International Conference on Innovative Data Communication Technologies and Application (ICIDCA-2023)
IEEE Xplore Part Number: CFP23CR5-ART; ISBN: 979-8-3503-9720-8

2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA) | 979-8-3503-9720-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIDCA56705.2023.10099779

Deception Detection using a Multimodal Stacked Bi-LSTM Model

Puneet Kumar Sehrawat
Department of Information Technology
Delhi Technological University
Delhi, India
puneetsehrawat2000@gmail.com

Rajat Kumar
Department of Information Technology
Delhi Technological University
Delhi, India
mathuria.rajatkumar@gmail.com

Nitish Kumar
Department of Information Technology
Delhi Technological University
Delhi, India
nitishkumar12c@gmail.com

Dinesh Kumar Vishwakarma
Department of Information Technology
Delhi Technological University
Delhi, India
dinesh@dtu.ac.in

***Abstract*— In the scientific community, researchers have recently become interested in the automatic identification of deceptive actions because of the range of fields where it might be advantageous, including criminology or security. Deception detection in conversational speech has drawn much attention in recent years. Given the significant risks associated with trial outcomes, using precise and efficient computational methods to assess the accuracy of court evidence may be very beneficial throughout the decision-making process. This study discusses about spotting deception in trial data from actual cases. Doing this has some challenges associated with creating a robust model that can accurately classify the deception and perform this action as fast as possible. Due to the limited number of videos and datasets available, some models overfit the data. A model should exist which can classify the data using various modalities, i.e., video, audio and text, and be able to work on multiple different datasets with excellent accuracy. This study has used videos from actual trials that were collected from open court proceedings and some videos from other datasets. To design a robust deception detection system that discriminates between witnesses and defendants, genuine and fraudulent testimony, this study investigates the utilization of text, audio and video modalities. By extracting and integrating information about the spoken words from audio, this study can achieve an accuracy of 80% approximately. The proposed model results with a classification accuracy of 96% approximately in an extended approach to perform video transcriptions. The Bag-of-Lies dataset, a multimodal database captured in real-world settings has achieved an accuracy of 85%. The Miami University Deception Detection Dataset focuses on people telling truths and lies about their social relationships, achieved an accuracy of 98.1% on the presented model. The proposed model employs LSTM (Long-Short Term Memory), Bidirectional LSTM (Long-Short Term Memory), CNN (Convolution Neural Network), and RestNet50. The results demonstrate that the proposed algorithm performs better at detecting deception than humans.**

Keywords— Deception Detection, Criminology, CNN, Long Short-Term Memory, RestNet50 (Residual Network 50), Multimodal.

I. INTRODUCTION

Deception is when someone tries to persuade others of a false fact by using misleading proof and engaging in fraud. This can be done by introducing dishonesty, factual misrepresentation, or omissions. Big deception tends to happen when the speakers are involved in their statements. It will have a significant impact, as in trial situations when a deceitful speech may lead to the release of a guilty person. Online reviews that aim to sway consumer decisions, social media posts, etc., may use casual deception to gather supporters for or opponents of a cause, indirectly impacting society. The globalization of economics and improvements in computer technology, which have created new chances for deception, are the main drivers of the growth of network communication in our daily lives. Formats based on text, audio, video and other multiple media fall under several categories of computer-mediated communication. Text-based network communication transfers textual information without audio or visual signals. The potential for receiving misleading communications rises with the volume of data carried over the Internet, making it ineffective and difficult to analyze and monitor such messages manually.

Without specialized tools, humans have a low detection rate of deceit, about 54% [1]. Deceitfully accusing the innocent and releasing the guilty can have serious consequences. For example, in the United States, many court proceedings are filed yearly. In US District Courts, 89,936 criminal cases were filed in 2013, while 80,262 cases were filed in 2014. Additionally, between 2000 and 2013, there were 4.29 exonerations on average each year, up from 3.03 between 1973 and 1999. According to the National Registry of Exonerations, 873 exonerations were reported from 1989 to 2012, each with a tragic history case [2].

Therefore, a reliable and effective method is needed to identify dishonest behaviour and distinguish between liars and truth-tellers. This study's objective highlights an effort to build a multimodal system to identify deception from real-life scenarios. We have used three comprehensive datasets (i) made up of 121 videos of court proceedings, of which 60 are truthful, and 61 are deceptive. (ii) 35 distinct subjects have contributed 325 annotated data sets, with 163 truthful and 162 false statements. (iii) a collection of 320 videos featuring male and female Black and White targets revealing both the truth and lies. These targets were videotaped, speaking openly and covertly about their interpersonal interactions. We have used numerous linguistic elements from the transcription extracted from the speaker's statements and the corresponding spectrograms of the audio signals, which are used to train our model. In the comprehensive approach, we used transcriptions from the videos. The model we created employs LSTM (Long-Short Term Memory), Bidirectional LSTM, CNN (Convolution Neural Network), and RestNet50 (Residual Network 50).

II. RELATED WORKS

Researchers have previously proposed several models for the detection of dishonesty. Examples include physiological techniques like the well-known Polygraph test or the more modern fMRI-based functional testing. According to the authors, these techniques have two drawbacks: (i) they need complex equipment setup. (ii) they are pervasive and demand a skilled operator. Their relevance to everyday life is limited. There is no scientific evidence that the frequently used polygraph test can identify individual deceit. Researchers have suggested behavioural strategies, including spontaneous facial expressions [3]. Since the typical individual lacks the training to recognize these micro-expressions, they are irrelevant to everyday society. Automated deception detection methods are required, and the increased computational capabilities have made it possible to build mechanical systems employing data-driven methodologies. Data-driven strategies also benefit from being discrete and using all accessible information. For instance, they are a perfect fit for deception detection systems since they can function with just video and, when available, text annotations. Other data-driven approaches for deception detection have been developed, using modalities including video, audio, text, electroencephalogram (EEG), and gaze. The identification of misleading information in various domains, such as social media sites, online dating websites, and consumer report websites, has been the subject of multiple research articles on verbal-based deception detection to date. Blob analysis, a technique for detecting non-verbal deceit, has pinpointed dishonest hand gestures by watching people's hand motions [4] or looking at geometric patterns connected to head and hand action [5]. Recently, elements from several modalities have been combined to obtain a set of multimodal features that performs optimally [6]. A multimodal deception dataset of linguistic, thermal, and physiological variables was first described in [7], and a multimodal deception detection algorithm was then created. Some research has employed machine learning methods to generate deception models utilizing linguistic inquiry and word count vocabulary [8]. They have demonstrated the value of using psycholinguistic data for automatic deception detection [9]. Numerous studies have also investigated the link between text syntactic complexity and deceitfulness, following the premise that liars could choose simpler words to hide the truth and make it easier for victims to remember their statements [10].

III. DATASETS USED

A. Real Life Court Trial Dataset

This dataset was collected by finding publicly available mixed media sources where Court hearings videos are available and can be reasonably observed and verified for truthful and deceptive behaviour. This dataset is targeted explicitly at court trial recordings on which part of restrictions established by current data processing methodologies may be applied and observed. There were some criteria for picking videos, like the accused face should be clearly visible in most clips. Video and audio quality should be clear enough to understand what the person is saying, and his facial expression is also identifiable. [16]

Three different court hearing outcomes were considered to correctly flag specific video clips of trials as misleading or truthful convictions. Thus, in the case of a sentence, deceptive videos were collected from the defendant at trial, whereas honest videos were gathered from witnesses at the same court trial. In some instances, misleading recordings were gathered denying the crimes committed by suspects, and truthful videos were filmed of the same suspects answering questions about facts confirmed by police to be true. For eyewitnesses, statements corroborated by police investigations were classified as truth, and statements supporting a guilty suspect were classified as misleading. Exoneration testimony was collected as a statement of truth.

Fig. 1. A sample screenshot showing face presentation and hand gestures from an actual trial clip.

The final dataset contains 121 court hearing recordings, including 61 in which a person is telling a lie and 60 in which a person is telling the truth. The mean duration of videos in the dataset is 28.0 seconds. The dataset consists of videos of 21 unique female and 35 exceptional male speakers. Their age varies approximately between 16 and 60 years. The average video duration is 27.7 seconds and 28.3 seconds for deceptive and truthful videos.

B. Miami University Deception Detection Database

The Miami University Deception Detection Database is a valuable resource for researchers studying deception and the techniques used to identify it, also known as the M3UD database for research studies. It's an open-source database that contains videos of different races and gender telling facts that may be true or a lie. [17]

The videos are accompanied by extensive metadata, providing researchers with detailed information about the circumstances of the deception, including the type of lie, the motivation for lying, and the level of experience of the liar. This database has been used to develop and test various deception detection technologies and techniques, including eye-tracking devices and facial recognition software. Researchers have also used the database to understand better the nonverbal cues people exhibit when they lie, which can help develop more accurate and reliable deception detection methods.

This database includes 320 videos of eighty people (20 White males, 20 Black males, 20 Black females and 20 White females) whose recordings were made while telling facts about their social relationship truthfully or deceptively. Every person was recorded four times telling (Positive, negative, positive, negative, negative truth), making 320 videos fully

crossing statement veracity, statement valence, target gender, and target race.

C. Bag of lies Dataset

The dataset delves into the cognitive dimension of deception and integrates it with visual components, providing a unique perspective on deception detection. Most of the primarily available datasets focus on a single modality, whereas Bag-of-Lies is a multimodal dataset consisting of recordings from 35 subjects. Recordings have been gathered through a carefully designed experiment for automatic lie detection purposes (truth/lie classification)[18]. The study also investigates the benefits of integrating multiple modalities for improved performance on the dataset. The authors contend that this dataset will aid in the development of more effective deception detection algorithms that are better suited to real-world situations

The dataset contains 325 manually annotated recorded videos, with 162 Deceptive and 163 truthful videos. Also, EEG data (13 channels) for 22 unique subjects is available. Participants in the video were volunteer University students who were fluent in English and belonged to different cultures. For the experiment, Unlike the traditional datasets in which Subjects were asked various questions and had to tell a lie or truth, here participants were shown 6 to 10 images.

They were free to respond in any way. 6-10 predetermined images were shown to every participant, and they were asked to briefly explain what they saw either truthfully or deceptively, depending upon what they wanted to say. The recordings duration varies from 3.5 seconds to 42 seconds.

IV. METHODS

Our model employs a Convolution layer, LSTM (Long-Short Term Memory), Bidirectional LSTM, RestNet50 (Residual Network 50) and one hot encoding.

A. Convolution layer

The CNN (Convolution neural networks) in deep learning consists of multiple layers used to discover patterns in various types of data. The convolution layer is one of the most crucial layers where main computations occur. It requires several components to work input data, a filter, and a feature map. This layer applies the convolution operations on the input matrix and passes the resultant matrix to the next layer. It changes all the pixels in the receptive field into a single value [11].

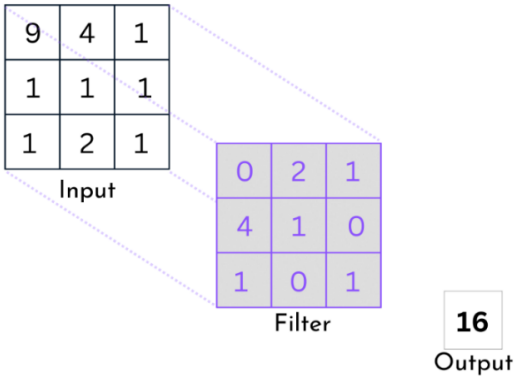

Fig. 2. Convolution Filter

For example, in a large Input image, a small section of the image is considered, and these layers convolve them into a single output using a filter (Kernel). Depending on the type of input and problems required to solve, there are different kinds of convolutions Transposed Convolutions, Dilated or Atrous Convolutions, Separable Convolutions, The 2D Convolution Layer, etc. For our problem, we have used separable convolutions.

B. Long Short-Term Memory (LSTM)

The LSTM (Long-Short Term Memory) is an advanced RNN (Recurrent Neural Network) method that's predominately used for sequential data predictions. It makes it easier to retain past data in memory. It also solves the associated vanishing gradient problem in traditional RNNs. Its applications are language modelling, sentiment analysis, video analysis and speech recognition [12]. LSTM network consists of three different gates:

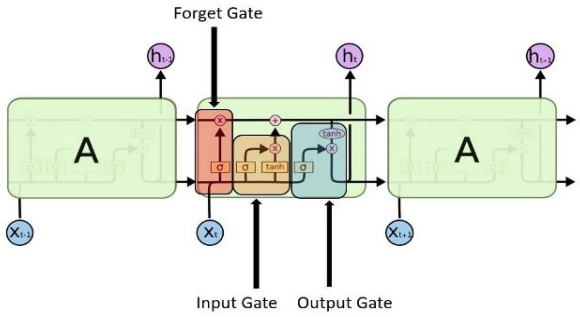

Fig. 3. LSTM cell

$$f_t = \sigma\left(x_t * U_f + H_{t-1} * W_f\right)$$
$$i_t = \sigma\left(x_t * U_i + H_{t-1} * W_i\right)$$
$$o_t = \sigma\left(x_t * U_o + H_{t-1} * W_o\right)$$
$$H_t = o_t(tanh(C_t))$$
$$Hidden\ state = Softmax(H_t)$$

Where t represents the current timestamp, x_t input at timestamp t, U_i Weight matrix of the input, H_{t-1} hidden state value at the previous timestamp, W_i Weight matrix of input associated with the hidden state, σ sigmoid function to bring

output 0 or 1, and i_t, f_t, o_t represents the input, forget, and output gates

C. Bidirectional LSTM

BI-LSTM (Bi-directional long short-term memory) is a recurrent neural network that consists of two LSTMS (long short-term memory). These LSTMs take input in both directions, backwards and forward. It consists of one more LSTM layer than the classical one, which reverses the order of information flow to effectively increase the amount of information given to the network, thereby improving the context available to the algorithm [13]. This implies that BI-LSTM consists of comprehensive and sequential details on every point before and after each end [14].

BI-LSTM architecture has many advantages in solving real-world problems, especially neural language processing. It can produce more meaningful output by combining the result from both the LSTMS. It can be used for NLP tasks like translation recognition, entity recognition and sentence classification; other than that, it also has applications in handwritten recognition, protein structure prediction, speech recognition and similar fields.

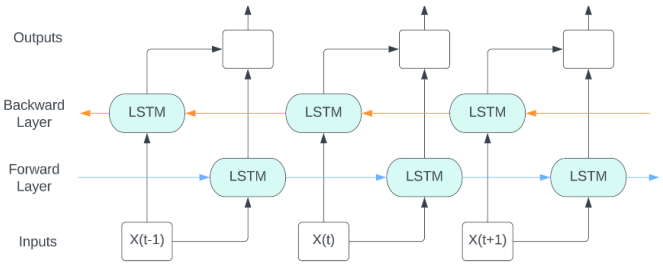

Fig. 4. Bi-LSTM Layer

D. Resnet50

Resnet50 (Residual Network) is a particular type of CNN (convolutional neural network) whose depth is 50 layers that consist of one intermediate pool layer, 1 MaxPool layer and 48 convolutional layers. Resnet 50 is a modified version of Resnet34. The architecture of Resnet50 is the same as that of Resnet 34. However, there is one big difference in the Resnet50 building block. It was modified into a bottleneck design to reduce the time it takes to train layers[15]. That's why the 2-layer building blocks in Resnet 34 are changed by the 3-layer building block making the Resenet50 architecture. Resnet50 is used as the base for many computer vision functions. The most significant advantage that Resnet50 provided was that it allowed the training of intense neural networks with 100+ layers.

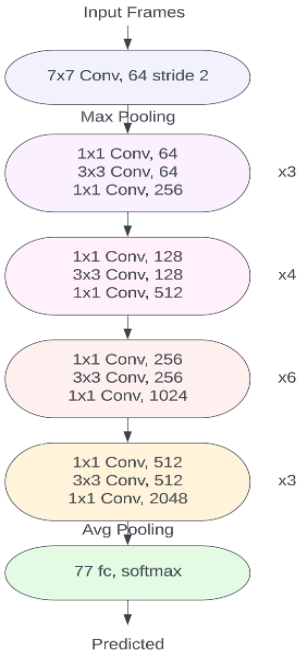

Fig. 5. Resnet50 architecture with Conv(Convolution layer) and fc(fully connected layer) with SoftMax.

V. EXPERIMENT AND RESULTS

We conduct our training experiments on the earlier introduced real-life instances datasets. Given the distribution of false and true clips, we extracted the audio from each video and processed it further. Because stereo audio channels include two channels of audio, meaning that when we're listening, two or more sound sources are concentrated to the left and right. Making it challenging to comprehend the file, we transformed our video's extracted audio fragment to the mono channel. The processing of a mono-channel audio file is more straightforward than that of a multi-channel audio file because mono audio is single-channel audio, which implies that all sound sources are perceived through just one channel.

The Audio files produced are of different durations where audio files might like the voice is loud in some areas and quiet in others. This variation in volume might make transcription difficult. So, we computed maximum audio seconds of about 34 seconds using the summation audio length duration of each audio and flooring against the constant sample rate of the audio. Then, to make all the audio samples the same length, we truncate or expand their length by padding it with silence. We initially tried extracting features using the VGG-16 pretrained model on our dataset. We got a validation accuracy of approximately 75% on the court proceedings dataset. In our alternative method of investigating fine-tuning, instead of utilizing a pretrained model as a feature extractor, we have employed another pretrained model, ResNet50 (Residual Network 50). These transfer learning models have been used because the datasets contain limited videos, resulting in an overfitting model. Hence, these models avoid overfitting and extract essential features from audio.

Deep Learning models rarely accept the raw audio directly as input. Thus we transform the padded audio samples to the Mel

Spectrogram next. Mel spectrogram refers to a spectrogram in which the frequencies are converted to the Mel scale, a pitch measurement system that makes equal gaps in pitch appear to the listener to be equally distant. Mel Spectrograms are frequently the best method for supplying audio data to deep learning models since they capture the fundamental aspects of the audio. The spectrogram height and width used is 224, which is acceptable by the ResNet50 Model. Using overlapping windowed portions of the audio signal, we applied the fast Fourier transform to translate the audio input from the time domain into the frequency domain. We translated the frequency on the y-axis onto the Mel scale. We stored these spectrograms in PNG format in "RGB" mode, as shown in (Figure-6), to feed as input to our model. The Mel-spectrogram offers additional details in the representation of our audio file. With the help of our Mel-spectrograms, we can provide our convolutional neural network model with more information to accurately distinguish between the classes we are training on that are deceptive and truthful in this case.

We extracted the transcriptions from the audio. These transcriptions are the statements made by the speaker during the process. There are 121 statements, comprising 60 truthful and 61 deceptive statements, in the court proceedings dataset. There are 325 statements in the Bag of the Lies dataset, 163 of which are false and 162 which are true. The Miami University Deception Detection dataset has 320 statements, 160 for fraud and 160 for the truthful category clips.

We will use a vocab size of 2000 words for vectorizing and applying one-hot encoding on these statements. The output of each statement will be a vector with values between 0 to 1999 for the distinct set of words in these statements. We are also declaring the size of max length to the floored value of summation of all sentences against the total number of sentences and, on the other side, adding pre-padding to it as 0.

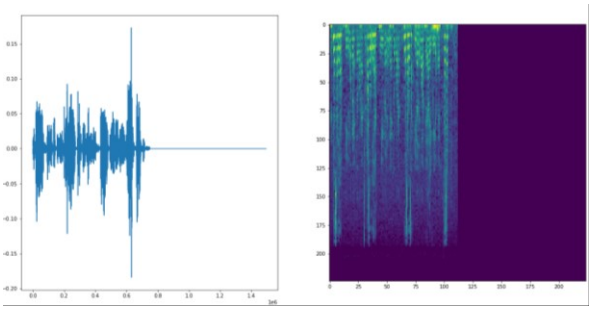

Fig. 6. On the Left, we have the Audio signal waveform, and On the Right, we have Mel Spectrogram (Deceptive Audio)

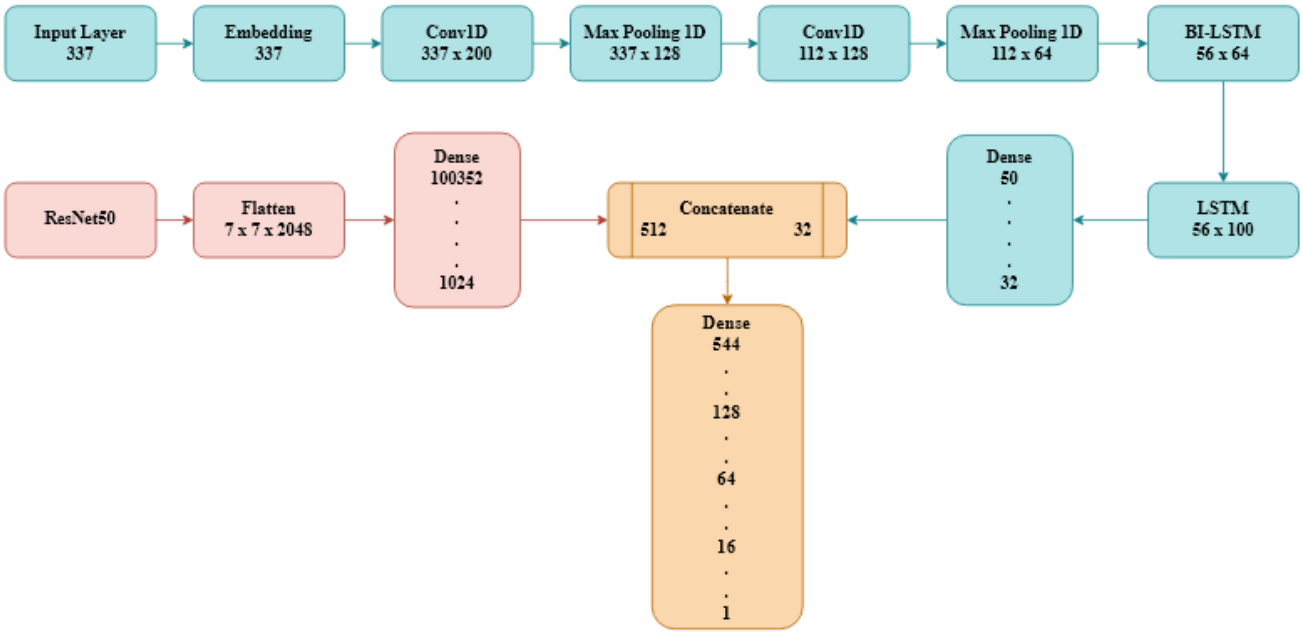

Fig. 7. Working of Model using ResNet50 pre-trained model (Audio + Text based method)

We have used the ResNet50(Residual Network 50) model in our analysis with weights of the trained ImageNet dataset. We have turned each word into a fixed-length vector with a predetermined size due to the embedding layer in the text model branch. In addition to having fewer dimensions, word vectors' constant length aids in improving word representation. With a vocabulary size of 2000 words and a word vector length of 200, we have provided precomputed sentence length as input to the embedding layer. A convolution 1D layer produces a tensor of outputs by combining a convolution kernel with the layer input over a solo temporal dimension. After convolution, some information gets eliminated, so we use the padding "same" such that the output has the same width and height dimensions as the input feature by uniformly adding padding

with zeros to the right and left or down and up of the information. As the convolution process is linear, we use the ReLU (Rectified Linear Unit) activation function to introduce non-linearity. The ReLU function's primary use over other activation functions is that it does not simultaneously activate all neurons. This indicates that the neurons won't stop functioning, except the linear transformation's result is less than 0. The neuron is not triggered if the result is zero for negative input values. The ReLU function is much more computationally efficient because it only activates a limited number of neurons.

By contributing an abstracted version of the illustration, max pooling reduces over-fitting. It chooses the most prominent element from the area of the input feature that the filter has covered. As a result, the output following the max-pooling Layer would consist of a feature map with the most visible elements from the prior feature map. This MaxPooling 2D Layer down samples by the pool size of 3 followed by another Conv1D layer with the same parameters of padding and activation function with 64 output filers in the convolution space and length of 1D convolution window of 2. A second MaxPooling Layer down samples with a pool size of 2. We made two copies of the LSTM hidden layer by enclosing it in a bidirectional layer. In our situation, we will obtain 100 output nodes rather than 50. This stacked Bi-LSTM model finds features in forward as well as backward sequences. Hence Bi-LSTM finds context in the whole series and not limits itself to only the forward direction, which LSTM does. The simple model LSTM in one layer ends with a dense layer of 64 nodes. After which, we drop out 20 per cent of nodes and again end with a dense layer of 32 nodes using the ReLU activation function. The ResNet50 model pre-trained output is flattened, culminating in two dense layers of 1024 and 512 nodes and performing a dropout of 10 per cent after each dense layer. This result of 512 nodes is then concatenated with the text model branch result of 32 nodes to construct a final model branch. We have used an ADAM(Adaptive Moment Estimation Optimizer) optimizer, and the activation function used in the last dense layer is Sigmoid. We perform a split of 80:20 on the dataset. The parameters chosen for predictions are the Mel spectrograms of padded audio samples, encoded transcriptions, and vector of the class-labelled audio spectrogram in which 0 denotes deceptive, and one represents truthful.

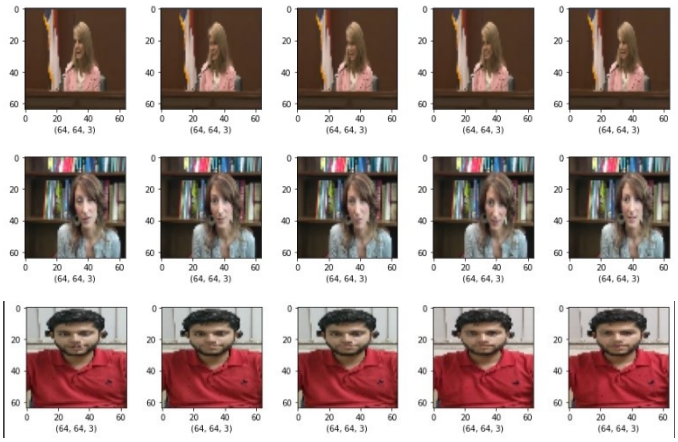

Fig. 8. First five frames of the video resized to dimension 64*64*3 (1st two rows of frames are from the Real-Life Court Trial Dataset. Followed by two rows of Bag of Lies Dataset, and the last two rows contain frames of MU3D Dataset)

During the training of our model, we used a batch size of 4 and employed ReduceOnPlateau call-back with a learning rate of 0.1, which we reduce after every two epochs if there is no improvement till the lower bound. Additionally, we are saving the best model with the highest level of validation accuracy using the Model Checkpoint call-back. We have accomplished a validation accuracy of 80% approximately in this analysis of our model for the court proceedings dataset.

In the comprehensive approach, we know the sequence of images that makes up a video. The human eye will see motion when these pictures, known as frames, are presented continually one at a time at a specific rate. We will construct a parameter to modify how many frames we want to extract and store every second because not all videos have the same duration and FPS. For example, it is set to 10 because, for the time being, it will only store ten frames per second of the video, even though the video FPS is, let's say, 30. A total of 300 frames will be preserved for a video that lasts 30 seconds. We resized the frame to 64*64 size and stored it in a list, as shown in the figure below.

TABLE I. Accuracies obtained for Real Life Court Trails Dataset on different methods.

MODALITIES USED	ACCURACY(Approx.)
Audio + Text using VGG16 method	75%
Audio + Text using ResNet50 method	80%
Video + Text	96%

As we have used different modalities for training our models like text, audio and video, we have encountered that text modality-based model training is faster than the other two as its processing time is less. However, in text-based method learning, very few features are extracted from the input data, leading to the model's poor efficiency. Whereas in model training, in which we used audio and text as our core modality, we were able to accomplish better accuracy. Because the audio

signal as an input is a sound waveform, it requires less processing time than models which uses video and text as their core modality. Still, due to the lack of visual elements, such audio-based trained methods cannot extract more features. Due to this, accuracy could be higher compared to video modality-based trained models. The models which use video modality for learning can extract more insightful features like face gestures which provide more essential elements of the input data, which gives us much better accuracy than the other two. But as we know, video comprises many frames. Hence video-based model training requires more time to process and extract valuable, helpful insights from the data.

We conduct our training experiments on the real-world instances dataset that was previously introduced. We took the audio from each video and processed it further after extracting it, using the same distribution of deceptive and actual clips. From the audio, we retrieved the transcriptions. The speaker's responses throughout the court hearings are captured in these transcriptions. There are 121 assertions, 60 of which are true and 61 of which are false.

The video and text modality model is shown in Fig. 9, where we apply a combination of convolution, normalization, and max pooling followed by Bi-LSTM and dense layers. Convolution is used to extract only the essential features from the frames, and max pooling reduces the size of those extracted features. Bi-LSTM layers find context in both directions as the frames of a video depend on the forward and backward edges. Dense layers then find important information from nodes obtained from Bi-LSTM layers.

We performed the same procedures on the Bag of a Lies dataset, which consists the 325 videos consisting of 163 deceptive and 162 truthful clips of 35 unique subjects to obtain their transcriptions. We collected the transcriptions from the audio segment we took from each subject's video. The same actions are performed on the Miami University Deception Detection Database, containing 320 videos in which 160 clips are truthful and 160 are deceptive. As introduced earlier, this dataset focuses on the truths and lies of males and females speaking about their social relationships. The audio segment is extracted from these clips, and transcriptions are retrieved from the audio part.

During the Training of our model, we employed the Early Stopping call-back method, which stops the Training further after three epochs if there is no improvement, as the goal is to minimize the loss. We have used an ADAM optimizer, and the activation function used in the last dense layer is Sigmoid. We perform a split of 85:15 on the dataset. The parameters chosen for predictions are concatenated array of resized frames of dimension 64*64*3 comprised of deceptive and truthful videos and a connected collection of corresponding labels of these frames.

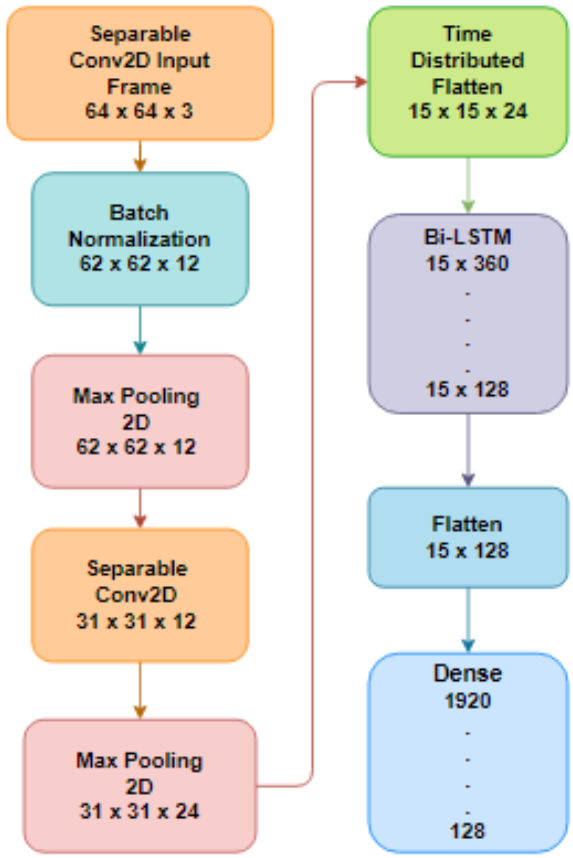

Fig. 9. Working of Model (Video + Text-based method)

The same model was run on the same parameters on all three datasets for Video and Text Modalities. The approximate accuracy recorded for all the datasets is in the table below.

TABLE II. Accuracies were obtained for the three datasets on video and text.

DATASET USED	ACCURACY(Approx.)
Bag of Lies Dataset	85%
Miami University Deception Detection Dataset	98.1%
Real-Life Court Trails Dataset	96%

VI. COMPARISON TO OTHER MODELS

Our paper integrates different modalities with their respective models to find deception in a dataset. Our best performing gave an accuracy of approx. 85%, 95% and 96% for the datasets Bags-of-lie, Miami University Deception Detection Dataset, and Real-Life Court Trails Dataset, respectively.

Many previous models have also been created on the same datasets using different methods. M. Umut S¸en, Veronica P´erez-Rosas, B.Yanikoglu, Mohd. Abouelenien, M. Burzo, and R. Mihalcea, in their paper [19], used Random Forest, SVM(Support Vector Machine) with Radial Basis Function kernel and Neural Network classifiers among which the neural

network achieved the highest accuracy of 84.1%, which consisted of two hidden layers of 100 and 500 nodes with the activation function set as softmax and cross-entropy loss. The model did outperform the average non-expert human performance, but the resulting accuracy is still a little less for a task as crucial as deception detection in court trials. M. Jaiswal, S. Tabibu, and R. Bajpai, in their paper [20], used OpenFace and OpenSmile toolkit, which extracted the audio modality from the video and then used an SVM classifier to get an accuracy of 78.95. The models use facial expressions only and don't consider the person's body language. Both of these models were used for the Real-life court trail Dataset.

V Gupta, M Agarwal, M Arora, T Chakraborty, R Singh and M Vatsa, in their paper [18], used Various models and modalities on the Bags of Lies Dataset. They employed a Random Forest classifier for the video, raw EEG and the gaze, and a KNN (K-Nearest Neighbors) classifier for audio modality, resulting in an accuracy of 66.17%. They evaluated various models for different modalities and, in the end, used the best model for each modality in union with each other. Despite using multiple modalities, the model's accuracy is still pretty low. However, these different models for each modality can be improved or used with other parameters to increase the use cases and accuracy. Ascensión Gallardo-Antolín and Juan Monter, in their paper [21], used an AP Fusion and Late fusion model. Late fusion results are averaged out after undergoing Masking, Attention LSTM and a Hidden layer, while in AP Fusion, results are combined after masking and attention, followed by hidden layers. Of these two, AP Fusion gave the maximum accuracy of 70.55%.

Kun Bu and Kandethody Ramachandran, in their paper [22], used random forest with Stochastic Gradient Boosting to evaluate a model for deception in the Miami University Deception Detection Database (MU3D), which resulted in an accuracy of approx—97%. In summary, existing research on these datasets has used different modalities and different methods to evaluate deception, resulting in different accuracy. Our study uses Video, Audio and text modality with LSTM and Convolution models.

TABLE III. Accuracies obtained in other papers on the datasets (i) Real-Life Court Trails Dataset, (ii) Bags of Lies (BoL), and (iii) Miami University Deception Detection Database (MU3D).

DATASET	METHOD	ACCURACY
Real-Life Court Trails Dataset	Late Fusion in NN	84.1%
Real-Life Court Trails Dataset	SVM with OpenFace and OpenSmile Toolkit	78.95%
BoL Dataset	RF (video, raw EEG and the gaze) + KNN(audio)	66.17%
BoL Dataset	AP fusion with Attention LSTM	70.55%
MU3D	RF(Random Forest) with Stochastic Gradient Boosting	97%

VII.CONCLUSION

This research discussed multidimensional deception detection experiments that used actual, high-stakes deception cases. We used a real-life scenarios dataset to conduct descriptive and analytical tests. The dataset illustrates deceit in a natural environment with little outside interference and numerous modalities, providing opportunities for intriguing research in detecting deception that will provide the basis for developing algorithms to address the issue in real-time. Our examination of false and true videos revealed insights that contribute to deception. We achieved an accuracy of 80% by extracting and integrating information from the spoken words from the audio. In a further comprehensive approach of using transcriptions from videos, we obtain an accuracy of 98.1%. We have accomplished an accuracy of 85% approximately for the Bag of Lies Dataset, 96% and 98.1% for the Real–Life Court Trial Dataset and the Miami University Deception Detection Dataset, respectively, in this analysis of our model.

In the future, these models can further be used to check which question, when asked, leads to high accuracy in detecting deception. A bot can then ask these questions to completely automate the interrogation process in court. The accuracy of the model can also be increased using large datasets for creating a state-of-the-art model which can be used in multiple scenarios.

VIII. REFERENCES

[1] C. F. Bond and B. M. DePaulo, "Accuracy of deception judgments," *Personality and Social Psychology Review*, vol. 10, no. 3. SAGE Publications Inc., pp. 214–234, 2006. doi: 10.1207/s15327957pspr1003_2.

[2] M. Shaffer, S. R. Gross, and R. Warden, "Exonerations in the United States, 1989-2012 Report by the National Registry of Exonerations National Registry of Exonerations," 2012.

[3] T. O. Meservy *et al.*, "IEEE INTELLIGENT SYSTEMS Deception Detection through Automatic, Unobtrusive Analysis of Nonverbal Behavior," 2005. [Online]. Available: www.computer.org/intelligent

[4] S. Lu, G. Tsechpenakis, D. N. Metaxas, M. L. Jensen, and J. Kruse, "Blob Analysis of the Head and Hands: A Method for Deception Detection," 2005.

[5] T. O. Meservy *et al.*, "IEEE INTELLIGENT SYSTEMS Deception Detection through Automatic, Unobtrusive Analysis of Nonverbal Behavior," 2005. [Online]. Available: www.computer.org/intelligent

[6] J. K. Burgoon *et al.*, "Detecting concealment of intent in transportation screening: A proof of concept," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 1, pp. 103–112, 2009, doi: 10.1109/TITS.2008.2011700.

[7] V. Pérez-Rosas, R. Mihalcea, A. Narvaez, and M. Burzo, "A Multimodal Dataset for Deception Detection." [Online]. Available: http://www.thoughttechnology.com/physsuite.html

[8] J. W. Pennebaker and M. E. Francis, "Linguistic Inquiry and Word Count (LIWC)." [Online]. Available: www.erlbaum.com

[9] R. Mihalcea and C. Strapparava, "The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language," 2009.

[10] M. Yancheva and F. Rudzicz, "Automatic detection of deception in child-produced speech using syntactic complexity features."

[11] J. Gu *et al.*, "Recent Advances in Convolutional Neural Networks," Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.07108

[12] S. Hochreiter and J. ¨ Urgen Schmidhuber, "Long Short-Term Memory."

[13] M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," 1997.

[14] Z. Yu *et al.*, "USING BIDIRECTIONAL LSTM RECURRENT NEURAL NETWORKS TO LEARN HIGH-LEVEL ABSTRACTIONS OF SEQUENTIAL FEATURES FOR AUTOMATED SCORING OF NON-NATIVE SPONTANEOUS SPEECH."

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[16] Veronica Perez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Mihai Burzo, Deception Detection using Real-life Trial Data, in Proceedings of the ACM International Conference on Multimodal Interaction (ICMI 2015), Seattle, November 2015.

[17] Lloyd, E. P., Deska, J. C., Hugenberg, K., McConnell, A. R., Humphrey, B., & Kunstman, J. W. "Deception Detection video database" (2017). Miami University

[18] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, M. Vatsa. "Bag-of-Lies: A Multimodal Dataset for Deception Detection", IEEE Conference on Computer Vision and Pattern Recognition Workshop on Challenges and Opportunities for Privacy and Security, 2019.

[19] Sen, Umut Mehmet; Perez-Rosas, Veronica; Yanikoglu, Berrin; Abouelenien, Mohamed; Burzo, Mihai; Mihalcea, Rada (2020). Multimodal Deception Detection using Real-Life Trial Data. IEEE Transactions on Affective Computing, (), 1–1. doi:10.1109/TAFFC.2020.3015684

[20] M. Jaiswal, S. Tabibu, and R. Bajpai, “The truth and nothing but the truth: Multimodal analysis for deception detection,” in Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016.

[21] Gallardo-Antolín, Ascensión, and Juan M. Montero. 2021. "Detecting Deception from Gaze and Speech Using a Multimodal Attention LSTM-Based Framework" Applied Sciences 11, no. 14: 6393.

[22] Kun Bu, Kandethody Ramachandran. Deception Detection using Random Forest-based Ensemble Learning, 18 January 2023, PREPRINT (Version 1) available at Research Square doi:10.21203/rs.3.rs-2460665/v1