Using Artificial Intelligence to Detect Lies During Court Hearings

Bilal Wajid[1], Hamza javed[1], Imran Wajid[2], Danish Wajid[1], Hafsa Rafique[3]

[1]Muhammad Ibn Musa Al-Khwarizmi Research and Development Division, Sabz Qalam, Lahore, Pakistan

[2]Institute of Social Sciences, Istanbul Commerce University, Istanbul, Turkey

[3]Department of Computer Science, University of Management and Technology, Lahore, Pakistan

Correspondence:

Bilal Wajid: bilal.wajid@sse.habib.edu.pk

Article Link: https://www.brainetwork.org/index.php/jcai/article/view/144

DOI: https://doi.org/10.69591/jcai.3.1.4

Volume 3, Issue 1, 2025

Funding
No

Citation: Wajid, B., Javed, H., Wajid, I., Wajid, D., & Rafique, H. (2025). Using artificial intelligence to detect lies during court hearings. *Journal of Computing and Artificial Intelligence, 3*(1).

Conflict of Interest: Authors declared no Conflict of Interest

Acknowledgment: No administrative and technical support was taken for this research

Article History

Submitted: Mar 02, 2025

Last Revised: Apr 27, 2025

Accepted: May 29, 2025

An official Publication of
Beyond Research Advancement & Innovation Network, Islamabad, Pakistan

Using Artificial Intelligence to Detect Lies During Court Hearings

Bilal Wajid[*1], Hamza Javed[1], Imran Wajid[2], Danish Wajid[1], Hafsa Rafique[3]

[1]Muhammad Ibn Musa Al-Khwarizmi Research and Development Division, Sabz Qalam, Lahore, Pakistan

[2]Instiiute of Social Sciences, Istanbul Commerce University, Istanbul, Turkey

[3]Department of Computer Science, University of Management and Technology, Lahore, Pakistan

Abstract

Globally, the judicial system is overworked and under-resourced, with fewer judges, longer trial durations of trials, an ever-increasing number of filed cases, coupled with defendants challenging verdicts of lower courts in upper courts. As false testimonies unnecessarily jeopardize administering justice, this paper focuses on developing an AI-based framework for detecting lies. Our proposed 'AI-based lie detection framework' utilizes text, audio, and visual footage of the courtroom to differentiate between truth and falsehood. In particular, we employ a multilayer perceptron for textual signals, a long-short-term memory neural network model for audio feed, and a two-stream convolution neural network for video classification. Experiments on different neural architectures conclude that our proposed hybrid model outperformed existing techniques with an AUC score of 0.93 for text, 0.88 for audio, and 0.98 for video-based deception detection.

Keywords*:* Deception, CNN, Long Short-term memory networks, Spatio-temporal Information, Acoustic Signals, Spectrograms.

Introduction

"Mr. Lorry asks the witness questions: Ever been kicked? Might have been. Frequently? No. Ever kicked downstairs? Decidedly not; once received a kick at the top of a staircase, and fell downstairs of his own accord." Charles Dickens, A Tale of Two Cities

Lying is a problem, and will forever remain one, despite legal repercussions and social sanctions. In the corporate sector, almost 50% of the companies have experienced fraud in the last two years [1]. With rising social media services, fake news has become rampant [2]. People lie in courts all the time. Politicians and government officials lie even more. Even heads of state will lie to wage war.

"All warfare is based on deception. Hence, when we are able to attack, we must seem unable; when using our forces, we must appear inactive; when we are near, we must make the enemy believe we are far away; when far away, we must make him believe we are near." Sun Tzu, The Art of War Within the legal system, the number of pending cases has risen sharply. Among South-Asian countries, Pakistan has over 2 million cases pending hearing [3], Bangladesh presents a backlog of 3.7 million, while India poses a behemoth of 47 million pending cases. Within the judicial system, the two primary reasons for these large numbers are (i) a smaller number of judges, (ii) the large duration of the trial, and (iii) an ever-increasing number of filed cases.

While the authors feel unable to solve all legal challenges, this humble study focuses on detecting false testimonies in courtrooms [4] where witnesses often lie to hide their involvement, wrongly shift blame, or falsely accuse others for minor gains [5]. As false testimonies undoubtedly jeopardize administrating justice, the authors hope that by introducing computational methods one may detect a liar.

Previously, researchers have employed a system of sensors, and associated hardware to measure physiological indicators to detect whether a person is telling the truth. However, despite considerable work [6], [7], [8], [9], [10], these frameworks are often biased, display poor results [11], and their validity is often questioned by the judges [12].

More recently, researchers are employing textual data [13], visual feed [14], and acoustic information [15] with a machine learning (ML) or artificial intelligence (AI) based system to differentiate between what is true and what is false. For instance, one study presented micro-expressions, such as furrowed eyebrows, as a significant indicator [16], while others showed certain hand gestures [17], modulated frequencies in speech [18], pitch, speaking rate, lexical, and prosodic features [19] as potential markers to differentiate truth from falsehood.

Among learning models, scientists have employed recurrent neural networks [20], [21], convolutional [22], [23], long short-term memory [24], attention [25], and multi-modal networks to detect lies in witnesses' testimony [26]. In general, these models perform better with speech and video taken over time, as opposed to working on static frames [27], [28].

Researchers generally employ the CSC deceptive speech corpus [19] and real-life courtroom trials [11] to develop and verify their frameworks.

Here, we propose a deep-learning framework that employs text, speech, and video taken from courtroom trials [11] to extract relevant features and help detect lies uttered in during hearings.

It is pertinent to state that AI-based frameworks have their limitations and must not function without judicial oversight, as they come with significant legal and ethical implications. Moreover, AI-based tools cannot speed up the legal process; doing otherwise may push harsh sentences or even incorrect judgments altogether, because the underlying datasets that model these AI tools pose significant bias.

The rest of the paper is distributed as follows. After introducing the subject, Section II provides details of the proposed model, with Section III comparing our proposed model with other frameworks. Section IV discusses both legal and ethical implications of an AI-based framework, and lastly, Section V summarizes the contributions of this study.

Methodology

This section elaborates on the proposed architecture summarized in Figure 1.

Dataset

Data comprised real trial hearings, where truthful and deceptive behavior were observed, verified, and annotated [11]. For instance, both the defendant and the witness were identified, their facial expressions were clear and visible and their audio distinct and understandable. The final data comprises 121 videos with 35 male and 21 female speakers, as highlighted below in Table I.

Figure 1: *Lie detection framework: The figure highlights the proposed 'Lie detection framework' which takes the footage as an input and utilizes three separate schemes, (i) Multi-Layer Perceptron (MLP) for text, (ii) Long-Short Term Memory (LSTM) network model for audio, and (iii) Two-stream convolutional neural network for visual inputs.*

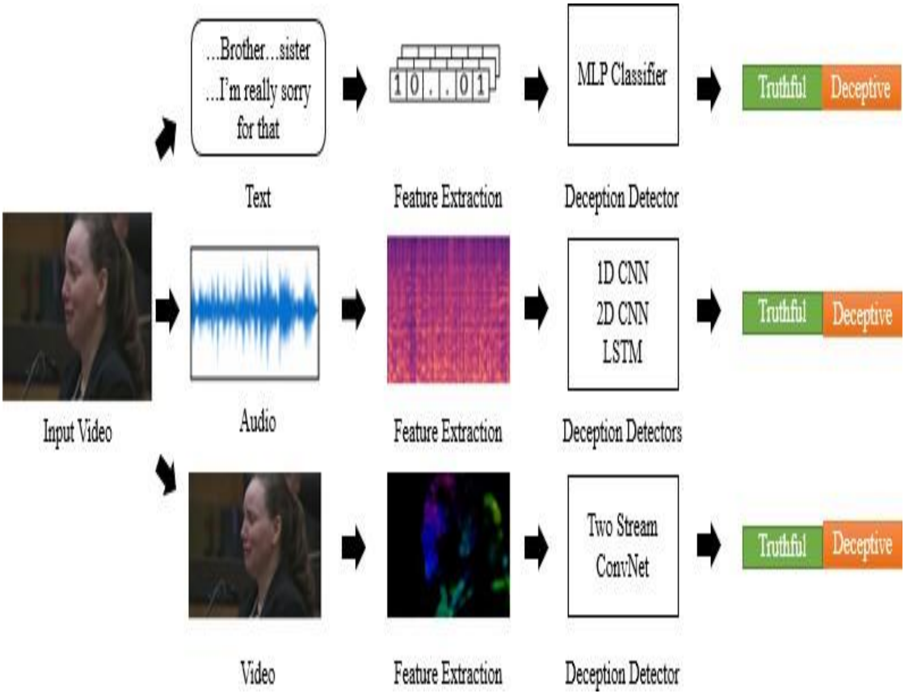


Table 1: *Dataset: The dataset comprises 121 videos, complete with audio and textual components. Cumulatively, it contains 60 truthful and 61 deceptive statements, having an average of 134 and 129 words, respectively.*

S. No.	Type	Number	Ave. duration	Ave. length
1	Truthful	60	28.3 seconds	134 words
2	Deceptive	61	27.7 seconds	129 words

For the purpose of this paper, the authors focus on all three components of the footage, i.e., (i) textual, (ii) acoustic, and (iii) video elements. The text from the videos is transcribed and presented by Perez-Rosas et. al. with careful attention to detail which includes repeated words, fillers, and intentional silence [11]. The final set contains 8,055 words, with an average of 66 words per transcript, where each transcript is classified as either 'truthful' or 'deceptive,' see Table I.

The data diversity statistics of the trial dataset are presented in Table II, showing the distribution of videos across truthful and deceptive categories, along with the gender, age range, and average video lengths of the speakers.

Table 2: *Data diversity: Diversity statistics of the trial dataset.*

Category	Truthful	Deceptive	Total
Number of Videos	60	61	121
Male Speakers	20	15	35
Female Speakers	21	-	21
Age Range (years)	16 – 60	16 – 60	16 – 60
Average Video Length	28.3 seconds	27.7 seconds	28.0 seconds

Pre-processing Transcripts

Although the transcripts provided in the dataset are classified in a binary fashion, i.e., {Truthful ≡ 0, Deceptive ≡ 1}, they are nonetheless of different word lengths. Since machine learning algorithms prefer numbers, a 'vector of numbers' data structure helped store the transcripts in a numeric format. The actual conversion of text to numbers is done via a feature extraction method referred to as a 'bag-of-words' (BOW) model [10].

The BOW model is simple, as it only concerns itself with whether 'known words' are present in the document and not how or where they are used. The intuition of the BOW model is that two texts are similar if they have similar content. It has two components: (i) a vocabulary of known words and (ii) a scoring framework for measuring the presence of known words. Table III showcases an example of building a BOW model.

Table 3: *BOW Example: (A) Document: The above represents some couplets of a poem written by Farid ud-Din Attar in his book Conference of the Birds [29]. Here, the owl (column 1 ≡ document 1) presents itself as an analogy of a man engrossed in the world. Whereas, the hoopoe (column 2 ≡ document 2) represents itself as a guide explaining why the owl's preoccupation is a diversion and that the owl should be making its journey towards Simorgh (an analogy of God). Together, columns 1 and 2 represent two documents. (B) The vocabulary of words: Here, column 1 comprises the entire list of words derived from both documents 1 and 2. Whereas column 2 presents unique words sorted in ascending order. (C) Vector of numbers: This portrayal of text is depicted in terms of {1 ≡ presence, 0 ≡ absence} of words present in the 'vocabulary.' Note that the BoW representation does not care about the order of words and their frequency of occurrence. It only cares whether or not a particular word is present or absent from the document.*

A - The Text	
The owl's excuse The owl approached with his distracted air, Hooting: "Abandoned ruins are my lair, Because, wherever mortal congregate, Strife flourishes and unforgiving hate; A tranquil mind is only to be found Away from men, in wild deserted ground. These ruins are my melancholy pleasure, Not least because they harbour buried treasure. Love for such treasures has directed me To desolate, waste sites; in secrecy I hide my hopes that one fine day my foot Will stumble over unprotected loot. Love for the Simorgh is a childish story; My love is solely for gold's buried glory.	The hoopoe answers him: The hoopoe answered him: "Besotted fool, Suppose you get this gold for which you drool What could you do but guard it night and day While life itself - unnoticed - slips away? The love of gold and jewels is blasphemy; Our faith is wrecked by such idolatry. To love gold is to be an infidel, An idol-worshipper who merits hell. On Judgement Day the miser's secret greed States from his face for everyone to read."

B - Vocabulary

Unsorted list with repeated words	Sorted List with Unique words
the, owl, approached, with, his, distracted, air, the, hoopoe, answered, him, besotted, fool, hooting, abandoned, ruins, are, my, lair, suppose, you, get, this, gold, for, which, you, drool, because, wherever, mortal, congregate, what, could, you, do, but, guard, it, night, and, day, strife, flourishes, and, unforgiving, hate, while, life, itself, unnoticed, slips, away, a, tranquil, mind, is, only, to, be, found, the, love, of, gold, and, jewels, is, blasphemy, away, from, men, in, wild, deserted, ground, our, faith, is, wrecked, by, such, idolatry, these, they, this, to, these, ruins, are, my, melancholy, pleasure, to, love, gold, is, to, be, an, infidel, not, least, because, they, harbour, buried, treasure, an, idol-worshipper, who, merits, hell, love, for, such, treasures, has, directed, me, on, judgement, day, the, miser’s, secret, greed, to, desolate, waste, sites, in, secrecy, states, from, his, face, for, everyone, to, read, I, hide, my, hopes, that, one, fine, day, my, foot, will,	a, abandoned, air, an, and, answered, approached, are, away, be, because, besotted, blasphemy, buried, but, by, childish, congregate, could, day, deserted, desolate, directed, distracted, do, drool, everyone, face, faith, fine, flourishes, fool, foot, for, found, from, get, glory, gold, greed, ground, guard, harbour, has, hate, hell, hide, him, his, hoopoe, hooting, hopes, i, idol-worshipper, idolatry, in, infidel, is, it, itself, jewels, judgement, lair, least, life, loot, love, me, melancholy, men, merits, mind, miser, mortal, my, night, not, of, on, one, only, our, over, owl, pleasure, read, ruins, secrecy, secret. Simorgh, sites, slips, solely, states, story, strife, stumble, such, suppose, that, the, these, they, this, to, tranquil, treasure, unforgiving, unnoticed, unprotected, waste, what, wherever, which, while, who, wild, will, with, wrecked, you

stumble, over, unprotected, loot, love, for, the, Simorgh, is, a, childish, story, my, love, is, solely, for, gold, buried, glory

C - The Vector of Numbers

The owl’s excuse:	The hoopoe answers him:
{1, 1, 1, 0, 1, 0, 1, 1 ,1, 1, 1, 0, 0, 1,	{0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0,
0, 0, 1, 1,	1, 1, 0, 0,
0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0,	1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1,
1, 1, 1	0, 1, 0,
1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1,	1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1,
0, 0, 0, 0, 1, 1,	1, 1, 1, 1, 0, 0,
0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0,	1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1,
0, 1, 1, 0, 1, 1, 1, 0,	1, 0, 0, 1, 0, 0, 0, 1,
1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1,	0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0,
1, 1, 1, 0, 1,	1, 0, 0, 1, 1,
1, 1, 1, 0, 1, 1, 0 1, 0, 0, 0, 1, 1, 1,	0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0,
0, 0}	1, 1}

Text Classification

A feed-forward Multi-Layer Perceptron (MLP) classifier helped differentiate between truthful and deceptive statements. The MLP framework is the simplest deep learning architecture that establishes a relationship, f(x)=y, between input x, which is the Bag-of-Words (BOW) representation of transcripts as a [1273 × 1] vector, and an output class y = Truthful or Deceptive, as shown in Figure 2.

Here, the MLP’s head of size [1273×1] reflects the vocabulary size used in the BOW model, where each entry represents the presence (1) or absence (0) of a unique word from the transcript. Here, stop-words are excluded, and the model does not explicitly account for any specific trigger words, nor does it account for frequency of utterance. It simply transforms the transcript as a single [1273×1] vector.

Pre-processing Audio Signals

Similar to converting texts into their equivalent numeric format, audio signals also need to be transformed into their equivalent numerical representation. As the proposed framework uses three different models for audio classification, each model required a distinct type of input transformation. Specifically, the authors chose the following:

1. Raw audio waveform, sampled at 16 KHz, for a 1D convolution neural network (CNN).
2. Logarithmic frequency spectrogram for 2D-CNN.
3. Mel-spectrogram for long-short-term memory-based NN (LSTM).

Digitally, an audio signal represents a change of amplitude over a period of time. The authors used raw audio signals (in the time domain) for use in their CNN. As for 2D-CNN, the authors applied Fast Fourier Transform (FFT) to convert the time-domain input audio signals to their equivalent frequency domain spectrum while keeping the frequency domain logarithmic, as opposed to linear. As for the LSTM, the mel-spectrogram representation of the audio signals was used. Here ordinary frequency scale is converted into the mel-frequency scale by using the following formula:

$$mel(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right)$$

Eq. (1) helps convert the frequency axis into its mel-scale equivalent. As for the signal's amplitude, it is transformed into decibels to form a mel-spectrogram, which is 2D joint time-frequency representation of the audio signal.

Audio Classification

As discussed above, the framework utilizes three different models for audio classification. Specifically, we employed (i) 1D-CNN with raw audio waveform, (ii) 2D-CNN with Logarithmic frequency spectrogram, and (iii) LSTM with Mel-spectrogram as input, as shown in Figure 3.

1. 1D Convolutional Neural Network (1D-CNN): 1D 1D-CNN architecture is utilized to classify 'deceptive' and 'truthful' testimonies [28]. The architecture comprised an Input layer, an intermittent convolution and pooling layer, followed by a fully connected and output layer as described below:

   (I) Input layer: 1 × 16384 resolution.

   (C1) Convolution layer: 8 → 4 × 1 filters.

   (M1) Max pooling layer: 2 × 2

   (C2) Convolution layer: 16 → 4 × 1 filters.

   (M2) Max pooling layer: 2 × 2

   (C3) Convolution layer: 32 → 4 × 1 filters.

(M3) Max pooling layer: 2 × 2

(C4) Convolution layer: 64 → 4 × 1 filters.

(M4) Max pooling layer: 2 × 2

(C5) Convolution layer: 128 → 4 × 1 filters.

(M5) Global Max pooling layer

(D) Dropout layer: 0.1

(F) Fully Connected layer: 64 neurons that are fully connected to previous layers.

(O) Output layer: 2 neurons connected to previous fully layered neurons.

In the above architecture, all hidden layers were ReLU activated, the convolution layer's filter had adopted a stride of 1; whereas, the output layer used Softmax activation, and the Categorical cross entropy loss function is utilized for Adam optimizer.

2. 2D Convolutional Neural Network: As shown in Table I, the data comprises 60 truthful signals with an average length of 28.3 seconds and 61 deceptive audio signals with an average length of 27.7 seconds. These audio signals were converted into their equivalent log-spectrogram representations using the Scipy and Librosa libraries as input to a 2D-Convolutional Neural Network (2D-CNN). Details of the 2D-CNN architecture are enumerated as follows:

Input layer: The audio signals are presented as a 128 logscales × 124 time window matrix.

(C1) Convolution layer: 8 → 7 × 7 filters.

(M1) Max-pooling layer: 2 × 2

(C2) Convolution layer: 16 → 5 × 5 filters.

(M2) Max-pooling layer: 2 × 2

(C3) Convolution layer: 16 → 3 × 3 filters.

(M3) Max-pooling layer: 2 × 2

(C4) Convolution layer: 32 → 3 × 3 filters.

(M4) Max-pooling layer: 2 × 2

(C5) Convolution layer: 32 → 3 × 3 filters.

(M5) Max-pooling layer: 2 × 2

(C6) Convolution layer: 128 → 4 × 1 filters.

(D) Dropout layer: 0.2

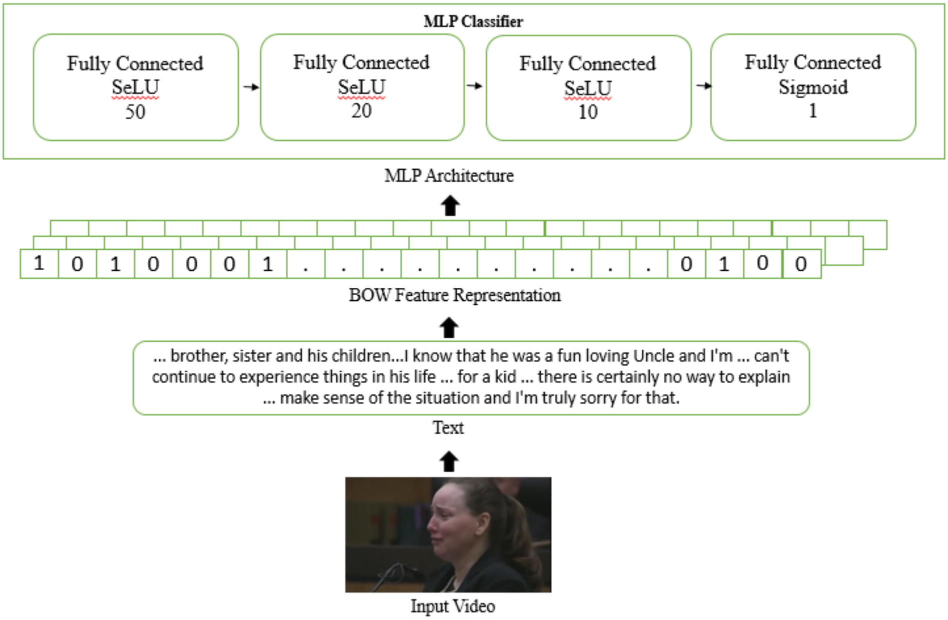
(F) Fully Connected layer: comprising 64 neurons.

(O) Output layer: composing 2 neurons.

In the above 2D-CNN, all neurons of C-1 to C-6 have a stride of 1, whereas all hidden layers use ReLU, while the output layer employs Softmax activation. Here, categorical cross-entropy is minimized via the Adam optimizer.

For each model, the Dropout layer helped reduce overfitting by randomly setting neurons to zero during training, allowing the network to learn robust features. Moreover, MaxPool down-samples the feature maps, retaining the most significant features and reducing spatial dimensions for computational efficiency. Additionally, Batch Normalization normalizes each layer's output across the batch, reducing internal covariant shifts and thereby improving convergence.

Figure 2: *Text Classifier: the framework transformed input transcripts into their equivalent BOW description, which was later fed onto the MLP classifier for segregating between truthful and deceptive statements*.



... brother, sister and his children...I know that he was a fun loving Uncle and I'm ... can't continue to experience things in his life ... for a kid ... there is certainly no way to explain ... make sense of the situation and I'm truly sorry for that.

As highlighted in Fig. 3 (by the token 3×), the proposed architecture repeats the Dropout, MaxPool, and BatchNorm layers three times before reaching the Fully Connected (FC) layer and final Sigmoid activation for binary classification. Here, the 1-D CNN processes a [16384 × 1] vector, while the 2-D CNN and LSTM employ inputs of size [128×124], derived from the time-frequency representations of the audio input.

3. Long-short-term memory neural network model: The proposed framework uses a mel-spectrogram depiction of input audio signals for the LSTM. The mel-spectrogram encompassed 128 bands with frequencies ranging from 0 to 16 kHz. As for LSTM(s), they are a type of Recurrent Neural Network (RNN) such that the output is dependent on current as well as past inputs. This is accomplished by adapting weights and biases throughout the network. Details of the model are presented below, where categorical cross-entropy is minimized via the Adam optimizer:

(I) Input layer: 128 × 124 Mel-Spectrogram feature vector extracted from raw

(II) audio signal.

(L1) LSTM Layer 1: 8 LSTM blocks with Relu activation.

(L2) LSTM Layer 2: 4 LSTM blocks with Relu activation.

(D) Dropout layer: 0.2 to prevent overfitting.

(F) Fully Connected layer:

(O) Output layer: composed of 2 neurons, one for each class.

Video Classification

Visual data contains both temporal and spatial information. Therefore, for classifying video signals, a two-stream convolutional neural network model was used. Here, the first stream captures static data while the second stream captures motion information. This collaborative learning framework is described below:

1)Spatial Stream ConvNet (SSCN): This network recognizes action from still images derived from the video frames. This is particularly useful as certain actions are strongly connected with specific objects.

2)Temporal Stream ConvNet (TSCN): This framework, although similar to that of the SSCN, employs a multi-frame optical flow as an input. This type of input is capable of capturing the trajectory of movements across consecutive frames of the video. Together, both SSCN and TSCN are employed in collaborative learning.

3)Collaborative learning: It may be described as two-way learning where, at any given time t, features extracted from static frames guide the SSCN, whereas dynamic elements extracted from time {t−1, t, t+1}, guide the TSCN. Together, they collaborate (hence, the name collaborative learning) to improve the classification framework of the combined model. The architecture is described as:

(I) Input Layer: 254 × 254 × 3 for input frame.

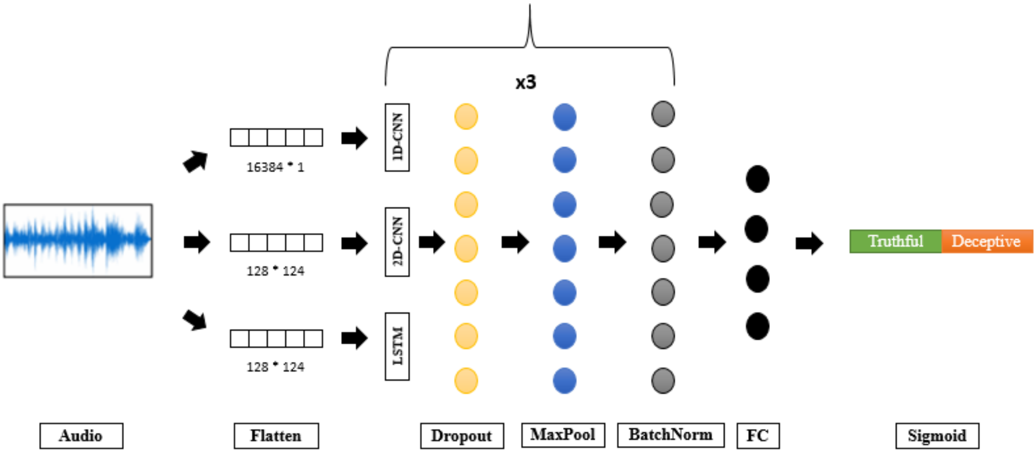
(C) Convolution Layer: 3 convolution layers with filters of 8,32 and 64.

(M) Max Pooling: 2 × 2 concatenated between convolution layers.

(F) Fully Connected Layers: 54 neurons fully connected with previous layers.

(O) Output Layer: 2 neurons, one for truth and the other for deceit.

Figure 3: *Proposed Framework: For audio, our input vector spans 16384 × 1 for 1-D CNN, and 128 × 124 for both 2-D CNN, and LSTM (via its equivalent mel-spectrum). The respective vectors for each architecture go through a Feed-Forward process comprising Dropout, Max Pooling, and Batch Normalization layers three times to finally become our binary output, i.e., truthful, or deceptive.*



The approach to video classification relies on a dual-stream ConvNet architecture, comprising both Spatial Stream ConvNet and a Temporal Stream ConvNet. Here, the Spatial Stream ConvNet operates on single frames capturing static information such as body configurations and not dynamic movements. While the Temporal Stream ConvNet, which uses optical flow, appropriately capturing motion information, as optical flow inherently represents the displacement of key points across multiple frames. This multi-frame analysis is essential for recognizing dynamic actions. Since collaborative learning combines outputs from both streams, spatial (appearance-based) and temporal (motion-based) features, it improves accuracy for classification of 'Truth' or 'Lie.'

Results

Baselines

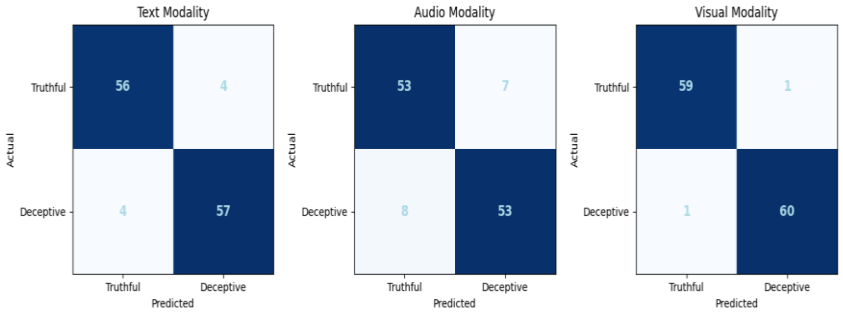
Rosas et al. [11] performed experiments with traditional machine learning (ML) techniques to identify deception in real-life trial data. In particular, the study analyzed visual, text, and acoustic modality for the task of identifying deceit. For text classification, Rosas et al. [11] used unigrams and bigram representation of text, followed by decision trees (DT) and a random forest (RF) classifier to report an accuracy of 60.33%. For capturing lies with visual modality, facial and hand gestures were analyzed [11] with DT and RF reporting an accuracy of 76.03% and 62.04% respectively.

Gogate et al. [33] also performed experiments with the same data [11]. Using an unimodal convolutional neural network, their work outperformed Rosas's framework [11] with 83% accuracy.

Wu et al. [27] utilized traditional ML algorithms and employed all three modalities. Krishnamurthy et al. [28] used CNN to extract textual features and an MLP architecture as a baseline model.

We compare our proposed model with the above-mentioned frameworks for all three text, audio, and visual modalities, as shown in Table V. In addition, we evaluate the model's performance using confusion matrices, measuring the True Positive Rate (TPR) and False Positive Rate (FPR) for each modality as depicted in Figure 4. Furthermore, the Precision-Recall (PR) curve, as shown in Figure 5 for the text, audio, and visual modalities, allows for a deeper analysis of the trade-off between precision and recall.

Figure 4: *Confusion Matrices: For the text, audio, and visual modalities, the True Positive Rate (TPR) and False Positive Rate (FPR) are measured using the Confusion Matrix*



Model Training & Results

For training each data modality on its particular network, we split the data into a 60% training set, 10% validation set, and 30% test set, while employing 5× cross-validation. In addition, we used accuracy to help update the weights of the neurons during back-propagation, along with employing binary cross-entropy loss. For each modality, we used the hyperparameters depicted in Table IV, which were searched using Grid Search. We evaluated the architecture using (i) accuracy and (ii) area under the ROC curve (AUC). Table V shows that the AUC score of the proposed methodology for each text, audio, and visual modality outperformed CNN and traditional machine learning classifiers employed by Wu et al. [27], Krishnamurthy et al. [28], Prez-Rosas et al. [11], and Gogate et al. [33].

Figure 5: *Precision-Recall curves for different modalities: The curves demonstrate the performance of each modality (text, audio, and visual), with the area under the curve (AUC) used to evaluate overall effectiveness.*

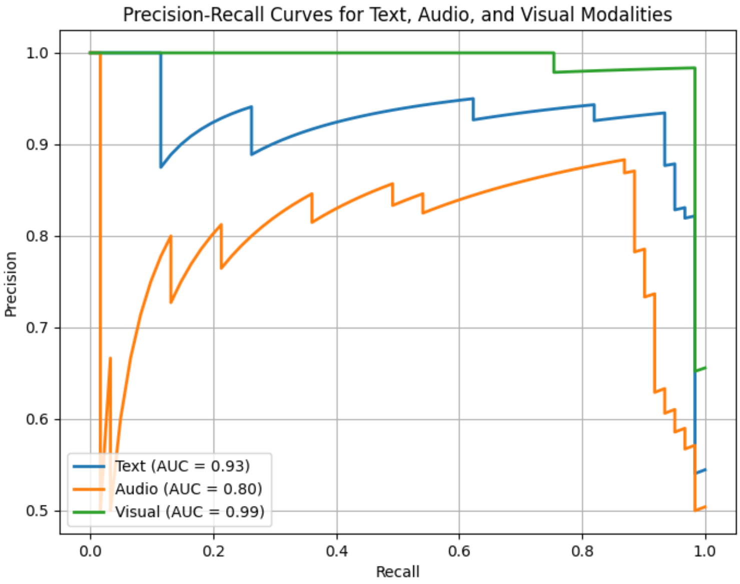


Table 4: *Hyperparameters for proposed framework: Hyperparameters for text, audio, and video streams training*

Hyperparameter	**Text MLP**	**Audio CNN**	**Video 2D CNN**
Learning Rate	0.001	0.0001	0.0001
Batch Size	32	8	4
# of Epochs	50	100	100
Optimizer	SGD	Adam	Adam
Dropout Rate	0.2	0.3	0.5
Weight Decay	0.01	0.01	-
Gradient Clipping	-	5.0	-
Momentum	-	-	0.9
Early Stopping	No	Yes	Yes

Table 5: *Comparison: Comparison of AUC scores of the proposed model with other baseline architectures for deception detection with textual, acoustic, and visual modalities.*

S. No	Method	AUC		
		Text	Audio	Visual
1.	Wu et al. [27]	0.66	0.81	0.75
2.	Krishnamurthy et al. [28]	0.82	0.76	0.95
3.	Prez-Rosas et al. [11]	0.60	-	-
4.	Gogate et al. [33]	0.83	-	-
5.	Proposed	0.93	0.88	0.98

Referring again to Table V, it is evident that among text, audio, and video, the visual modality serves as the most important cue in detecting lies with an AUC of 0.98. Please note, with such a high AUC (0.98) by simply using visual cues, the authors did not need to merge different modalities, as it would increase the number of comparisons and fusions, including (text, audio), (text, visual), (audio, visual), and (text, audio, visual) for improvement. Moreover, as highlighted in Table III, none of the ladies lied in the trial court. Therefore, the mere absence of females from lying renders the proposed solution also gender-biased.

Discussion

Among the methods adopted above, one requires further deliberation. The conversion of text to numbers is done via a 'bag-of-words' (BOW) model, where 1 denotes the presence, and 0 indicates the absence of words in the vocabulary. The BoW approach, while being simple and effective, has several limitations. Firstly, it disregards word order, leading to a loss of syntactic and semantic meaning. This can result in different sentences with similar words being treated identically, despite conveying distinct meanings. Secondly, BoW does not capture contextual relationships, making it ineffective for handling words with multiple meanings. Thirdly, BOW produces sparse and high-dimensional vectors, especially when dealing with large vocabularies, leading to inefficiencies in storage and computation. Lastly, BoW lacks an inherent mechanism for understanding the importance of words within a document, treating all terms equally regardless of their significance in conveying meaning [30], [31], [32], [33].

Moving away from science, one must bear in mind that automated lie detection is inevitably 'high risk' as it incorporates both legal and ethical implications.

Legal perspective

Within the legal domain, courts are indeed overworked and under-resourced; it is not necessarily true that a judicial activity, like determining whether the witness is speaking the truth or not, is the cause for such a predicament. For instance, in South

Asian countries, of whom these authors belong to, much delay is caused by (a) a huge increase in litigation; (b) too few judges to hear such large number of cases; (c) prioritizing criminal over civil cases on the court calendar, (d) an inherent philosophy of procrastination among many lawyers and judges [34], as well as (e) challenging verdicts of lower courts in higher courts.

At the same time, as the study assesses truthfulness in open court, the inherent need to hear the witness/defendant does not go away. Rather, the proposed solution shows as a proof-of-principle, a resource which either raises a 'red flag' in the event of a false testimony, or confirms one, in addition to the judgment of the lawyers arguing, and the jury hearing the case.

In addition, 'all lies are not equal.' For instance, there is a difference when someone said something wrong, but also knew it was wrong, as opposed to when they were unaware of it being wrong.' Moreover, the mere stress of a legal proceeding both facilitates and undermines 'stress-induced' lie indicators. As such, the proposed solution does not address these issues and simply states when the testimony is false.

Lastly, both the practice of lying and its identification is both culture-specific and gender-biased [35], [36]. As the underlying dataset is limited, the proposed solution is biased towards one culture.

Sample Cases where AI-based framework worked poorly

Lawyers and judges are trained to assess consistency of evidence, witness testimonies, and the behavior of individuals across the whole case, while AI simply cannot evaluate with the same depth [37]. Herein below are three cases, where AI-based framework swayed the course of justice:

(I) COMPAS: In Eric Loomis vs. Wisconsin, Eric was harshly sentenced due to a risk assessment, where an AI tool called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) assessed him to be high-risk. The problems with this approach are (i) COMPAS's methodology is proprietary, and hence not transparent. (ii) COMPAS may have been trained upon a dataset that introduced both racial and socioeconomic bias. Lastly, (iii) without human correction, reliance on an AI framework can potentially lead to unjust sentences [38].

(II)PredPol: PredPol was deployed in several U.S. cities to predict crime hotspots and optimize police patrols. However, PredPol relied on historical crime data that reflected racial and socioeconomic bias, resulting in areas already subject to heavy policing continued to be flagged as high-risk, leading to a feedback loop that disproportionately targeted neighborhoods and individuals, often communities of color, leading to residents facing increased scrutiny and arrests [39].

(III) Child Welfare Algorithm: In Tennessee, the Department of Children's Services used an AI-based tool to predict the risk of child abuse and neglect. The algorithm flagged families for potential investigation based on socioeconomic indicators and historical data, which disproportionately affected low-income families and families of color. In some cases, families faced separation due to the AI's risk assessments despite lacking sufficient evidence. When social workers later reviewed these cases, it became clear that the AI's assessments were both biased and inaccurate, leading to unwarranted disruptions in families' lives [40].

Societal Impact

As for societal impact, there are potential benefits and some ethical concerns.

For instance, border check posts may employ AI to help screen individuals at border crossings [41]. In addition, they may help by offering an additional layer of verification in investigations and courtrooms [42], this humble effort being one of them. AI may also be employed to assess candidates for a job by monitoring signs of dishonesty [43].

At the same time, there are serious ethical concerns. For instance, with constant surveillance, privacy is lost [44].

Moreover, AI approaches are as good as the foundational dataset training them. Very often, as highlighted in this paper as well, these datasets have significant gender and cultural biases, rendering one to be unfairly judged [45], leading to discriminatory outcomes. Imagine manipulating the sentiment of the general public towards a new candidate for an election, just because the candidate looks suspicious in an AI model.

Furthermore, fundamental elements like false positives/negatives could lead to wrong convictions within the criminal justice system [46].

Need for datasets

An important element that hinders development in this area is the sheer scarcity of datasets for detecting truths and falsehoods in trial courts. As of now, there are several databases (some subscription-based) providing transcripts of court cases of law for different countries, as shown in Table VI.

There is a need for an effort that extracts key information from the transcripts of the law cases in these databases, to develop an extensive dataset that helps classify 'truth' from 'falsehood.' Such efforts require significant ingenuity and patience to collect a more diverse dataset, one free from gender and cultural biases.

Table 6: *Datasets of sample countries: The table enlists three countries from each continent, and their corresponding source from where transcripts of court cases are available.*

Continent	Country	Name of Website	Website
Asia	Pakistan	Pakistan Law Site	www.pakistanlawsite.com
	Bangladesh India	Supreme Court of Bangladesh	www.supremecourt.gov.bd
		Indian Kanoon	www.indiankanoon.org

Africa	South Africa	Southern African Legal Information Institute	www.saflii.org
	Kenya	Kenya Law	www.kenyalaw.org
	Nigeria	Nigeria Law Reports	www.nigeria-law.org
Europe	United Kingdom	British and Irish Legal Information Institute	www.bailii.org
	Germany	German Legal Information	www.gesetze-im-internet.de
	France	LegiFrance	www.legifrance.gouv.fr
North America	United States	CourtListener	www.courtlistener.com
	Canada	Canadian Legal Information Institute	www.canlii.org
	Mexico	National Supreme Court of Justice	www.scjn.gob.mx
South America	Brazil	JusBrasil	www.jusbrasil.com.br
	Argentina	Centro de Información Judicial	www.cij.gov.ar
	Colombia	Consejo Superior de la Judicatura	www.ramajudicial.gov.co
Oceania	Australia	Australasian Legal Information Institute	www.austlii.edu.au
	New Zealand	New Zealand Legal Information Institute	www.nzlii.org
	Papua New Guinea	Pacific Islands Legal Information Institute	www.paclii.org

Conclusion

This humble effort makes the following contributions, The proposed MLP classifier for text classification outperforms traditional machine learning algorithms utilized for textual deception detection in [11], [27], The Bag of Words representation for textual features with MLP performs better in detecting lies as compared to CNN-based feature extraction techniques in [28], The Mel-Spectrogram representation of audio performs better in detecting falsehood as compared to MFCC and openSMILE-based feature extraction techniques in [27], [28], 2D-CNN and LSTM networks outperform machine learning classifiers and MLP networks in [27], [28] for audio signals, Two two-stream convolutional neural networks for video classification outperform handcrafted feature extraction-based techniques in [27] and [11], Lastly, our proposed 'Lie detection framework' exhibits an AUC of 0.98, which is very promising, Despite employing audio, textual, and visual feeds in catching lies with an AUC of 0.98, there are several limitations to our work. For instance, there is a need for (i) broader range of datasets containing real-life annotated footage of courtrooms, (ii) datasets in different languages and cultures, (iii) datasets for different courts, for instance, family courts, high courts and supreme court, and (iv) datasets pertaining to different types of civil and criminal concerns. By having the above suggested datasets, data scientist like us may be able to build more robust architectures for detecting lies in courtroom trials.

References

[1] M. G. Nakitende, A. Rafay, and M. Waseem, "Frauds in Business Organizations: A Comprehensive Overview," *Handbook of Research on Theory and Practice of Financial Crimes*, pp. 21–38, 2021.

[2] B. Hu, Z. Mao, and Y. Zhang, "An overview of fake news detection: From a new perspective," *Fundamental Research*, vol. 5, no. 1, pp. 332–346, 2025. doi: **10.1016/j.fmre.2023.12.004**

[3] M. Qadeer, Z. Zafarullah, and J. Riaz, "The right to a fair and public hearing in Pakistan," *Contemp. J. Soc. Sci. Rev.*, vol. 3, no. 2, pp. 829–838, 2025.

[4] F. Maisarah, "Witness Trapping in False Testimony Cases: Criminal Law and Qanun Perspectives," *Al-Qadha: Jurnal Hukum dan Syar'iah*, vol. 1, no. 1, 2025. [No DOI available].

[5] N. Jacquemet, S. Luchini, J. Rosaz, and J. F. Shogren, "Truth telling under oath," *Management Science*, vol. 65, no. 1, pp. 426–438, 2019. doi: **10.1287/mnsc.2017.2900**

[6] G. Nguyen, L. Wang, Y. Jiang, and T. Gedeon, "Truth and Trust: Fake News Detection via Biosignals," *arXiv preprint*, arXiv:2505.16702, 2025. doi: **10.48550/arXiv.2505.16702**

[7] B. N. Taha, M. Baykara, and T. B. Alakuş, "Neurophysiological approaches to lie detection: A systematic review," *Brain Sci.*, vol. 15, no. 5, p. 519, 2025. doi: **10.3390/brainsci15050519**

[8] H. Delmas, V. Denault, J. K. Burgoon, and N. E. Dunbar, "A review of automatic lie detection from facial features," *J. Nonverbal Behav.*, vol. 48, no. 1, pp. 93–136, 2024. doi: **10.1007/s10919-023-00438-6**

[9] F. Salice et al., "Nonintrusive Monitoring and Detection of Sitting and Lying Persons: A Technological Review," *IEEE Access*, 2024. doi: **10.1109/ACCESS.2024.3374076**

[10] S. Akhtar, "The Evolution of Natural Language Processing: From Bag of Words to Generative AI," *Journal of Computer Science and Technology Studies*, vol. 7, no. 4, pp. 307–312, 2025. [No DOI available].

[11] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," *Proc. ACM Int. Conf. Multimodal Interaction*, pp. 59–66, 2015. doi: **10.1145/2818346.2820758**

[12] J. A. Oravec, "From Polygraphs to Truth Machines: Artificial Intelligence in Lie Detection," *Critical Humanities*, vol. 2, no. 2, p. 3, 2024.

[13] R. Loconte et al., "Verbal lie detection using large language models," *Scientific Reports*, vol. 13, p. 22849, 2023. doi: **10.1038/s41598-023-50165-y**

[14] J. B. Hirschberg et al., "Distinguishing deceptive from non-deceptive speech," 2005.

[15] R. Mihalcea and C. Strapparava, "The lie detector: Explorations in the automatic recognition of deceptive language," *Proc. ACL-IJCNLP Short Papers*, pp. 309–312, 2009. doi: **10.3115/1667583.1667679**

[16] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*, rev. ed. New York, NY, USA: W.W. Norton, 2009.

[17] E. V. Goncharenko et al., "Gestures-Adapters for persons involved in the crime in non-instrumental lie detection," *ВЕСТНИК*, p. 196.

[18] F. K. Al-Dhaher, D. Y. Mohammed, and M. Khalaf, "The Most Important Features of Lie Detection Using Voice Stress," *Al-Iraqia Journal for Scientific Engineering Research*, vol. 3, no. 1, pp. 93–110, 2024.

[19] S. V. Fernandes and M. S. Ullah, "Use of machine learning for deception detection from spectral and cepstral features of speech signals," *IEEE Access*, vol. 9, pp. 78925–78935, 2021. doi: **10.1109/ACCESS.2021.3083961**

[20] F. M. Talaat, "Explainable enhanced recurrent neural network for lie detection using voice stress analysis," *Multimedia Tools and Applications*, vol. 83, no. 11, pp. 32277–32299, 2024. doi: **10.1007/s11042-024-18653-0**

[21] E. H. Neiterman, M. Bitan, and A. Azaria, "Multilingual Deception Detection by Autonomous Agents," *WebConf Companion*, pp. 480–484, 2020. doi: **10.1145/3366424.3383546**

[22] H. M. Zangana, A. K. Mohammed, and F. M. Mustafa, "Advancements and applications of convolutional neural networks in image analysis: A comprehensive review," *Jurnal Ilmiah Computer Science*, vol. 3, no. 1, pp. 16–29, 2024.

[23] S. Venkatesh et al., "Video based deception detection using deep recurrent convolutional neural network," *CVIP 2019*, pp. 163–169, 2020. doi: **10.1007/978-981-15-4018-9_16**

[24] D. Vavola et al., "LieToMe: Preliminary study on hand gestures for deception detection via Fisher-LSTM," *Pattern Recognition Letters*, vol. 138, pp. 455–461, 2020. doi: **10.1016/j.patrec.2018.10.007**

[25] A. R. Bhamare et al., "Deep Neural Networks for Lie Detection with Attention on Biosignals," *ISCMI 2020*, pp. 143–147, 2020. doi: **10.1109/ISCMI51676.2020.9311591**

[26] V. Karpova et al., "Was It You Who Stole 500 Rubles?-The Multimodal Deception Detection," *ICMI Companion*, pp. 112–119, 2020. doi: **10.1145/3395035.3425237**

[27] Z. Wu, B. Singh, L. Davis, and V. Subrahmanian, "Deception detection in videos," *AAAI*, vol. 32, no. 1, 2018. doi: **10.1609/aaai.v32i1.11342**

[28] D. Grabowski et al., "Multimodal Behavioral Sensors for Lie Detection," *Sensors*, vol. 25, no. 19, p. 6086, 2025. doi: **10.3390/s25196086**

[29] F. U. Attar, *The Conference of the Birds*. London, UK: Penguin, 1984.

[30] F. Ullah et al., "Detecting cybercrimes in accordance with Pakistani law," *LREC-COLING*, pp. 4717–4728, 2024.

[31] A. P. Tuan et al., "Bag of biterms modeling for short texts," *Knowledge and Information Systems*, vol. 62, no. 10, pp. 4055–4090, 2020. doi: **10.1007/s10115-020-01466-z**

[32] A. Muqadas et al., “Deep learning and sentence embeddings for detection of clickbait news,” *Scientific Reports*, vol. 15, p. 13251, 2025. doi: 10.1038/s41598-025-53713-5

[33] I. Amin et al., “Boosting Arabic fake reviews detection…,” *IEEE Access*, 2025.

[34] R. Steinwall, *Life Lessons for Lawyers*. Taylor & Francis, 2024.

[35] F. Quesque et al., “Does culture shape our understanding…,” *Neuropsychology*, vol. 36, no. 7, p. 664, 2022. doi: 10.1037/neu0000792

[36] M. K. Mandal and N. Ambady, “Laterality of facial expressions of emotion,” *Behavioural Neurology*, vol. 15, pp. 23–34, 2004. doi: 10.1155/2004/587059

[37] S. Kim and M. Schulz, “Limits of artificial intelligence in courtrooms,” *Law and Society Review*, vol. 56, no. 4, pp. 587–611, 2022. doi: 10.1111/lasr.12615

[38] J. Angwin et al., “Machine Bias,” *ProPublica*, 2016. [Online]. Available: propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[39] R. Richardson et al., “Dirty data, bad predictions,” *NYUL Rev. Online*, vol. 94, p. 15, 2019. [No DOI].

[40] E. Minoff et al., “Algorithmic Injustice,” CSSP, 2021. [Online]. Available: cssp.org/resource/algorithmic-injustice/

[41] T. Tsujii, “Ethical implications of automatic deception detection technologies,” *Journal of Law and Technology*, 2019.

[42] R. S. A. Faqir, “Digital criminal investigations in the era of AI,” *Int. J. Cyber Criminol.*, vol. 17, no. 2, pp. 77–94, 2023. doi: 10.5281/zenodo.8429904

[43] D. Lyon, *Surveillance Society: Monitoring Everyday Life*. Open Univ. Press, 2016.

[44] D. Murray et al., “The chilling effects of surveillance,” *J. Human Rights Practice*, vol. 16, no. 1, pp. 397–412, 2024. doi: 10.1093/jhuman/huae014

[45] J. K. Burgoon et al., “Application of expectancy violations theory…,” *Int. J. Human-Computer Studies*, vol. 91, pp. 24–36, 2016. doi: 10.1016/j.ijhcs.2016.02.002

[46] L. F. Jones et al., “The validity of reconviction…,” in *Challenging Bias in Forensic Psychological Assessment and Testing*, Routledge, 2022, pp. 69–94.