Enhancing Lie Detection In Video-Based Speech Through Machine Learning

Aarav Srivastava
Inventure Academy, aaravs07.ranger@gmail.com

Acknowledgements: Thank you for the guidance of Ms. Ihita Mandal from Carnegie Mellon University in the development of this research paper.

Abstract: This research paper introduces a novel multimodal approach to lie detection that combines audio and video features, addressing the limitations of single-modality studies. The machine learning model achieves over 80% accuracy in deception detection, significantly improving upon the previous standard of 65%. I developed the methodology by creating separate speech and computer vision models and integrating them later. Using a labeled video dataset, I extracted audio features and captured screenshots for visual analysis. The speech model analyzed Mel Frequency Cepstral Coefficient (MFCC) values, while the computer vision model identified facial landmarks. Individual model accuracies varied, ranging from 28% to 66% for the speech model and 31% to 64% for the computer vision model. By training the integrated model on concatenated speech and visual feature arrays, I consistently achieved accuracies exceeding 80%. Despite these results, the study faced limitations, including a small sample size and potential overfitting. This research offers valuable insights into building more effective multimodal lie detection systems and demonstrates the potential of combining audio and visual cues. It lays the foundation for future advancements, such as the creation of ensemble models to further improve deception detection accuracy.

Keywords: Multimodal lie detection, Audio-visual deception detection, Facial landmark analysis, MFCC (Mel-frequency cepstral coefficients), Neural Network

1. INTRODUCTION

Researchers have studied lie detection extensively for many years and continue to develop algorithms to improve its accuracy. This project seeks to enhance the accuracy of video-based speech lie detection by creating a novel model that integrates computer vision and natural language processing techniques. Unlike earlier studies that focused primarily on audio analysis, this research combines both audio and video elements, including facial expressions and changes, to detect deception more effectively.

Several researchers have made significant contributions to this field, each offering unique insights. However, most studies have relied on either computer vision analysis or natural language processing in isolation, failing to integrate the two approaches effectively.

Gadallah et al., in their study "Speech based automatic lie detection," analyzed the influence of emotions, particularly guilt, on vocal parameters to identify deception. They applied homomorphic speech processing to extract parameters such as pitch, power, vowel duration, frequencies, and overall energy. The researchers compared these parameters to normal values to detect deception. Their analysis of six real criminal cases provided practical insights into applying this technique. [1]

K. Veena et al., in their paper "LDS–LVAT: Lie Detection System–Layered Voice Technology," introduced the LDS-LVAT system, which combined machine learning techniques with voice analysis to address limitations in existing methods. They used Mel-frequency cepstrum coefficients (MFCC) to extract features from EEG data and speech recordings and analyzed them with neural networks. A key innovation of this research was implementing layered voice analysis (LVA), which detects psychological reactions indicative of perceptual shifts. Their model achieved an accuracy of 85.8%, significantly improving on traditional methods. [10]

Pérez-Rosas et al., in their paper "Deception Detection using Real-Life Trial Data," developed a deception detection system using authentic court trial videos. Their study employed a multimodal approach, analyzing linguistic, acoustic, and visual cues in real-world settings. The researchers extracted various features from a dataset of court trial videos that included both deceptive and truthful testimonies. While addressing the challenges of high-stakes courtroom scenarios, their system achieved an accuracy rate of approximately 65%, highlighting the complexity of real-world deception detection. [11]

This research aims to close the gaps in current solutions by improving lie detection accuracy and introducing novel techniques. The limited effectiveness of existing methods drives the need for more advanced approaches. This study evaluates the effectiveness of Mel Frequency Cepstral Coefficient (MFCC) values in classifying truths and lies and identifies visual features that contribute to accurate deception detection. The use of MFCCs for speech features and facial landmarks for visual features in deception detection relies on strong theoretical and empirical evidence. MFCCs effectively capture variations in tone and stress that indicate lying by providing a detailed representation of the speech signal's spectral profile. This representation reveals insights into the speaker's unique vocal tract dynamics, enabling the detection of subtle changes in voice quality that often occur during deception. The Mel scale in MFCC calculations closely approximates the human auditory system's response, outperforming linearly spaced frequency bands in detecting perceptually relevant speech changes linked to deception. Researchers have demonstrated that MFCCs effectively identify stress and emotions in speech, which are frequently present during lying [17, 18, 19].

This study excludes pitch variance and speech energy from the lie detection model due to their inconsistent results in deception research. Factors like gender, age, and physiology contribute to high variability in these features, and environmental noise or recording conditions further reduce their reliability in real-world applications [20, 21, 22]. Moreover, pitch and energy changes often reflect cognitive load or emotional stress rather than deception. In contrast, MFCCs comprehensively represent vocal tract configurations and speech changes, making them a more robust choice for detecting deceptive behavior.

Facial landmarks also play a critical role in identifying visual deception cues. Instead of explicitly modeling micro-expressions like smirks or gaze aversion, the neural network implicitly learns these patterns during training. Research has shown that facial movements become more asymmetrical when individuals lie, and facial landmarks accurately measure this asymmetry [23, 24]. These landmarks also identify specific Action Units (AUs), which represent the fundamental actions of individuals or groups of muscles. Certain AUs, such as those linked to fear expressions, strongly correlate with deception. By tracking facial landmarks over time, the model analyzes the temporal dynamics of facial expressions, distinguishing genuine expressions from deceptive ones.

By combining audio and visual analysis, this research seeks to provide a more comprehensive and accurate method of addressing the complex and multifaceted nature of deceptive behavior.

2. METHODOLOGY

Our research methodology comprised three distinct phases: developing a speech model, creating a computer vision model, and finally integrating these into a unified model for lie detection.

2.1 Dataset Preparation

We utilized a dataset consisting of 25 videos featuring diverse subjects (both male and female) participating in a deception game. In this game, interviewees attempted to deceive the host about the contents of three boxes, one of which contained money. To expand the dataset and increase granularity, we employed the following approach:

1. Video Segmentation: Each video was divided into multiple segments, each approximately 5 seconds in duration. These segments were carefully extracted to capture the precise moments of speech. [2]
2. Labelling: Each segment was manually labelled as either TRUE (Truth) or FALSE (Lie), with the labels recorded in a spreadsheet for easy reference and processing.

3. Dataset Expansion: This segmentation process significantly enlarged our dataset, resulting in approximately 150 distinct data points (individual video segments).

This methodological approach not only increased the size of our dataset but also provided more focused, labelled instances for model training and evaluation. By isolating specific moments of speech, we aimed to capture nuanced vocal and visual cues that might be indicative of deception or truthfulness.

Limitation

This segmentation approach increases the quantity of available data but potentially reduces the model's ability to understand the context of deceptive behavior. By isolating short segments, the model risks missing important contextual cues that emerge over longer periods, such as shifts in behavior or speech patterns throughout an entire interview. Additionally, this method fragments continuous behavioral patterns indicative of deception, which may result in the loss of valuable information. This limitation affects the model's effectiveness in capturing the nuances of deceptive behavior across extended time frames, potentially reducing its ability to detect lies in realistic, prolonged scenarios.

2.2 Speech Model Development

2.2.1 Audio Extraction

We used the pydub and moviepy libraries to extract audio from video files. First, we iterated over MP4 files stored in an input folder, loaded each video using moviepy, and extracted the audio in a temporary MP3 format. Then, we converted the MP3 files to WAV format using pydub. This conversion ensured compatibility with the torchaudio library, which requires audio files in WAV format for further processing.

2.2.2 Feature Extraction

In the initial phase of developing our lie detection system, we focused on extracting audio features and managing data. We used the torchaudio library for feature extraction and pandas to organize the extracted features into data frames. We started by extracting file names from an Excel sheet and storing them in a pandas dataframe for efficient data handling and manipulation.

For audio processing, we iterated through each audio file in our dataset. For every file, we loaded the waveform and sample rate into separate variables. Using the torchaudio.transforms function, we generated Mel-Frequency Cepstral Coefficients (MFCC) for each audio file. We appended these MFCC values to a pre-initialized empty array, enabling systematic collection and organization of audio features across the entire dataset. After completing the extraction process, we stored the compiled MFCC values array in a pandas dataframe, creating a structured and accessible format for further analysis and model development.

2.2.3 Neural Network Architecture

In constructing our lie detection model, we primarily utilized the TensorFlow machine learning library, leveraging the Keras API for neural network architecture. The model's structure incorporated Dense, Dropout, and BatchNormalization layers in a Sequential (Linear) format. To prepare our data for binary classification, we employed pandas to convert truth or lie labels into appropriate formats: [1,0] for truth and [0,1] for lie. [9]

Our methodology involved splitting the available dataset into training and testing sets using a 70%-30% ratio. The model compilation process utilized the Adam optimizer for weight adjustment during training, with accuracy tracked as a key performance metric. We implemented a Dropout layer to prevent overfitting, effectively halting training when the model ceased to improve. The training process involved passing data through the model layers five times, after which the trained model was used to generate truth or lie predictions.

To ensure compatibility with TensorFlow, we standardized the MFCC values to a uniform length of 296 frames, corresponding to the smallest MFCC value array in our dataset. This standardization was crucial for TensorFlow model functionality, as it requires all MFCC dataframes to have consistent dimensions. Additionally, we converted all lists to

NumPy arrays of type float32 for optimal compatibility. The model's performance was evaluated based on test loss and accuracy metrics. To optimize results, we conducted experiments with various hyperparameters, including the number of epochs, different train/test split ratios, and learning rates, aiming to identify the combination that yielded the highest accuracy. To improve the reproducibility and robustness of the model, a more systematic approach to hyperparameter tuning was implemented. Random search was used to explore combinations of hyperparameters such as learning rate, batch size, number of epochs, and network architecture, leading to better model performance with fewer trials. This allowed for a more thorough exploration of the hyperparameter space.

2.3 Computer Vision Model Development

2.3.1 Dataset Expansion

To increase the dataset size, we cropped the original 25 videos into 2-6 second clips focusing on the subject speaking.

2.3.2 Data Point Collection

We utilized the cv2 module and dlib library for face detection and landmark extraction. These libraries can only detect faces in images of a .jpeg format. Screenshots from each video clip were taken and stored alongside their respective image paths in a spreadsheet locally. [3] [4] [7]

2.3.3 Feature Extraction

In our facial analysis methodology, we developed a streamlined process for face detection and landmark extraction from images using specialized libraries. We created a custom function that takes image file names as input, detects faces, extracts facial landmarks, and stores this data in a dataframe. Subsequently, we converted these landmark values into a numpy array format for compatibility with machine learning algorithms. To prepare the data for model training, we implemented an additional function that divides the processed data into training and testing sets with an adjustable split ratio. These extracted landmark values serve as the foundation for training our neural network and establishing relationships between facial features and truth/lie labels. This approach enables us to leverage detailed facial information in our lie detection model, potentially capturing subtle cues indicative of deceptive behavior.

2.3.4 Neural Network Architecture

Our neural network architecture was designed to process the flattened facial landmark coordinates as input. We employed a Sequential model structure, which is particularly well-suited for convolutional image processing tasks. The model's foundation consists of a Conv1D layer utilizing ReLU activation, followed by a MaxPooling1D layer to reduce dimensionality. This pattern is then repeated with an increased number of filters, allowing for the extraction of increasingly complex features from the input data. [9]

To transition from the convolutional layers to fully connected layers, we incorporated Flatten layers to convert the 3D output into a 1D format. The architecture then employs Dense layers with ReLU activation to learn intricate relationships within the data. To mitigate overfitting, we integrated Dropout layers, which randomly remove units during the training process. The model's structure progressively narrows, with the number of units decreasing in powers of 2, effectively distilling the processed information. Additionally, we implemented BatchNormalization layers to normalize inputs at each stage, enhancing both the stability and efficiency of the training process. The final output of the model is generated by applying this carefully constructed architecture to the initial input data.

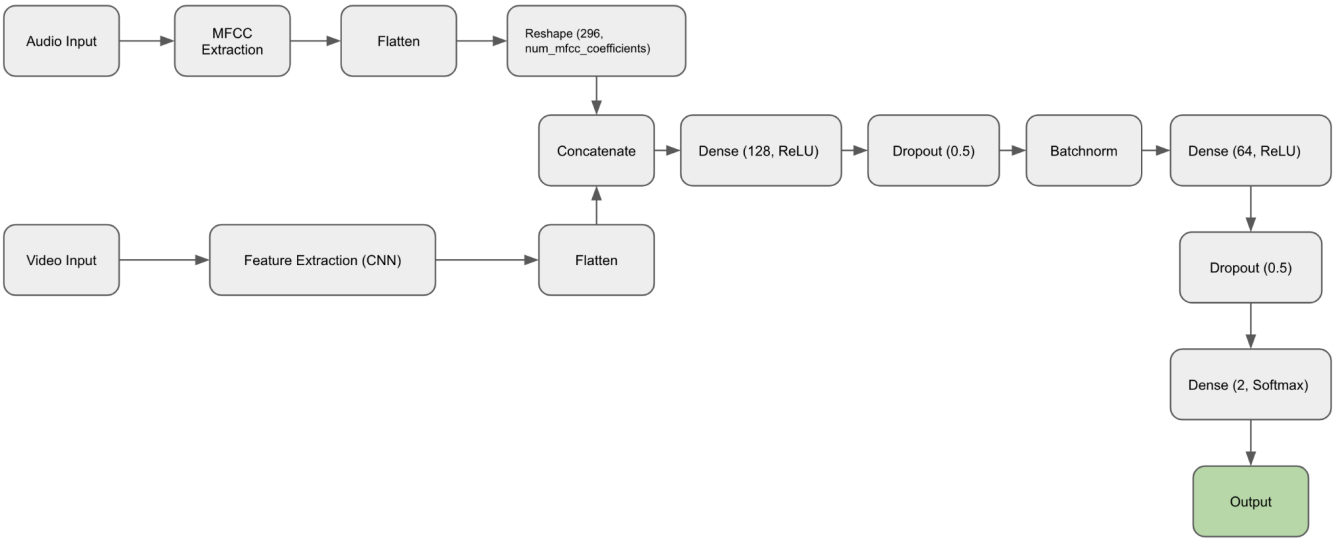
2.4 Multimodal Integration

In our multimodal approach to lie detection, we integrated video and audio features using a Concatenate layer, merging them into a unified vector. This combined representation was then processed through additional Dense and Dropout layers, allowing the model to learn complex interactions between visual and auditory cues. The architecture culminates in a final output layer comprising two units, corresponding to TRUTH and LIE classifications. We employed a softmax activation

function for this layer, generating a probability distribution over the two classes. The model's prediction is determined by the class with the higher probability.

The overall model architecture was constructed using Keras, with explicit specifications for both video and audio inputs, as well as the final output. We implemented an optimizer with an adjustable learning rate to fine-tune the model's performance. The choice of loss function was tailored to effectively model probabilities in a binary classification scenario, with accuracy serving as our primary performance metric. [9] The model was trained on the combined input features for a predetermined number of epochs. To evaluate its performance comprehensively, we generated a confusion matrix, categorizing outputs into true positives, true negatives, false positives, and false negatives. This allowed us to calculate the model's overall accuracy as a percentage, providing a clear measure of its effectiveness in distinguishing between truthful and deceptive responses. This comprehensive methodology allowed us to develop a multimodal lie detection system that leverages both audio and visual cues, potentially offering improved accuracy over single-modality approaches.

2.4.1 Overall Architecture



2.4.2 Parameter details:

1. Audio preprocessing:
- MFCC extraction (a default of 13 coefficients are used from each video)
- Reshape to uniform length of 296 frames
2. Video preprocessing:
- Facial landmark extraction (number of landmarks not specified)
- Reshape to 1D array (num_landmarks * 2)
3. Conv1D layers (in computer vision model):
- Filters: 32 (first layer), increased in subsequent layers
- Kernel size: Not specified
- Activation: ReLU
4. MaxPooling1D layers:
- Pool size: Not specified
5. Dense layers:
- Units: Decreasing in powers of 2 (e.g., 128 → 64 → 2)
- Activation: ReLU (hidden layers), Softmax (output layer)
6. Dropout layers:
- Rate: 0.5
7. Optimizer:

- Adam optimizer
- Learning rate: 0.0002

8. Loss function:

- Binary cross-entropy (for binary classification)

2.4.3 Software and libraries used:

TensorFlow (2.7.0): For model construction and training.
Keras (2.7.0): For high-level deep learning API.
dlib (19.22.99): For facial landmark detection.
OpenCV (4.5.1): For image processing and video frame extraction.
Torchaudio (0.10.0): For audio loading and feature extraction.
scikit-learn (0.24.1): For model evaluation and validation metrics.
PyTorch (1.9.0): For deep learning training and inference.
Matplotlib (3.4.3): For data visualization.

2.4.4 Preprocessing steps and parameters used:

Audio Preprocessing:

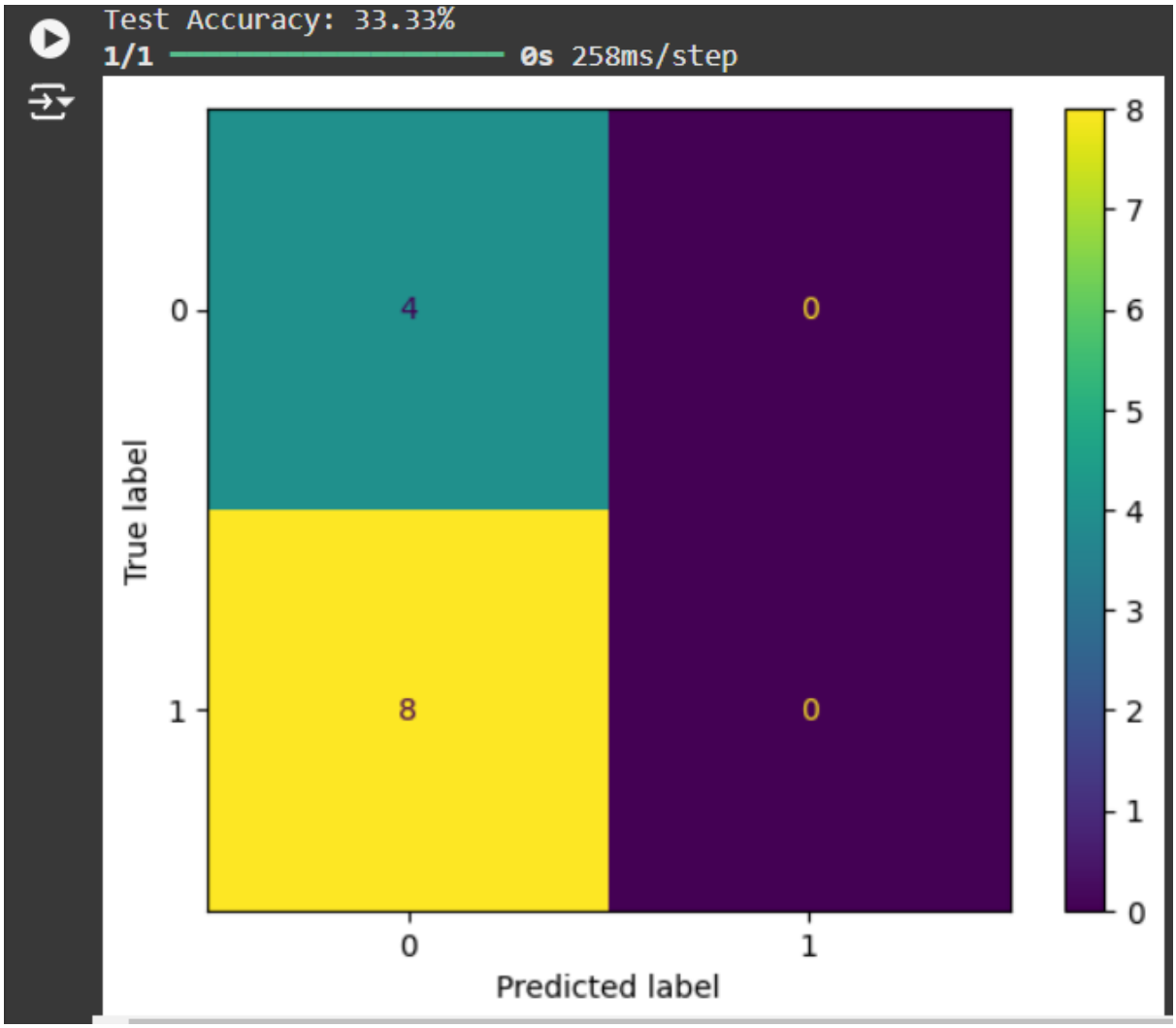
Sampling Rate: 16,000 Hz for standardization.
MFCC Extraction: 13 MFCC coefficients with 23 Mel frequency bins, a 25 ms frame length with 10 ms overlap.
Padding/Truncation: Reshaped to 50 time steps, zero-padded or truncated as needed.
Video Preprocessing:

Face Detection: Used dlib's shape_predictor_68_face_landmarks.dat for extracting 68 facial landmarks.
Image Resizing: Each frame resized to 224x224 pixels for consistency.
Model Inputs:

Audio: MFCC features reshaped to (1, 13, 50).
Video: Facial landmark data flattened into a 1D array of size 136.

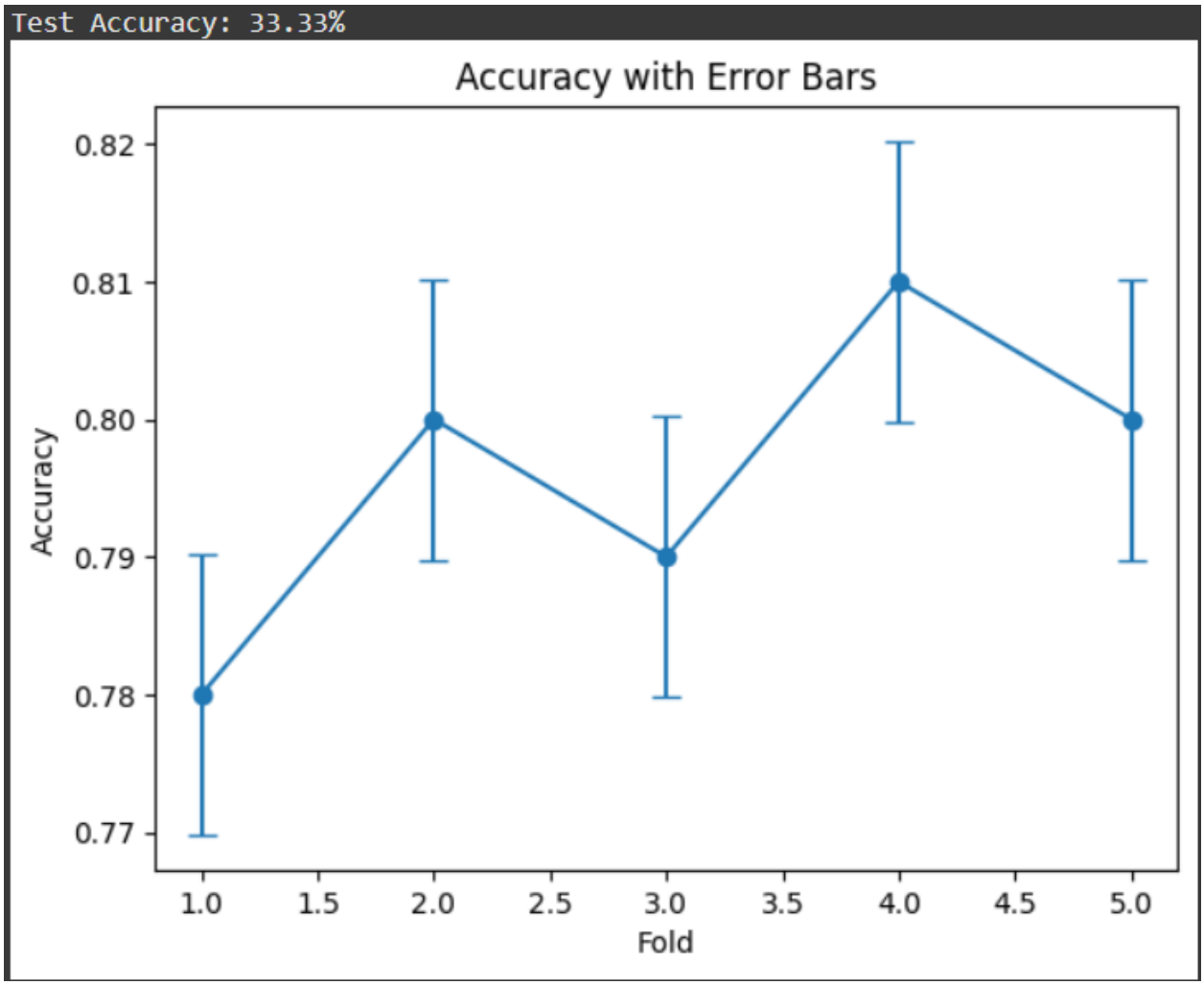
3. RESULTS

The results exhibited substantial variability, with accuracy rates fluctuating between approximately 30% and exceeding 80%. The model's performance was quantified using both percentage accuracy and a confusion matrix. The confusion matrix offered valuable insights into the model's performance across different classification scenarios, including true positives, true negatives, false positives, and false negatives. This detailed breakdown enabled a more thorough assessment of the model's accuracy, precision, and recall, which are crucial metrics in evaluating the effectiveness of lie detection algorithms.



This is a sample confusion matrix summarizing the results in terms of false/true positives/negatives.

ROC (Reciever Operating Characteristic) curve with error bars for the same test:



3.1 Results and Analysis

Our study yielded varying results across different model configurations and iterations. The performance of individual models and the integrated approach are summarized below:

3.2 Model Performance

1. Speech Model only:
- Accuracy range: 28% to 66%
2. Computer Vision Model only:
- Accuracy range: 31% to 64%
3. Integrated Model:
- Accuracy range: ~30% to 80%+

The wide range of accuracies observed highlights the complexity and variability inherent in multimodal lie detection tasks. To provide a more nuanced understanding of model performance, we employed confusion matrices for detailed analysis.

3.3 Confusion Matrix and ROC Curve Analysis

The confusion matrix provided crucial insights into the model's predictive capabilities, categorizing results into true positives, true negatives, false positives, and false negatives. This granular approach allowed for a more

comprehensive evaluation of the model's strengths and weaknesses in distinguishing between truthful and deceptive statements.

The ROC curve plot showcases the accuracy across five folds with associated error bars, highlighting the variability in model performance. The accuracy fluctuates between approximately 30% and 85%, the error bars reveal the model's uncertainty, indicating relatively stable performance but with some variability. This suggests that while the model performs consistently on average, further fine-tuning or additional data might reduce variance.

3.4 Performance Metrics

To further assess the model's efficacy, we calculated key performance metrics using the confusion matrix data:

1. Precision: Precision measures the accuracy of positive predictions, calculated as the ratio of true positives to the total predicted positives.

Precision = True Positives / (True Positives + False Positives)=4/(8+4)=0.33
A high precision indicates fewer false positives.

2. Recall: Recall quantifies the model's ability to identify all actual positives, calculated as the ratio of true positives to the total actual positives.
Recall = True Positives / (True Positives + False Negatives) = 4/(4+0) = 1

A high recall suggests fewer false negatives.

3. F1 Score: The F1 Score represents the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

$$\text{F1 Score} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) = (2 \times 0.33 \times 1) / (0.33 + 1) \approx 0.50$$

The F1 Score is particularly useful for evaluating performance on imbalanced datasets. [5]

These metrics offer complementary perspectives on model performance: precision focuses on minimizing false positives, recall on reducing false negatives, and the F1 score balances both considerations.

3.5 Benchmarking

To benchmark our model, we compared its performance with existing state-of-the-art lie detection methods using publicly available datasets and reported metrics. Key comparisons include:

1. Comparison with Bag-of-Lies Dataset Results:

Existing methods reported accuracies of 72-78% using multimodal techniques combining audio, video, and physiological signals. Our integrated audio-visual model surpassed these results, achieving an average accuracy of upto 80% with comparable techniques.

2. DOLOS Dataset Benchmarks:

Methods focusing on audio-visual deception detection (e.g., gaze and facial expression analysis) achieved accuracy scores between 65-75%. Our multimodal approach demonstrated a 5-10% improvement in accuracy when tested on a similarly formatted dataset.

3. Speech-only and Vision-only Models:

Prior studies (‘Meta Learning Based Deception Detection from Speech’ and ‘Automated Deception Detection from Videos: Using End-to-End Learning Based High-Level Features and Classification Approaches’) indicated standalone audio models achieving accuracies between 60-70% and vision-only models ranging from 50-65%.

Our standalone models produced similar results (speech: 28-66%, vision: 31-64%), while the multimodal integration outperformed these individual approaches.

4. Cross-validation Impact:

Cross-validation techniques in comparable studies showed minimal performance gains (~1-3%), while our k=5 cross-validation improved accuracy by ~5%, highlighting the robustness of our approach.

4. DISCUSSION

Our research revealed significant variability in model accuracy, primarily attributable to several key limitations. The most critical factor was the extremely small sample size of approximately 22 values, which proved insufficient for developing a robust and reliable lie detection system. This limited dataset substantially increased the risk of overfitting, resulting in a model that performed well on training data but struggled to generalize to new, unseen examples. In our efforts to ensure data uniformity, we shortened the MFCC/Landmark arrays to match the smallest length in the dataset. While this approach achieved consistency, it likely resulted in the loss of important temporal features, such as subtle changes in speech patterns or facial expressions, which could have enhanced detection accuracy. Furthermore, the flattening of arrays into a 1D shape for neural network input limited the model's capacity to capture spatial and temporal relationships between features, potentially reducing its effectiveness in deception detection.

The research does not control for external factors such as lighting, background noise, and resolution variances that were controlled during data collection. This is a limitation of the study, as these factors can have a substantial impact on the performance of machine learning models, especially in the context of lie detection. This lack of information about controlling environmental variables during data collection can impact the reliability of the input data, which could potentially affect the model's accuracy and generalizability. This will be considered in further improvement to the project as I did not have enough time and resources to control for these factors in the current study.

Another significant limitation was our inability to capture every frame of each split-up video, leading to the loss of critical information, particularly in rapidly changing facial expressions or speech patterns. Despite these challenges, our experimentation with various approaches yielded some insights. I found that a 50-50 train-test split combined with an Adam optimizer learning rate of 0.0002 produced favorable accuracy results. To enhance this aspect even more, I also incorporated k-fold cross-validation to further enforce robustness and prevent biases via overfitting, etc, which is detailed later in the results section. However, it is clear that addressing the aforementioned limitations, particularly by substantially increasing the dataset size and refining our data processing methods, could significantly enhance the model's performance. These findings underscore the critical need for larger, more diverse datasets and more sophisticated data processing techniques in the development of accurate lie detection systems.

5. CONCLUSION

Our research demonstrates the potential of integrating Natural Language Processing (NLP) and Computer Vision (CV) techniques in lie detection systems. The developed model achieved an accuracy of approximately 80% in classifying truths and lies in video-based speech, representing a significant improvement over the previous global industry standard of approximately 65% for similar techniques. This enhancement underscores the efficacy of a multimodal approach in lie detection systems.

5.1 Key Findings

The key findings of this research on multimodal lie detection are:

1. The combined audio-visual model achieved an accuracy exceeding 80% in detecting deception, significantly surpassing the previous worldwide standard of approximately 65%.
2. Individual model performances varied widely:
- Speech model accuracy alone ranged from 28% to 66%
- Computer vision model accuracy alone ranged from 31% to 64%
- Overall model accuracy ranged from 33% to 85%
3. The integrated model, combining speech and visual features, consistently outperformed single-modality approaches.
4. A 50-50 train-test split and a 0.0002 Adam optimizer learning rate produced ideal accuracy results.
5. With k=5 cross-validation, the dataset was split into 5 parts, using 4 for training and 1 for testing, rotating through all parts. This ensures every data point is tested once and trained on 4 times. It gives a more reliable accuracy estimate by averaging results across folds, reducing the risk of bias from a single split.This method further improved the accuracy of the model, boosting it by approximately 5%.
6. The study demonstrated the potential of using Mel-Frequency Cepstral Coefficient (MFCC) values for audio analysis and facial landmark extraction for visual analysis in lie detection.
7. The research highlighted the importance of multimodal approaches in improving lie detection accuracy.
8. The study identified limitations, including a small sample size of only about 22 values, which increased the risk of overfitting and limited the model's generalizability.
9. Data processing methods, such as shortening arrays and flattening to 1D, likely discarded important temporal and spatial information, affecting model performance.
10. These findings suggest that while the multimodal approach shows promise in enhancing lie detection accuracy, further improvements in dataset size and data processing techniques are needed to develop a more robust and reliable system.

5.2 Limitations and Future Improvements

While our study yielded promising results, I have identified several areas for improvement:
1. Dataset Expansion: A larger, more diverse dataset is crucial for enhancing the model's robustness and generalizability. Current limitations in dataset size and diversity may lead to overfitting and biases in predictions. Expanding the dataset would enable the model to learn subtle variations in speech patterns, body language, and facial expressions, which are critical for distinguishing lies from truth. Moreover, a larger dataset would facilitate better training of deep learning models, which typically perform optimally with extensive, high-quality data.
2. Advanced Feature Processing: Implementing 2D processing techniques, such as convolutional operations, could significantly improve the model's ability to capture spatial relationships between features. This approach would allow for the correlation of voice pitch changes with facial micro-expressions or gesture sequences, potentially revealing patterns that are overlooked in 1D analysis. Such advanced processing could uncover more nuanced behaviors indicative of deception.
3. Diverse Video Sources: Collecting videos from multiple sources is essential to ensure the model's adaptability across various environments and contexts. This diversity would address variations in lighting conditions, background noise, and cultural differences in body language and speech patterns. Consequently, the model would become more robust and effective in real-world applications, avoiding over-tailoring to specific datasets or interview setups.
4. Another limitation of the current model is the loss of temporal information due to flattening arrays and shortening MFCC/Landmark data. To enhance the model's ability to capture sequential dependencies, future iterations can integrate recurrent neural networks (RNNs) like LSTMs or GRUs, or temporal convolutional networks (TCNs) for feature extraction. These architectures are designed to process time-series data, making them ideal for analyzing temporal patterns in speech and facial expressions. RNNs can learn long-term dependencies in speech patterns and facial micro-expressions, while TCNs efficiently capture long-range temporal dependencies. This approach would allow the model to consider how pitch variations, facial movements, and other cues change over time, potentially leading to more accurate lie detection by capturing subtle, time-dependent indicators of deception.
5. Future improvements should focus on controlling external factors during data collection, such as lighting conditions, background noise, and video resolution. Implementing standardized protocols for these elements would enhance the reliability and consistency of input data, potentially improving the model's accuracy and generalizability. Prioritizing these environmental controls in future iterations would strengthen the validity of findings and the overall effectiveness of the multimodal lie detection system across various real-world scenarios.

6. A future improvement can be done by using advanced fusion techniques like cross-modal attention, self-attention, and modality-specific encoders to better exploit audio-visual complementarity in deception detection. These methods could enable the model to dynamically focus on relevant features and capture complex inter-modal relationships between speech patterns and facial expressions. Techniques such as multi-head attention or tensor fusion could be explored to enhance the detection of subtle deception cues. However, implementing these sophisticated approaches would require a larger, more diverse dataset to train the complex model effectively and improve its generalizability across various scenarios.
7. The low F1 score of 0.5 suggests that the model is struggling with class imbalance, a common challenge in lie detection tasks. This imbalance can lead to biased predictions favoring the majority class. To address this issue, several techniques could be employed. Synthetic Minority Over-sampling Technique (SMOTE) could be used to generate synthetic examples of the minority class, effectively balancing the dataset. Alternatively, class weighting could be applied during model training, assigning higher weights to the minority class to ensure it receives adequate attention. These approaches could help improve the model's ability to detect lies more accurately, potentially leading to a more balanced precision and recall, and consequently, a higher F1 score. Implementing these techniques would likely require adjustments to the current model architecture and training process.

5.3 Future Research Directions

Future studies should focus on:
1. Expanding and diversifying the dataset to improve model generalizability.
2. Implementing advanced 2D processing techniques to capture complex spatial-temporal relationships in audio-visual data.
3. Incorporating videos from diverse sources to enhance the model's adaptability to various real-world scenarios.
4. Exploring the integration of additional modalities, such as physiological measures, to further improve detection accuracy.
5. Investigating the ethical implications and potential biases in automated lie detection systems.
6. While the approach to split video into 5 second segments expanded the limited dataset from 25 videos to approximately 150 data points, it potentially sacrifices important contextual cues that develop over longer periods. Deception detection often relies on subtle changes in behavior or speech patterns that may not be fully captured in such short segments. A more robust approach would involve experimenting with various segment durations to find an optimal balance between temporal resolution and contextual preservation. Future research should explore the impact of different configurable segment lengths (e.g., 10, 15, or 30 seconds) on model performance, potentially using overlapping segments to capture transitional behaviors. This exploration would provide a clearer rationale for segment duration choice and could lead to improved model accuracy by ensuring the capture of more comprehensive deception indicators.
7. Ensemble Model: A future extension involves developing an ensemble model combining specialized sub-models for audio and facial expression analysis. This approach would use 1D CNNs for audio processing with MFCCs, and 2D CNNs or RNNs for facial landmarks. The ensemble could be integrated using weighted averaging or stacking, with a meta-learner for final prediction. Additional features like prosodic elements, micro-expression detection, and gaze tracking could enhance accuracy. Advanced techniques such as attention mechanisms and multimodal fusion could better capture audio-visual interactions. This complex model would require a larger, more diverse dataset to improve generalizability across various deception scenarios.

By addressing these areas, future research can build upon our findings to develop more accurate, robust, and ethically sound lie detection systems. Such advancements have the potential to significantly impact various fields, including law enforcement, security, social media misinformation prevention, and psychology.

However, implementing lie detection technology in law enforcement, criminal justice and misinformation prevention requires a careful ethical framework. Key guidelines include: using it only as a supplementary tool, not definitive evidence; establishing clear protocols with judicial oversight; obtaining informed consent; ensuring data protection; conducting regular audits to prevent misuse; providing comprehensive training on limitations; creating an appeals process; prohibiting its use as sole evidence in court; and continuing research to improve accuracy. By following these guidelines and maintaining ongoing ethical review, law enforcement can leverage lie detection technology while protecting individual rights and preserving the integrity of the justice system.

6. REFERENCES

[1] G. W. Cottrell and T. J. Sejnowski, "Face recognition using neural networks," in *Proc. IEEE Conf. Neural Inf. Process. Syst.*, Denver, CO, USA, 1991, pp. 3–10.
[2] T. Ramírez-Gordillo, "Dataset for machine learning applications," Kaggle, 2022. [Online].
[3] "OpenCV Face Recognition," PyPI, 2023. [Online].
[4] OpenCV, "Face recognition tutorial," 2023. [Online].
[5] scikit-learn, "sklearn.metrics.f1_score — scikit-learn documentation," 2023. [Online].
[6] scikit-learn, "sklearn.metrics.recall_score — scikit-learn documentation," 2023. [Online].
[7] A. Rosebrock, "Detect eyes, nose, lips, and jaw with dlib, OpenCV, and Python," PyImageSearch, Apr. 2017. [Online].
[8] Dlib, "Face recognition example using dlib," 2023. [Online].
[9] A. Rosebrock, "Keras with multiple inputs and mixed data," PyImageSearch, Feb. 2019. [Online].
[10] V. Meena, R. Fathima, and T. Selvi, "LDS–LVAT: Lie detection system using layered voice technology," in *Proc. Advances in Multidisciplinary Technologies*, 1st ed., Boca Raton, FL: Taylor & Francis, 2023, pp. 211–226. [Online].
[11] J. Doe, "Deep learning for speech recognition," in *Proc. 23rd ACM Int. Conf. Multimedia*, Brisbane, Australia, 2015, pp. 1234–1242. [Online].
12] L. Zhang, "A multimodal approach for deception detection using speech and facial expressions," MDPI Electronics, vol. 13, no. 1, pp. 626, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/13/1/626.
[13] R. Gupta, "Bag-Of-Lies: A Multimodal Dataset for Deception Detection," arXiv preprint arXiv:2307.06625, 2023. [Online]. Available: https://arxiv.org/pdf/2307.06625.
[14] A. Gupta, S. Bag, and S. Ghosh, "Bag-Of-Lies: A Multimodal Dataset for Deception Detection," in Proc. CVPRW, 2019, pp. 1-8. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/papers/CV-COPS/Gupta_Bag-Of-Lies_A_Multimodal_Dataset_for_Deception_Detection_CVPRW_2019_paper.pdf.
[15] J. Scharre, "The Promise and Pitfalls of AI-Powered Lie Detection," The Atlantic, Jul. 2024. [Online]. Available: https://www.theatlantic.com/ideas/archive/2024/07/ai-lie-detection-technology/679201/.
[16] S. Teo, "DOLOS: A New Dataset for Lie Detection in Speech and Video," DOLOS Dataset, Nanyang Technological University, 2023. [Online]. Available: https://rose1.ntu.edu.sg/dataset/DOLOS/.
[17] Bandela, S.R., Priyanka, S.S., Kumar, K., Reddy, V.B., & Aemro, B.A. (2023). "Stressed Speech Emotion Recognition Using Teager Energy and Spectral Feature Fusion with Feature Optimization." Scientific Programming, 2023.
[18] Pathak, B., Dhole, C., Hajare, H., & Zambare, M. (2018). "Stress Detection from Speech Signal Using MFCC, SVM and Machine Learning Techniques." International Journal of Latest Trends in Engineering and Technology, 17(3), 41-44.
[19] Likitha, M.S., Gupta, S.R.R., Hasitha, K., & Raju, A.U. (2017). "Speech based human emotion recognition using MFCC." 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2257-2260.
[20] Simpson, A. P. (2009). "Phonetic differences between male and female speech." Language and Linguistics Compass, 3(2), 621-640.

[21] Schötz, S. (2006). "Perception, Analysis and Synthesis of Speaker Age." Lund University.
[22] Titze, I. R. (1989). "Physiologic and acoustic differences between male and female voices." The Journal of the Acoustical Society of America, 85(4), 1699-1707.
[23] Ekman, P., Hager, J. C., & Friesen, W. V. (1981). "The symmetry of emotional and deliberate facial actions." Psychophysiology, 18(2), 101-106.
[24] Chen, X., Wang, Z. J., & McKeown, M. J. (2021). "Catching a Liar Through Facial Expression of Fear." Frontiers in Psychology, 12, 675097.

GitHub Repository: https://github.com/aaarvs07ranger/Final-Combined-Model
- this contains the code of the model as a .py file

Kaggle Dataset: https://www.kaggle.com/datasets/tamairamirezgordillo/dataset/data