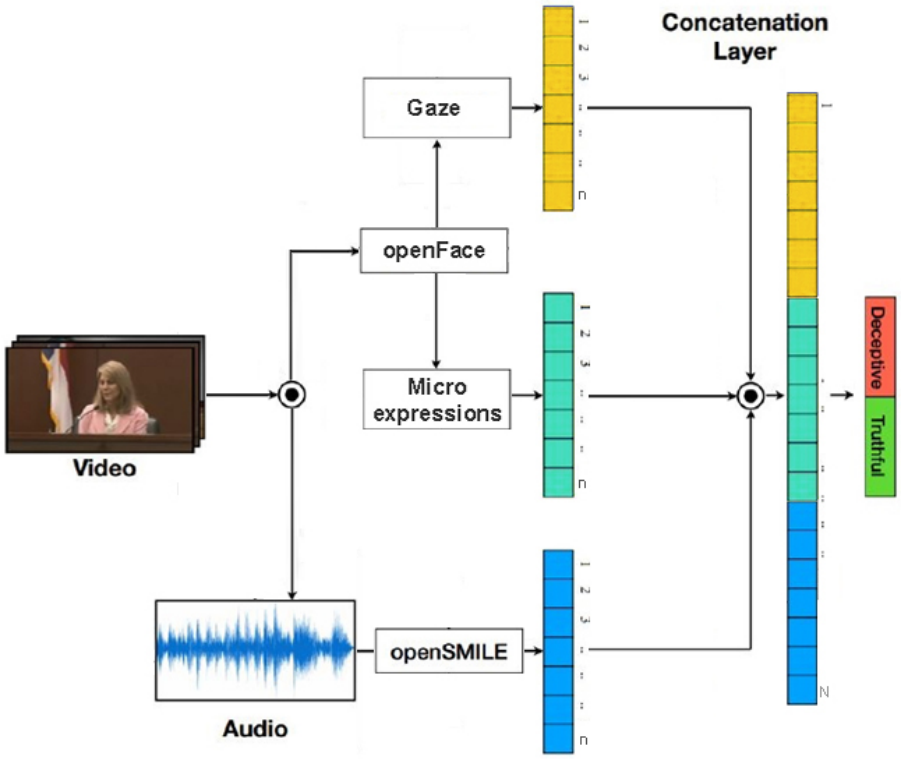


**Figure 3: The final architecture for our multi-modal approach**

pose location is also used.

**Audio**: openSMILE configuration file "emobase.conf" for emotion recognition base set of 988 features, was used to extract frame-wise low-level descriptors such as MFCC, Pitch, Prosody from audio data.

# 5 OUR APPROACH

## 5.1 Pre-processing

Since the data was collected from various sources so the length of the videos are not equal. This obviously led to un-even number of frames per video. Time series data needs careful handling during sampling. In order to achieve a uniform shape for all datapoints, different pre-processing techniques were used. Zero padding and concatenation with the data slice from the last frame was used in the initial model. However, some improvements in results were seen when linear interpolation on the data was done. The optimal number of frames was chosen after statistical analysis of the distribution of the number of frames over full dataset. 75% of data had around 1600 frames so all the datapoints with lesser number of frames was treated by linear interpolation to match up to the target number, and all the datapoints which had more than 1600 frames, the extra frames was discarded. For the multimodal approach we used a different pre-processing technique. Resample is a scipy function which uses Fourier method to resample a given x number of samples to a certain target value(y).The resampled signal starts at the same value as x but is sampled with a spacing of $len(x)/y * (spacingofx)$. As far as the Machine Learning approach is concerned, mean vector for each datapoint is calculated from all the generated frames for a particular video for all the modalities.

## 5.2 Deep Learning Model

For this approach we only considered audio features as the input since we simply did not have enough data for gaze and micro-expressions to train a deep learning model.

The Initial model consisted of two fully connected neural layers with dropouts in between, binary_crossentropy loss and an Adam Optimizer.

The model produced training loss of 0.7532 and training accuracy of 0.6044. The resulting loss kept decreasing while the accuracy stays constant. Test loss and accuracy values are 2.28 and 0.48. This showed that our initial model had a high variance where it trains well but is not able to generalize when introduced to data it has not seen before.

To improve on our initial approach, we added a L2 regularizer and early stopping. The final model use a LSTM layer followed by two simple Dense Layer with L2 regularizer, binary_crossentropy loss and Adam Optimizer.

As a result, our test loss has decreased to 0.855 and the accuracy increased to 0.629

We compared the results of our simple DL model with that of Krishnamurthy et.al. Table 1 shows that our approach, while having an accuracy of 0.62 shows an AUC of 0.54. We did a chi-square goodness test to confirm that the improvement in the performance

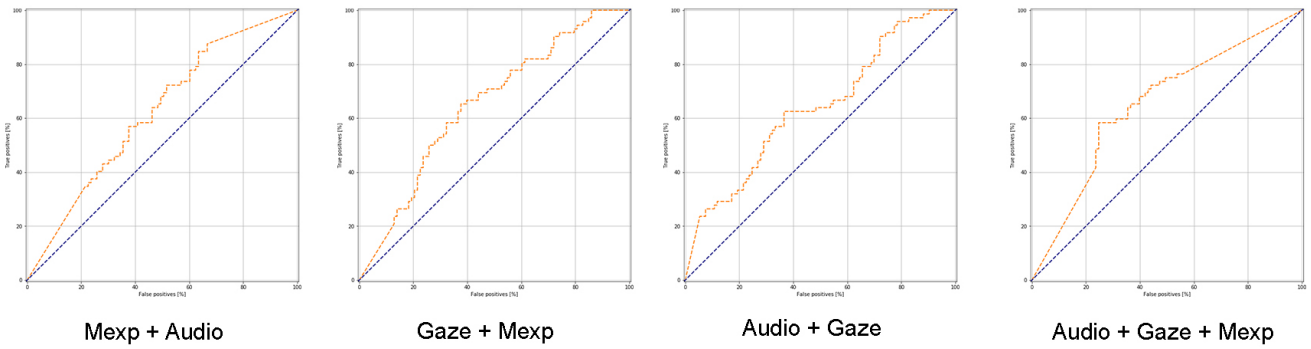


**Figure 4: ROC curves given for all possible combinations of our modalities using the multi-modal DL approach**

of the model are not merely random and are induced by the changes done on the architecture, p-value of $3.258e^{-11}$ obtained from the test establishes this.

### 5.3 Siamese Network

Unless we have sufficient data points, it is futile to use complex, data hungry deep learning models. Humans are able to learn from very less experience, mostly by trying to learn the difference between different objects they see. For example, when a child sees an image of a lion and a tiger, he/she needs very few examples of both the animals in order to be able to correctly differentiate between the two. This idea is used by meta-learning models, we use Siamese networks that consists of two symmetrical Deep Learning networks both sharing the same weights and architecture and both joined together at the end using some energy function E. The objective of our Siamese network is to learn whether the two inputs are similar or dissimilar. As Figure 5 shows, We give two input modality(either audio, gaze or micro-expression data) $X_1$ and $X_2$ and we want to learn whether the two datapoints are similar or dissimilar. The network that process two inputs consists of two fully connected neural network layers with dropouts, this outputs embeddings for the features given in the input, the difference between the embeddings is calculated by the energy function which is Euclidean distance in our case. Contrastive loss, which is a distance-based loss function used for optimizing the parameter values.
As shown in Table 1 we can see the improvement on our ROC AUC being 0.634.

### 5.4 Multimodal Network

The multimodal architecture has three inputs for each of the three modalities. The input will go through two dense layers of size 128 each. Then a concatenation layer comprising features from individual modalities are concatenated back to back, which finally goes through two more dense layers. Using Hamdard product(element wise multiplication) for the concatenation produced comparable results.
Figure 3 shows the architecture of the multimodal approach.
As a result of this approach we were able to achieve an Accuracy 0.666 with ROC AUC of 0.653 when all three modalities(gaze, micro-expression and audio) was combined.

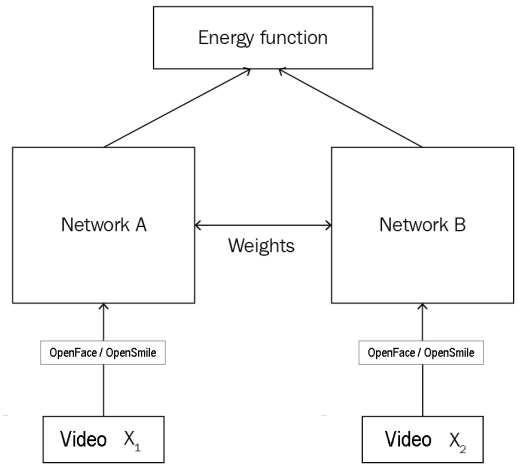


**Figure 5: Siamese network architecture**

## 6 MACHINE LEARNING APPROACH

Now we want to test how the proposed DL model would stand against traditional ML models. Again, we consider gaze, micro-expressions, and audio feature extracted from the raw data.
For the multimodal Deception Detection model, we implemented a Random Forest for both micro-expressions and gaze while using XGBoost for Audio. Then we combine the different scores using Late Fusion where we take a weighted average for each of the resulting hyperparameters:

$$S = \sum_{i=0}^{3} \alpha_i S_i \tag{1}$$

Where $\alpha_i$ are the weights applied for the different scores $S_i$.
Until now, the weights applied to tune our hyper-parameters were manually changed. This process involved a lot of trial and error to get the best possible solution.
Next, we implemented an architecture which voted for the weights

| Modality | RealLifeTrail | | BagOfLies | | Youtube | |
|---|---|---|---|---|---|---|
| | Accuracy | ROC AUC | Accuracy | ROC AUC | Accuracy | ROC AUC |
| Gaze | 0.627 | 0.658 | 0.458 | 0.437 | 0.563 | 0.597 |
| Mexo | 0.587 | 0.634 | 0.440 | 0.439 | 0.654 | 0.711 |
| Audio | 0.570 | 0.576 | 0.510 | 0.552 | 0.563 | 0.558 |
| Gaze + Mexp | 0.600 | 0.633 | 0.538 | 0.559 | 0.790 | 0.819 |
| Gaze + Audio | 0.720 | 0.726 | 0.676 | 0.669 | 0.767 | 0.823 |
| Mexp + Audio | 0.680 | 0.786 | 0.661 | 0.660 | 0.767 | 0.773 |
| Mexp + Audio + Audio | 0.720 | 0.746 | 0.692 | 0.668 | 0.813 | 0.846 |

**Table 2: Results of the ML approach over the different datasets**

| Modality | DL Approach | | Meta-Learning | | ML Approach | | ML Approach(Voting) | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | ROC AUC | Accuracy | ROC AUC | Accuracy | ROC AUC | Accuracy | ROC AUC |
| Gaze | - | - | 0.63 | 0.63 | 0.575 | 0.526 | - | - |
| Mexp | - | - | 0.60 | 0.60 | 0.523 | 0.555 | - | - |
| Audio | 0.620 | 0.530 | 0.652 | 0.634 | 0.597 | 0.524 | - | - |
| Gaze + Mexp | 0.636 | 0.637 | - | - | 0.659 | 0.677 | - | - |
| Gaze + Audio | 0.612 | 0.657 | - | - | 0.606 | 0.651 | - | - |
| Mexp + Audio | 0.551 | 0.561 | - | - | 0.666 | 0.712 | - | - |
| Gaze + Mexp+ Audio | 0.666 | 0.653 | - | - | 0.674 | 0.702 | 0.68 | 0.69 |

**Table 3: Comparison of all the approaches implemented with the combination of different modalities**

| | Accuracy | ROC AUC |
|---|---|---|
| Our ML Approach (RealLifeDeception) | 0.720 | 0.746 |
| Wu et. al.[9] | - | 0.8477 |

**Table 4: Comparisan of our ML approach with the approach of Wu et. al.[9]**

| | Accuracy | ROC AUC |
|---|---|---|
| Our ML Approach (BagOfLies) | 0.668 | 0.692 |
| Gupta et. al. | - | 0.646 |

**Table 5: Comparisan of our ML approach with the approach of Gupta et. al.[4]**

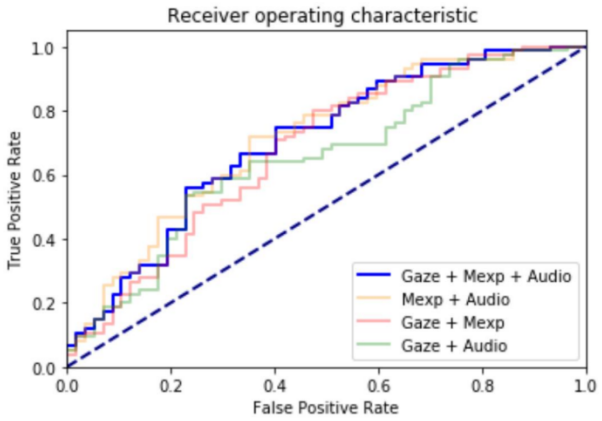


**Figure 6: ROC curves given for all possible combinations of our modalities using the ML approach**

given to the different modalities by using XGBClassifier and RandomForestClassifier. The corresponding weights for our XGBClassifier are 0.576, 0.217, and 0.205 while for RandomForestClassifier the weights 0.492, 0.352, and 0.154 for audio, micro-expression and gaze respectively. Clearly audio modality was given more importance by both classifiers

As seen in figure 3 we were able to reach an accuracy 0.674 for the combination of the three modalities with an AUC of 0.702. We went along and tested how each of our datasets will hold on their own. Figure 2 shows the results of each dataset with the corresponding Modality. The Youtube dataset gave us the highest accuracy of 0.813 with an AUC of 0.847.

Figure 4 shows the comparison of our model with Wu et. al.[9], knowing that the dataset tested on was the real-life trial dataset and which was edited for clarity by the authors.

Figure 5 compares the accuracy of our model against the one from Gupta[8] et. al. although the paper used the same dataset there were still differences in the pre-processing steps.

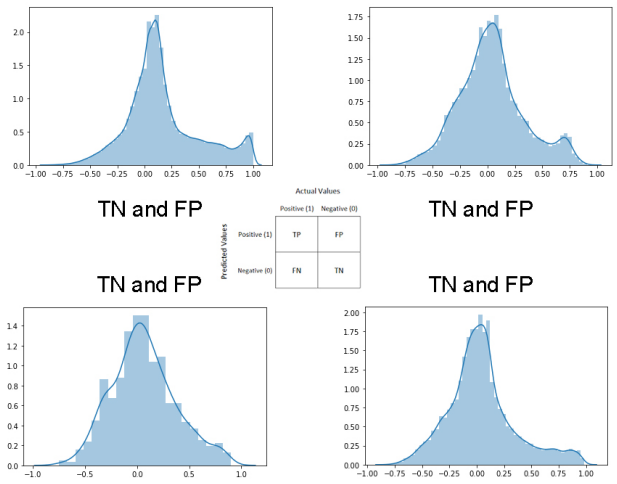


**Figure 7: Top left: cosine-similarity between TP and FP. Top right: cosine-similarity between TP and FN. Bottom left: cosine-similarity between FN and TN. Bottom right: cosine-similarity between FP and TN**

## 7 COMPARATIVE STUDY

Area Under Curve (AUC) is one of the most widely used metrics for evaluation for a binary classification problem. Now with the results obtained from both the DL and the ML approach at hand, we can compare both techniques. Figure 3 Shows a comparison between the simple DL approach, the Siamese Network, and the ML model, the table also shows how the various combination of all three modalities has performed. The Siamese Network performed best with singular Audio modality, since it is specifically tailored for sparse data, giving a 0.652 and a 0.634 for accuracy and AUC respectively. The multimodal DL approach for gaze, micro-expressions, and audio gave a 0.666 accuracy with 0.653 AUC. However, it was slightly out performed by the multimodal ML approach with a 0.674 accuracy and 0.702 AUC.

## 8 ANALYSIS

We intend to find out the similarity in the data that might be a probable cause why the model sometimes fail to differentiate between deceptive and truthful data. We want to see if there exist any similarity between true positive (TP) and False Positive (FP) and similarity between false negative(FN) and true negatives(TN). In order to find this, we implement an independent t-test where a pair-wise comparison is done for TP and FP, and for FN and TN. The resulting range of p-values for the pairs TP and FP is given from 0.13 to 0.98 and for FN and TN we get a range from 0.001 to 0.99. This implies that the similarity between data is spread out over a wide range which might be a possible cause for low accuracy.It would be interesting to see how many datapoints have very high similarity, this leads us to perform the next tests.

We then implement a one-way ANOVA test. As wanted to know if data classified as TP and FP drawn from the same sample. We also want to know if data classified as TN and FN are from same space. The resulting probability was 0.91 for data predicted to be deceptive came from the same population and 0.94 probability for data that was predicted as truthful being from the same population. This gave a an impression that the two different classes we are dealing with have very similar characteristic.

Figure 7 shows the cosine similarity for the corresponding confusion matrix. What strikes as odd is the similarity between all the combinations of TP and FP and TN. The two classes we are dealing with is quite inter-twined.

## 9 FUTURE WORK

As evident, our DL model falls short and suffers mainly through the scarcity of our data. We tried getting around this by implementing the Siamese network and compensating for the lack of data. Nonetheless what gave our model the boost was sampling the data using Fourier method in the pre-processing steps. We can look into imputation methods which can give us better results. Markov chain Monte Carlo (MCMC) methods comprise a class of algorithms for sampling from a probability distribution would be a good place to start with. In Gupta et. al. The EEG modality was also used for the predictions. Owing to the similarity between the data from the separate classes, it would be interesting to analyse the features that could efficiently differentiate them. This would mean exploring more modalities or trying to use co-related body movement data. Moreover, curating and adding more data for training might be fruitful.

## 10 CONCLUSION

In this paper, we were able to gather data from three different sources and curate one dataset from Youtube. While our simple deep learning model struggled because of the lack of data, we implemented a Siamese network that compensate for the loss. We decided to explore three features; gaze, micro-expressions, audio, and test the effect of the different combinations of these modalities on the prediction. It is also interesting to take into account the audio modality, which seems to be crucial in deception detection as voted by the classifiers of the Machine Learning method. The Machine Learning approach seems to outperform the deep learning model.

## REFERENCES

[1] 2016. Real-life Deception. http://web.eecs.umich.edu/~mihalcea/downloads/RealLifeDeceptionDetection.2016.zip [Online; accessed April 27, 2020].
[2] adas Baltrušaitis, Amir Zadeh, Yao Chong Lim, , and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *Proceedings of the IIEEE International Conference on Automatic Face and Gesture Recognition*. IEEE.
[3] Mingyu Ding, An Zhao, Zhiwu Lu, Tao Xiang, and Ji-Rong Wen. 2018. Face-Focused Cross-Stream Network for Deception Detection in Videos. *CoRR* abs/1812.04429 (2018). arXiv:1812.04429 http://arxiv.org/abs/1812.04429
[4] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractort. In *ACM Multimedia (MM)*.
[5] Theresa A Gannon, Anthony Beech, and Tony Ward. 2013. *Risk Assessment and the Polygraphk*. 129–154.
[6] Gangeshwar Krishnamurthy, Navonil Majumder, Soujanya Poria, and Erik Cambria. 2018. A Deep Learning Approach for Multimodal Deception Detection. *CoRR* abs/1803.00344 (2018). arXiv:1803.00344 http://arxiv.org/abs/1803.00344
[7] Columbia University, SRI International, and University of Colorado Boulder. 2013. CSC Deceptive Speech LDC2013S09. http://web.eecs.umich.edu/~mihalcea/downloads/RealLifeDeceptionDetection.2016.zip [Online; accessed April 27, 2020].
[8] Gupta Viresh, Agarwal Mohit, Arora Manik, Chakraborty Tanmoy, Singh Richa, and Vatsa Mayank. 2019. Bag-Of-Lies: A Multimodal Dataset for Deception

Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[9] Zhe Wu, Bharat Singh, Larry S. Davis, and V. S. Subrahmanian. 2017. Deception Detection in Videos. *CoRR* abs/1712.04415 (2017). arXiv:1712.04415 http://arxiv.org/abs/1712.04415