Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 19, no. 8, August 2023, pp. 547-558
DOI: 10.23940/ijpe.23.08.p7.547558

Enhancing Deception Detection with Exclusive Visual Features using Deep Learning

Victor Diaz[a], W. Eric Wong[b,*], and Zizhao Chen[b]

[a]*California Polytechnic State University, Humboldt, USA*
[b]*University of Texas at Dallas, Richardson, Texas, USA*

Abstract

A combination of nonverbal cues, verbal cues, and measurements of body abnormality make guidelines to determine deceitfulness. The combination of these guidelines will vary from person to person, making deception detection a complex challenge. Research has demonstrated that the accuracy of the latest computerized polygraph testing techniques is 98% accurate. Several human-controlled variables help to achieve this level of accuracy, such as being properly trained and must use an accepted procedure and scoring system from the British Polygraph Society. This causes a lack of availability for Deception detection as the implementing these techniques have training from the British Polygraph Society. Hence this research aims to reduce the requirements of lie detection by relying on Visual Features tracked with computer vision. The proposed multi-modal will track facial and body movements to classify whether a person is Deceiving or telling the Truth. The model proposed will use data consisting of videos collected from public court trials. The data will be cleaned with Facial Action Units (AU) with OpenFace, and then augmented with various rotations. The features extracted from the videos are the Movement with Holistic landmarks and Unique features from deep learning extraction. The Multi-model will consist of three pathways: a 3D-CNN pathway, a CovLSTM2D Pathway, and a dense pathway. The outputs of the three paths are concatenated and fed into a dense layer with SoftMax activation for classification. With a continuous emphasis on examining the proposed methodology for model creation, we discovered that higher accuracy can be achieved by leveraging deep learning algorithms for visual inputs as complex as the human body.

Keywords: deception detecting; exclusive visual; computer vision

1. Introduction

Computer vision constitutes a specialized field within the realm of artificial intelligence, concentrating on the extraction of pertinent information from diverse forms of visual input such as images and videos. In stark contrast to the innate ability of humans to perceive objects and ascribe significance to them, computers are inherently devoid of this capability. However, through the application of machine learning techniques, computers can be empowered to emulate the cognitive processes of humans [1]. Within the domain of computer vision, the integration of these machine learning algorithms serves as a conduit for the classification, extraction, and discernment of data, leading to the proficient identification and labeling of objects [2].

Machine learning has become an indispensable tool in the realm of computer vision, facilitating remarkable advancements in various applications. This synergy between machine learning and computer vision is exemplified by the utilization of convolutional neural networks (CNNs), a class of deep learning models, which have demonstrated exceptional performance in tasks such as image classification and object detection [3, 4]. The ability of CNNs to automatically extract hierarchical features from raw pixel data has revolutionized image analysis, allowing systems to recognize patterns and objects with remarkable accuracy. Furthermore, the application of recurrent neural networks (RNNs) in tasks like image captioning has enabled machines to describe visual content with human-like fluency [5]. These machine learning techniques are instrumental in enhancing the capabilities of computer vision systems, enabling them to decipher complex visual information and opening doors to a wide array of applications, from autonomous vehicles to medical image analysis.

* Corresponding author.
E-mail address: ewong@utdallas.edu

Behavior/Emotions are one of the interests of computer vision. There are many challenges when annotating emotions with machine learning [6]. According to the American Psychological Association: “Emotions are conscious mental reactions (such as anger or fear) subjectively experienced as strong feelings usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body. [7]” The state of mind of an individual affects the data used in machine learning training, which makes notating emotions with machine learning challenging [8].

In a recent study, computerized polygraph testing techniques have a 98% average accuracy in field examinations, 92% in field-independent analysis, 80% in laboratory simulations, and 81% in laboratory-independent analysis [2]. While the accuracy level makes deception detection a viable tool, the methodology limits the availability of such techniques. The implementation of the methodology is limited to examiners who undergo continuous professional development and have qualified certification. The methodology consists of using the latest techniques and methods in the field, being a current member of the British Polygraph Society or the American Polygraph Association, and submitting work for quality checks as part of a commitment to maintaining the highest testing standards possible. In addition, having the minimum recording equipment is needed for cardiovascular, respiratory, and electrodermal activity [8]. With computer vision, the goal is to reduce the input requirement to only video footage.

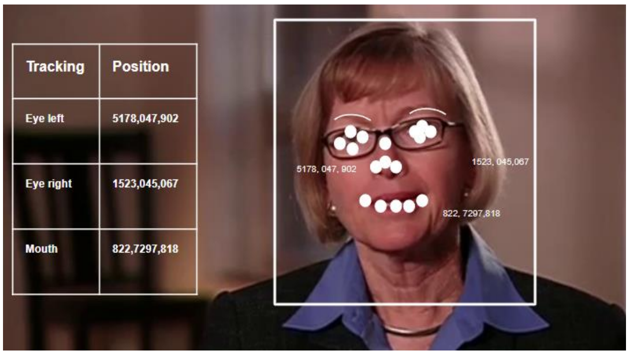
To narrow the scope of this research, we focus on safety-critical environments. Safety-critical environments are scenarios where deception or intentionally misleading someone can cause safety concerns. In court, misinformation can cause an increase in the credibility of certain testimony [9]. Misinformation can affect the fairness of trial cases [10]. This research aims to increase the availability of computer vision in critical environments such as courts and trials. In a critical environment where a person's future is at stake, having a tool that can help make more informed decisions and is accessible at any time is valuable. Hence this research aims to increase the availability of deception detection with computer vision by proposing a multimodal which only uses Visual feature extraction.

The paper is structured as follows: Section 2 provides a comprehensive overview of computer vision fundamentals, including an assessment of previously proposed models and related research. Section 3 outlines the research methodology and implementation specifics, while Section 4 details the experimental procedures, including the creation of both the primary and sub-models. Section 5 focuses on presenting and analyzing performance results during validation and testing. Finally, Section 6 offers conclusions drawn from the research and identifies future directions for further investigation, ensuring a logical and coherent flow of information throughout the paper.

2. Related Work

2.1 Visual Features Detection

Studies on visual cognition show that humans do not simultaneously concentrate on an entire scene. Instead, they sequentially direct their attention to various portions of the scene to extract pertinent information [11]. Action recognition concerns finding meaning from a set temporal visual field information. Similar to how the brain works. Visual features are the image frame-based information derived from videos. In Deception detection, we use different features extracted from visual data, such as Acoustical features (Voice or audio), textual features (text encoding or transcripts), or Visual features (image frame-based information). The filtering stage is a standard start with biological models [12]. And depending on what feature filter is applied, that data can improve the Machine learning dome [13]. From the Visual input of frame data in Figure 1 we can see an example of information extraction. Now know that the frame contains facial features.



Tracking	Position
Eye left	5178,047,902
Eye right	1523,045,067
Mouth	822,7297,818

Figure 1. Example of information being extracted from a frame

Numerous publications on deception detection have undertaken the extraction of visual features and subsequent classification. These studies have included experiments as Table 1:

The Face Landmark Detection System and Azure Machine Learning utilized two algorithms: the Two-Class Support Vector Machine and Linear Regression. The control method involving human subjects achieved an accuracy of 76.2% [14]. Facial displays and Non-Verbal features achieved an accuracy of 68.59% through Decision Tree classification and 73.55% through Random Forest tree classification [15].

The support Vector model achieved an accuracy of 67.20% when utilizing Action Units extracted from OpenFace. To ensure the multimodal model's validity and performance, various models and classifiers were tested against it. The model being tested are a subset of the proposed model. Including single modality video classifying with 3D-CNN, Convolutional LSTM 2D, and Holistic landmark Random Forest Classification. Muli-modal model video classification with the combination of 3D-CNN and ConvLSTM-2D. Muli-modal model video classification with the concatenation of 3D-CNN and ConvLSTM-2D and Random Forest Classification. To ensure performed enhancements at each step, the test set being compared is constant.

By employing PittPatt to locate the facial region and an Anthropometric face model, we achieved an accuracy of 76.92% when analyzing macro expressions. Additionally, they obtained an accuracy of 56.92% when examining micro-expressions using a Random Forest model [16]. Visual Features extracted visual features from the videos using a 3D CNN and employing a 3D CNN classifier; we achieved an accuracy of 78.57% [17]. Utilizing a CNN for extracting visual features, combining IDT with a Linear Support Vector Machine achieved an accuracy of 77.31%. Furthermore, using Micro-expressions with a Random Forest model resulted in an accuracy of 80.64%. Using a linear regression model, the integration of both Micro-expressions and IDT reached an accuracy of 89.88% [18]. Visual features extracted from 3D-CNN, which achieved an accuracy of 95.96% when utilized in a Multi-Layer Perceptron model [19]. Visual Features, specifically face landmark detection and head pose estimation using Convolutional Experts Constrained Local Model (CE-CLM), achieved the following accuracies with different classification models: Logistic Regression - 54.86%, Random Forest - 58.42%, SVM (Linear) - 65.56%, SVM (Sigmoid) - 71.99%, SVM (RBF) - 76.84% [20].

Visual features extracted from OpenFace and applied Long Short-Term Memory (LSTM), which achieved an accuracy of 56%. Additionally, utilizing a Support Vector Machine (SVM), an accuracy of 57.4% was obtained [21].

Visual features are extracted from an R-CNN network, enabling the creation of a Face-focused cross-stream network (FFCS). The FFCS incorporates fusion with two inputs: focusing on the face and movement within the network. By incorporating adversarial learning, meta-learning, and cross-stream correlation learning into the model, they have achieved an accuracy of 93.16%. Additionally, the model reached a face detection accuracy of 84.33% [22].

Table 1. Accuracy Comparison of Detection Methods/Models

Method/Model Description	Accuracy	Reference
Two-Class Support Vector Machine and Linear Regression	76.20%	[14]
Decision Tree classification (Facial displays)	68.59%	[15]
Random Forest tree classification (Facial displays)	73.55%	[15]
PittPatt and Anthropometric face model (Macro expressions)	76.92%	[16]
Random Forest model (Micro-expressions)	56.92%	[16]
Visual Features with 3D CNN	78.57%	[17]
Integration of Micro-expressions and IDT (Linear Regression)	89.88%	[18]
Visual Features from 3D-CNN (Multi-Layer Perceptron)	95.96%	[19]
Visual Features (CE-CLM) - Logistic Regression	54.86%	[20]
Visual Features (CE-CLM) - Random Forest	58.42%	[20]
Visual Features (CE-CLM) - SVM (Linear)	65.56%	[20]
Visual Features (CE-CLM) - SVM (Sigmoid)	71.99%	[20]
Visual Features (CE-CLM) - SVM (RBF)	76.84%	[20]
Visual Features from OpenFace - LSTM	56%	[21]
Visual Features from OpenFace - SVM	57.40%	[21]
Face-focused cross-stream network (FFCS)	93.16%	[22]
Face detection accuracy (FFCS)	84.33%	[22]

2.2 Multimodal Features Deception Detection

The multimodal approach uses a combination of different features to train the model:

Verbal Features consisting of unigrams and bigrams, along with Non-Verbal Features involving Facial and Gesture display, an accuracy of 75.20% reached by Decision Tree classification and 50.41% with Random Forest Tree classification [15].

A combination of Lexical Feature Extraction, Audio-based Features, and Visual Features reached an accuracy of 78.95% with Feature-level Fusion, 76.12% with Decision-level Fusion, and 74.02% with Utterance-based Feature Fusion in a Vector Machine model [23].

A combination of Macro and Micro-expression reached an accuracy of 76.92% with Random Forest classification [16].

Audio, visual, and textual features extracted from a Convolutional Neural Network (CNN) and openSMILE were combined using early and late feature fusion techniques. The result was an accuracy of 92% with early fusion and 96.4% when using late fusion [17].

A combination of Acoustic, Gestures, Text Modality, and IDT (Improved Dense Trajectory) features achieved the following accuracies with different classifiers: 87.73% with L-SVM (Linear Supported Vector Machine), 82.33% with K-SVM (K-supported vector Machine), 77.76% with NB (Naive Bayes), 77.77% with DT (Decision Tree), 84.77% with RF (Random Forest Tree), 78.94% with LR (Linear Regression), and 78.99% with AdaBoost [18].

A combination of feature extraction by Acoustic, Gestures, and Text Modalities was used. Gesture feature extraction involved the utilization of a 3D-CNN to identify facial expressions. Textual feature extraction employed a Convolutional Neural Network (CNN). For the audio feature, openSMILE was utilized. The facial expressions were manually annotated into binary features. The algorithm achieved an accuracy of 75.20% for DT (Decision Tree), 50.41% for RF (Random Forest), 90.49% for MLPC (Multi-Layer Perceptron Classifier), and 90.99% for MLPH+C (Multi-Layer Perceptron with Hidden Layer Count) [19].

A combination of Visual, Acoustical, and Textual modalities was employed. OpenFace was utilized for Visual modality extraction, FFmpeg was used to extract the Acoustics modality, and Watson's Speech to Text was employed for textual modality extraction. Through early and late fusion techniques, an accuracy of 0.665 was achieved using Deep Learning, and an accuracy of 0.610 was achieved using LSTM [15].

3. Methodology

3.1 Machine Learning Models

The Multi-Modal model structure (as shown in Figure 2) is a multi-model that uses deep-learning feature extraction to detect deception. Deception Detection requires an architecture that implements sequencing or the Temporal Dimension. Events such as looking away or stuttering are not always cues of deception [24]. While deception in itself does not affect someone's behavior, deceptive indicators are signs of attempting behavior control. This implies that deceptive behavior may be visible if a liar experience telling the truth. There is no generalized deceptive behavior for everyone; some behaviors are more likely to occur than others [25]. Knowing there can be some nonverbal information from deception behavior, the proposed ML model will detect Truthful and deceptive behavior.

The proposed model must be attuned to frame sequences/video and can aptly identify a person. Furthermore, it is why 3D-CNN, 2D-LSTM, and Holistic landmark features are used. 3D-CNN and 2D-LSTM have shown generalized success in representing temporal data [26, 27] Holistic landmarks accurately represent people with set landmarks [28]. Random Forest classifiers determine the relative importance by considering the collective predictions of all constituent decision trees [29]. Since nonverbal deception detection does not generalize and is situational, Random Forest classification is used for the final classification

To ensure the multimodal model's validity and performance, various models and classifiers were tested against it. The model being tested are a subset of the proposed model. Including single modality video classifying with 3D-CNN, Convolutional LSTM 2D, and Holistic landmark Random Forest Classification. Muli-modal model video classification with the combination of 3D-CNN and ConvLSTM-2D. Muli-modal model video classification with the concatenation of 3D-CNN and ConvLSTM-2D and Random Forest Classification. To ensure performed enhancements at each step, the test set being compared is constant.

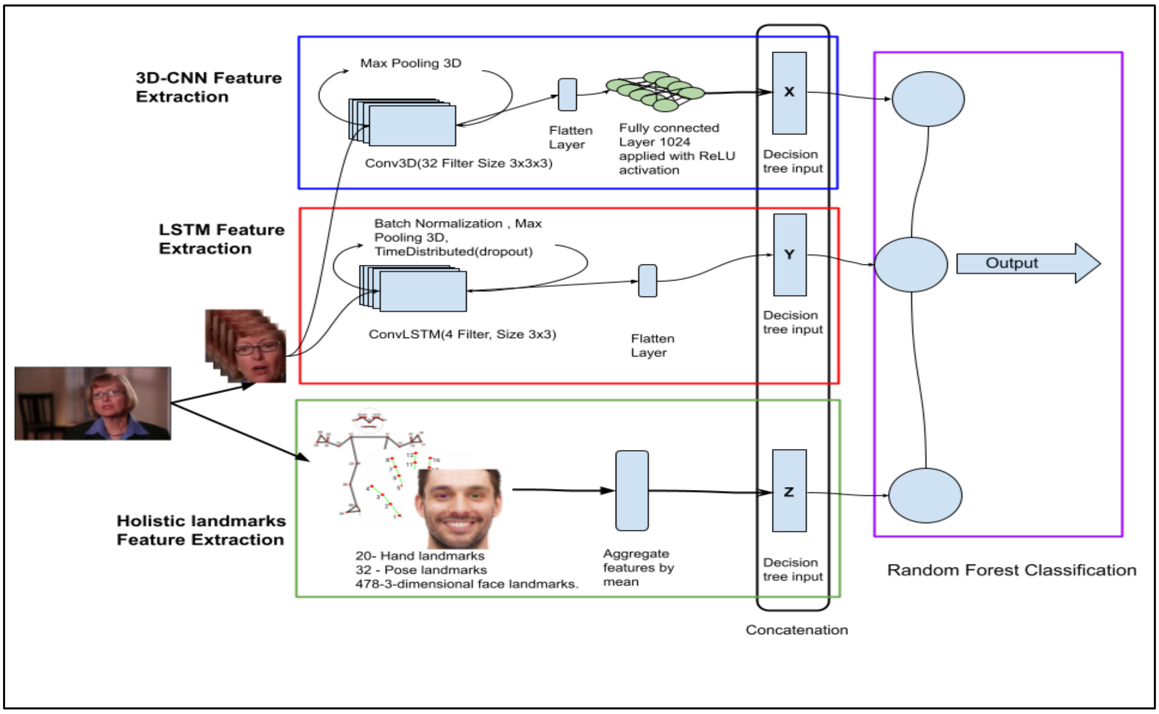

3-CNN Feature Extraction
Max Pooling 3D
Conv3D(32 Filter Size 3x3x3)
Flatten Layer
Fully connected Layer 1024 applied with ReLU activation
X
Decision tree input
LSTM Feature Extraction
Batch Normalization , Max Pooling 3D, TimeDistributed(dropout)
ConvLSTM(4 Filter, Size 3x3)
Flatten Layer
Y
Decision tree input
Output
Holistic landmarks Feature Extraction
20- Hand landmarks
32 - Pose landmarks
478-3-dimensional face landmarks.
Aggregate features by mean
Z
Decision tree input
Concatenation
Random Forest Classification

Figure 2. High-level diagram of the proposed model

3.2 Feature extraction

Studies have shown that there exist some indications of deception [30, 31]. The issues with non-verbal deception detention stem from humans' cognitive complexities and the inability to have every characteristic in a controlled environment. [31, 32].

Three significant gaps have been identified when dealing with Non-verbal Deception research; increase the methodology domain by examination, increase the diagnostic of cues for deception work, and heighten focus where there is no alternative to making nonverbal veracity assessments [3].

With AI, we address the first concerns by creating and testing a variety of ML methodologies against each other; the second concern and third concerns by using deep learning feature extraction, which removes the issues of previous diagnosis cues and lets the model create its own identify features from the given data.

Feature extraction allows a machine learning model to consume identity characteristics for classification. This technique yields better results when compared to feeding raw data to an ML model. [33, 34]. This model uses two Deep learning feature extraction methods, a Conv LSTM 2d and 3D CNN and a Holistic Landmark extraction. The use of automatic feature constructions from Deep learning addresses the issues of finding identified characteristics [33, 34]. We can identify features from the data given with these deep learning models [34].

The last feature used in the modal is MediaPipe Holistic Landmarks which combines components of the pose, face, and hand landmarks to create a complete landmark for the human body. With 3DCnn and CovLSTM2d, we input face frames to get a complete view of the human body Holistic Landmarks were added.

3.3 Paths for Extraction

A Convolutional LSTM 2D (CovLSTM2d) is a deep learning model. It Combines the convolutional structures (CNN) with the feedback connection of the LSTM [33]. Combining the input sequence data from a CNN makes it well-suited for image and video Classification [35]. ConvLSTM2d models explicitly model temporal dependencies using LSTM cells [33]. The Convolutional LSTM, as shown in Figure 3 is a network defined with three different layers. The first layer has 4 filters and uses a 3x3 kernel size with a hyperbolic tangent (tanh) activation function. It employs the 'channels_last' data format and includes a recurrent dropout of 0.2 while returning sequences. After that, a Batch Normalization layer is applied. Following that, a MaxPooling3D layer is used with a pool size of (1,2,2) and 'same' padding in the 'channels_last' data format.

The same structure is repeated two more times but with different filter numbers (8 and 16) for the ConvLSTM2D layers. After each ConvLSTM2D layer, Batch Normalization and MaxPooling3D layers are used again, along with a Time Distributed layer and Dropout layer with a dropout rate of 0.2 applied to each time step.

Finally, a flattened layer is employed to transform the output feature map into a one-dimensional vector, which can be used for further processing or classification tasks.

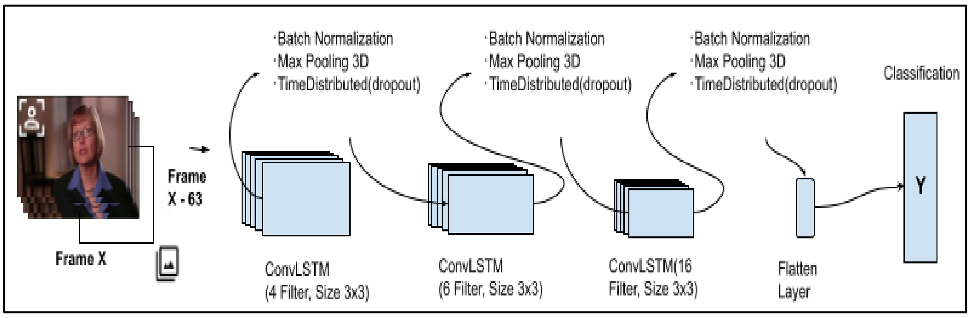


Figure 3. Diagram of Convolutional LSTM 2D path

MediaPipe Holistic Landmarks combines components of the viewing angle of the purpose, face, and hand landmarks to create a complete landmark for the human body.

In this process for Holistic Landmark extraction, the color space is converted to RGB using the function Convert_ColorSpaceToRGB(). Then, the pose landmarks are obtained using the Results.pose_landmarks() method, and the coordinates of each landmark are collected and appended to a list using the append(coordinates_of_each_landmark) operation. After gathering the individual landmark coordinates, they are aggregated to extract additional features using the Aggreatege_feateurs_appended() function. Finally, the resulting output features are obtained and ready for further analysis or usage.

A 3D CNN as shown in Figure 4 uses a three-dimensional filter to perform convolutions. It learns to recognize patterns in videos from Height, Width, and Depth. The depth in this context will be the temporal dimension. 3DCNN treats the temporal dimension in the same way, as the spatial dimension(time)[34, 36]. a 3D convolutional neural network (Conv3D) is described with three layers. The first Conv3D layer has 32 filters and uses a 3x3x3 kernel size with the Rectified Linear Unit (ReLU) activation function. It employs 'same' padding to preserve the spatial dimensions. After that, a MaxPooling3D layer is used with a pool size of (1, 2, 2) to reduce the spatial dimensions while keeping the depth unchanged.

The same structure is repeated two more times, but with different filter numbers (64 and 128) for the Conv3D layers. Each Conv3D layer is followed by a MaxPooling3D layer with the same pooling size.

After the last MaxPooling3D layer, a flattened layer is applied to transform the output feature map into a one-dimensional vector. This vector is then fed into a Dense layer with 1024 units and ReLU activation to perform fully connected processing. The final output represents the extracted features that can be utilized for various tasks like classification or regression.

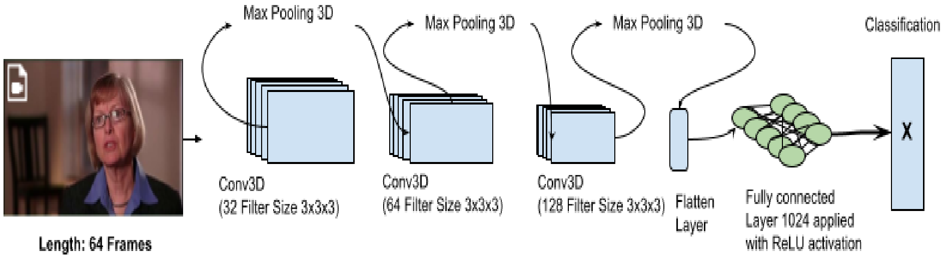


Figure 4. Diagram of 3D CNN path

4. Experiment

4.1 Experiment Setup

The core dataset utilized for detecting deceitful behavior comprises real-life trail-based videos, which have been employed in numerous studies [15]. The database is composed of 121 trial videos split between 61 Deceptive and 60 Truthful. Video labeling was done manually primarily based on the court verdicts, posterior exenteration, verification of police reports against declarations, and other relevant factors [15].

The data was divided as seen in Table 2 into sections to prevent the testing data from mixing with the training data and creating Train-Test Contamination.

Table 2. Division of unedited video files

	Truth	Deception
Training	48 Videos	49 Videos
Testing	12 Videos	12 Videos

There are prominent issues with training data in machine learning, such as Insufficient Data, Non-Representative Training data, Poor Quality Data, or Underfitting the Data [34]. Input diversity and quantity of the dataset are essential for machine learning [24].

4.2 Data augmentation

Data augmentation and cleaning as seen in Figure 5 will be used on the data set to increase our training set. The data cleaning will work in two parts: The first step is to divide all the videos into videos of 64 frames with OpenCV, which creates 97 inputs into a data set of 1095.

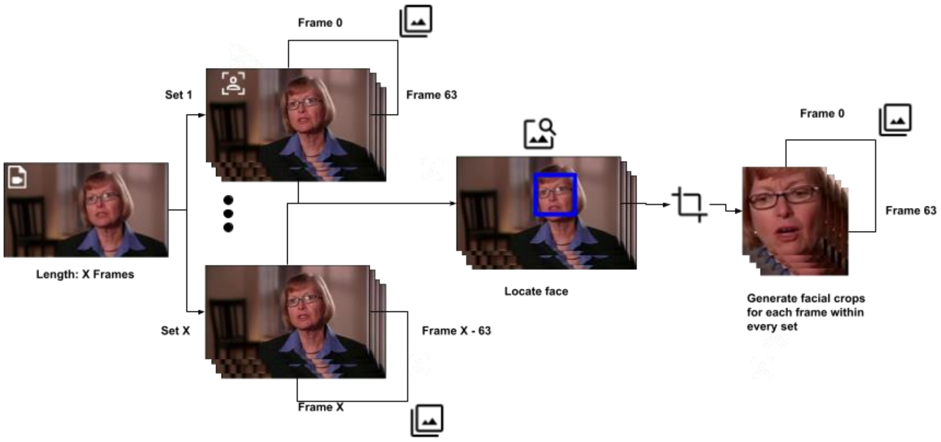


Figure 5. Data Augmentation Diagram

The second step is taking those 1095 videos and capturing all the faces from them using OpenCV. This reduces our data set to 1027 as some faces can not be identified in every video. The full division of the Test/ Train set is seen in Table 3.

Table 3. Division of augmented video files

Instance	Set	Truth	Deception
Training	64 Frames Full Image	493 Videos	534 Videos
	64 Frames Face Crop	493 Videos	534 Videos
Testing	64 Frames Full Image	138 Videos	121 Videos
	Raw Test	12 Videos	12 Videos
	Split-Test	98 Videos	106 Videos

4.3 Uni-Modal model

To ensure the enhanced performance and validity of the multimodal model, each step and classifier within the model was rigorously tested. Firstly, the 3D-CNN deception detection step involved training a 3D-CNN model on face-cropped videos consisting of 64 frames. A total of 20 tests were conducted, with 10 tests performed on the testing set containing 64-frame videos and the remaining 10 on the original video set.

Next, the ConvLSTM-2D deception detection step utilized face-cropped videos with 64 frames to train a ConvLSTM-2D model. Three tests were conducted: the first test involved a 64-frame test set, the second utilized the original video set, and the third involved a training split extracted from the original training set.

For holistic landmarks with random forest classification, face-cropped videos with 64 frames were used to extract holistic landmark features. A random forest classifier with 100 trees was trained using these extracted features. Similarly, three tests were conducted: the first with a 64-frame test set, the second with the original video set, and the third with a training split from the original training set.

Overall, each component of the multimodal model was meticulously tested, and the evaluation was performed under different conditions to ensure its effectiveness and generalization across various datasets and scenarios.

4.4 Multi-Modal model

Each step and classifier were tested to ensure the validity and increased performance of the multi-modal model. The proposed approach involves a Multi-modal model that fuses two distinct pathways, namely, a 3D Convolutional Neural Network (3D-CNN) and a Convolutional LSTM-2D model. The input to this model consists of face-cropped videos containing 64 frames, which are utilized during the training process. The model comprises two parallel paths: Path 1 represents the 3D-CNN model, and Path 2 represents the ConvLSTM model. These two paths operate independently on the input data.

To create a unified representation, the outputs from both Path 1 and Path 2 are concatenated and fed into a dense layer. This dense layer further transforms the aggregated information and produces initial outputs, which are then passed through a softmax function to obtain the final predictions.

To evaluate the performance of the proposed Multi-modal model, three different tests were conducted. The first test used a test set containing 64 frames per video. The second test utilized the original video set as the input. Lastly, the third test involved a training split taken from the original training set. These tests were carried out to assess the model's effectiveness and generalization across various scenarios and datasets.

A multi-modal model was developed, which integrates feature extraction from both a 3D-CNN and a ConvLSTM-2D model, and the classification is performed using a Random Forest classifier. The input to the model consists of face-cropped videos containing 64 frames, which are utilized in a two-path Multi-Modal model. Path 1 represents the 3D-CNN model, while Path 2 represents the ConvLSTM model. These two paths operate independently on the input data.

To create a unified representation of the features extracted by both paths, the outputs from Paths 1 and 2 are concatenated and fed into a dense layer. This dense layer further processes the combined features, and the initial outputs are obtained using a softmax function.

Subsequently, the features extracted from the multi-modal model are used for classification. These features are fitted to a Random Forest classifier comprising 100 trees. The Random Forest classifier allows for effective classification based on the combined information obtained from both the 3D-CNN and ConvLSTM-2D models.

To assess the performance and generalization of the multi-modal model, three different tests were conducted. The first test involved using a test set with 64 frames per video. The second test utilized the original video set as the input. Lastly, the third test involved using a training split extracted from the original training set. These tests were crucial in evaluating the model's capability to detect and classify deception across various datasets and scenarios.

The proposed multi-modal model combines feature extraction from three distinct sources: a 3D-CNN model, a ConvLSTM-2D model, and Holistic landmarks. These features are then integrated and classified using a Random Forest classifier. The model is trained on face-cropped videos containing 64 frames, and it operates within a tree-path Multi-Modal framework.

Path 1 represents the 3D-CNN model, while Path 2 represents the ConvLSTM model. These paths process the input data independently, extracting relevant features from each frame of the video. Additionally, Path 3 involves Holistic landmark feature extraction from the face-cropped video, where the extracted landmarks are aggregated to form meaningful features.

Table 4. Accuracy report of Models

Modality	Model	Model	Set of 64 frame test set	Raw Test set	Split-test
Single	3D-CNN classifier	CNN	0.45173745	0.4166667	0.4854369
	ConvLSTM2D classifier	ConvLSTM2D	0.509652495	0.41666666	0.91176471
	Holistic landmarks with a Random Forest classifier	H + F	0.39	0.38	0.79
Multi	3D-CNN and ConvLSTM-2D	CNN + LSTM	0.6061776	0.625	0.44385027
	3D-CNN and ConvLSTM-2D with a Random Forest Classifier	CNN + LSTM + F	0.49034749	0.54166667	0.7486632
	3D-CNN, ConvLSTM-2D, and Holistic landmarks and was classified with a Random Forest Classifier	Proposed Model	0.558510638	0.54166667	0.87165775

The outputs from Path 1 and Path 2 are concatenated and passed through a dense layer with an initial softmax function, resulting in a unified representation of the extracted features from both 3D-CNN and ConvLSTM-2D models.

To incorporate the Holistic Landmark features, all the extracted features, including those from the 3D-CNN, ConvLSTM-2D, and Holistic landmarks, are concatenated to form a single feature set for each video.

Subsequently, the combined features from all three paths are utilized as input to a random forest classifier consisting of 100 trees. This classifier leverages integrated information to perform deception detection with enhanced accuracy and robustness.

To evaluate the performance and generalization of the multi-modal model, three different tests were conducted. The first test involved using a test set with 64 frames per video. The second test utilized the original video set as the input. Lastly, the third test involved using a training split extracted from the original training set. These tests were crucial in assessing the model's effectiveness in detecting deception across various datasets and scenarios.

5. Results

As shown in Table 4, after completing the experiment, the first subset single model, a 3D-CNN classifier trained with face augmentation, achieved an accuracy of 0.4166667 when tested against raw unprocessed test files. When applied to the processed test set consisting of 64 frames, it reached an accuracy of 0.45173745. Furthermore, when tested with the split training set of videos, it achieved an accuracy of 0.4854369.

The second subset single model, a ConvLSTM2D classifier also trained with face augmentation, achieved an accuracy of 0.4166666567 with raw unprocessed test files. When tested with the processed set of 64 frames, it reached an accuracy of 0.509652495. Furthermore, testing with the split training set of videos yielded an accuracy of 0.91176470588.

The third subset single model employed Holistic landmarks with a Random Forest classifier. Trained with face augmentation, this model achieved an accuracy of 0.38 when tested against raw unprocessed test files. The accuracy rose to 0.39 when the model was applied to the processed set of 64 frames. Additionally, when tested with the split training set of videos, the model reached an accuracy of 0.79.

The first subset multi-modal model, which combines 3D-CNN and ConvLSTM-2D, was trained with face augmentation and achieved an accuracy of 0.625 when tested against raw unprocessed test files. When the model was applied to the processed test set of 64 frames, it reached an accuracy of 0.6061776. Furthermore, testing with the split training set of videos led to an accuracy of 0.4438502673.

The second subset multi-modal model combined feature extraction from 3D-CNN and ConvLSTM-2D and was classified with a Random Forest Classifier. Trained with face augmentation, it reached an accuracy of 0.54166667 when tested against raw unprocessed test files. When applied to the processed test set of 64 frames, it achieved an accuracy of 0.490347490. Furthermore, when tested with the split training set of videos, it reached an accuracy of 0.74866319972.

The proposed multi-modal model integrates three distinct feature extraction methods: 3D-CNN, ConvLSTM-2D, and Holistic landmarks, employing a Random Forest Classifier for classification. Its performance surpasses that of individual models such as CNN or ConvLSTM-2D.The initial accuracy of 0.54 achieved on raw test data indicates the model's proficiency in pattern recognition. A more noteworthy performance emerges when the model is applied to processed data containing 64 frames, yielding an accuracy of 0.56. This indicates an enhanced capacity to comprehend intricate temporal sequences. Additionally, the model demonstrates an accuracy of 0.87 when tested on a segregated training video dataset, highlighting its robust generalization capabilities to unseen data instances. In contrast to standalone CNN or ConvLSTM-2D models, the multi-modal model's merit lies in its amalgamation of techniques, harnessing their respective advantages while mitigating limitations. This amalgamation facilitates the acquisition of both spatial and temporal features, resulting in improved predictive abilities.

6. Conclusion

This study aimed to create a machine-learning model proficient at detecting deception solely through video inputs. We designed an experimental model that leverages 3D-CNN, ConvLSTM-2D, and MediaPipe Holistic Landmark feature extraction methods, followed by classification through Random Forests. We used a dataset of real-life, trial-based videos to train and test our model.

Based on the results in Table 4, single models like 3D-CNN and ConvLSTM2D achieved modest accuracy ranging from 0.42 to 0.91. However, the real breakthrough came with the multi-modal model, integrating 3D-CNN, ConvLSTM-2D, and Holistic landmarks with a Random Forest Classifier. It demonstrated remarkable accuracy, reaching 0.87 on a separate training video dataset, highlighting its ability to generalize effectively. This underscores the potential of multi-modal approaches, surpassing the limitations of individual models like CNN or ConvLSTM-2D and significantly improving predictive performance.

Our findings suggest that while the model performed with lower accuracy on raw test files, it showed promise in the face of inherent challenges. Despite these challenges, the model using Deep learning did surpass the 80% accuracy as shown in Figure 6, indicating that our study's high-level structure with deep learning Emphasis can yield promising results.

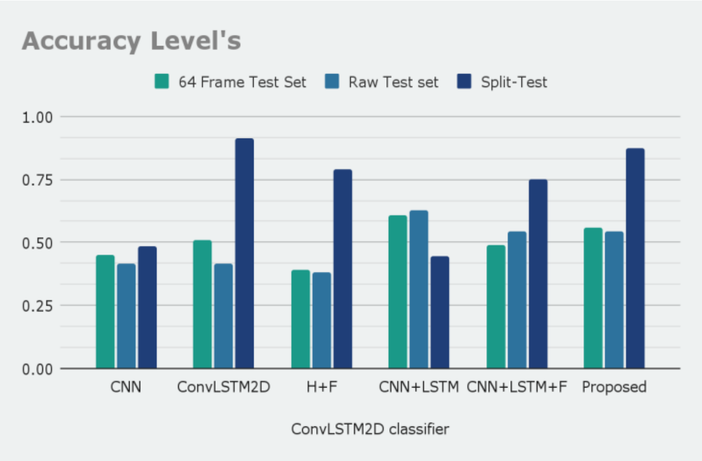


Figure 6. Comparison of accuracy using different models

7. Future Work

However, it is crucial to recognize that certain limitations in our approach could enhance our findings' overall accuracy and generalizability. One such limitation lies in the absence of an exhaustive parameter-tuning process and data limitation. The unpredictable factors in raw test files present a unique obstacle, yet our proposed model exhibited a general accuracy trend in these conditions. Machine learning models typically perform better with larger, more diverse datasets - a factor that might have yet to be fully realized with our current dataset. Parameter tuning is a critical aspect of machine learning models that enhance performance by optimizing their internal parameters or hyperparameters. Hyperparameters are the model elements we adjust before the training begins, such as the learning rate, regularization parameters, the depth of decision trees, etc.

Our results show that the availability of deception detection could increase with the work of machine learning and AI. Nonetheless, it is essential to remember that deception detection is a complex and multi-faceted task. Therefore, the predictions made by a machine learning model should not be the sole determinants of truthfulness but should be used as supplementary tools within a broader context.

8. Acknowledgment

This work was supported in part by the U.S. National Science Foundation under Grant 1757828, Grant 1822137, and Grant 2050869. This work was completed solely by the author at the University of Texas at Dallas Research Experience for Undergraduates (REU) under the supervision of Dr. W. Eric Wong and Mr. Zizhao Chen.

References

1. Shirai, Y. *Three-dimensional computer vision*. Springer Science & Business Media, 2012.
2. British Polygraph Association. (n.d.). Accuracy and Validity of Polygraph Testing. https://polygraph.org.uk/accuracy-and-validity-of-polygraph-testing/, Last accessed in August 2023.
3. LeCun, Y., Bengio, Y. and Hinton, G. Deep learning. *nature*, vol. 521, no. 7553, pp. 436-444, 2015.
4. Krizhevsky, A., Sutskever, I. and Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, vol. 25, pp. 1-9, 2012.
5. Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 3156-3164, 2015.
6. Vrij, A. and Granhag, P.A. Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, vol. 1, no. 2, pp. 110-117, 2012.
7. American Psychological Association. (n.d.). Emotions. APA. https://www.apa.org/topics/emotions, Last accessed in August 2023.
8. Lie Detector Test UK. (n.d.). Lie Detector Test Accuracy. https://liedetectortest.uk/lie-detector-test accuracy#:~:text=Research%20has%20shown%20that%20the,achieve%20this%20level%20of%20accuracy, Last accessed in August 2023
9. Puddifoot, K. Re-evaluating the credibility of eyewitness testimony: the misinformation effect and the overcritical juror. *Episteme*, vol. 17, no. 2, pp. 255-279, 2020.
10. Schwartz, J. As jurors turn to web, mistrials are popping up. *New York Times*, vol. 3, no. 17, pp. 9, 2009.
11. Rensink, R.A. The dynamic representation of scenes. *Visual cognition*, vol. 7, no. 1-3, pp. 17-42, 2000.
12. Ghebreab, S., Scholte, S., Lamme, V. and Smeulders, A. A biologically plausible model for rapid natural scene identification. *Advances in Neural Information Processing Systems*, vol. 22, 2009.
13. Khalid, S., Khalil, T. and Nasreen, S. A survey of feature selection and feature extraction techniques in machine learning. In *2014 science and information conference*, IEEE, pp. 372-378, 2014, August.
14. Azhan, S., Zaman, A. and Bhuiyan, M.R. *Using machine learning for lie detection: classification of human visual morphology* (Doctoral dissertation, BRAC University), 2018.
15. Pérez-Rosas, V., Abouelenien, M., Mihalcea, R. and Burzo, M. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 59-66, 2015.
16. Su, L. and Levine, M. Does "lie to me" lie to you? An evaluation of facial clues to high-stakes deception. *Computer Vision and Image Understanding*, vol. 147, pp. 52-68, 2016.
17. Gogate, M., Adeel, A. and Hussain, A. Deep learning driven multimodal fusion for automated deception detection. In *2017 IEEE symposium series on computational intelligence (SSCI)*, pp. 1-6, IEEE, 2017
18. Wu, Z., Singh, B., Davis, L. and Subrahmanian, V. Deception detection in videos. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
19. Krishnamurthy, G., Majumder, N., Poria, S. and Cambria, E. A deep learning approach for multimodal deception detection. In *International Conference on Computational Linguistics and Intelligent Text Processing*, Cham: Springer Nature Switzerland, pp. 87-96, 2018, March.
20. Avola, D., Cinque, L., Foresti, G.L. and Pannone, D. Automatic deception detection in rgb videos using facial action units. In *Proceedings of the 13th International Conference on Distributed Smart Cameras*, pp. 1-6, 2019, September.
21. Rill-García, R., Jair Escalante, H., Villasenor-Pineda, L. and Reyes-Meza, V. High-level features for multimodal deception detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-9. 2019.
22. Ding, M., Zhao, A., Lu, Z., Xiang, T., and Wen, J.R. Face-focused cross-stream network for deception detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7802-7811, 2019.
23. Jaiswal, M., Tabibu, S. and Bajpai, R. The truth and nothing but the truth: Multimodal analysis for deception detection. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, IEEE, pp. 938-943, 2016, December.
24. Martínez-Fernández, S., Bogner, J., Franch, X., Oriol, M., Siebert, J., Trendowicz, A., Vollmer, A.M. and Wagner, S. Software engineering for AI-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 31, no. 2, pp. 1-59.
25. Vrij, A. Telling and detecting lies. *Applying psychology*, pp. 179-241, 2002.
26. Grinciunaite, A., Gudi, A., Tasli, E. and Den Uyl, M. Human pose estimation in space and time using 3d cnn. In *European Conference on Computer Vision*, Cham: Springer International Publishing, pp. 32-39, 2016, October.
27. Hu, W.S., Li, H.C., Pan, L., Li, W., Tao, R. and Du, Q. Spatial–spectral feature extraction via deep ConvLSTM neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6) pp. 4237-4250, 2020.
28. Zou, X., Zhong, S., Yan, L., Zhao, X., Zhou, J. and Wu, Y. Learning robust facial landmark detection via hierarchical structured ensemble. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 141-150, 2019.
29. Guns, R. and Rousseau, R. Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics*, vol. 101, pp. 1461-1473, 2014.

30. Johnson, L., Lee, K., Fletcher, R., and Wilson, D., Combining Convolutional and Recurrent Neural Networks for Efficient Image Classification.
31. Vrij, A. and Granhag, P.A. Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, vol. 1, no. 2, pp.110-117, 2012.
32. Ekman, P. Lying and nonverbal behavior: Theoretical issues and new findings. *Journal of nonverbal behavior*, vol. 12, pp.163-175, 1988.
33. DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K. and Cooper, H. Cues to deception. *Psychological bulletin*, vol. 129, no. 1, pp. 74, 2003.
34. Dara, S. and Tumma, P. Feature extraction by using deep learning: A survey. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, IEEE, pp. 1795-1801, 2018, March.
35. Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L.A. eds. *Feature extraction: foundations and applications*, vol. 207, Springer, 2008.
36. Yu, Z., Liu, G., Liu, Q. and Deng, J. Spatio-temporal convolutional features with nested LSTM for facial expression recognition. *Neurocomputing*, vol. 317, pp. 50-57, 2018.