Deception Detection: An Ensemble and Spatio-Temporal Approach

Junaid Sayani*
Computer Science
National University of Computer and Emerging Sciences
Karachi, Pakistan
junaidsayani03@gmail.com

Bilal Hassan*
Computer Science
National University of Computer and Emerging Sciences
Karachi, Pakistan
bilalhassan2103@gmail.com

Daniyal Khan*
Computer Science
National University of Computer and Emerging Sciences
Karachi, Pakistan
daniyalgopang58@gmail.com

Dr. Muhammad Atif Tahir
Computer Science
National University of Computer and Emerging Sciences
Karachi, Pakistan
atiftahir@iba.edu.pk

***Abstract*—Deception detection in police interrogations has traditionally relied on subjective assessments based on verbal cues and intuition. This research presents a video-based deception detection system leveraging deep learning to analyze facial micro-expressions and action units (AU's) for an objective evaluation. Initially implemented using a simple Long Short-Term Memory (LSTM) network, our approach evolved to incorporate ensemble learning techniques such as bagging, boosting, and stacking, as well as Vision Transformers, specifically TimeSformers, to enhance temporal and spatial feature extraction. Our system processes low-resolution video feeds, extracts meaningful AUs, and classifies behavior as truthful or deceptive. By leveraging advanced deep learning architectures, our method improves interpretability, scalability, and robustness while ensuring ethical compliance. The findings contribute to the development of AI-driven tools for forensic psychology and investigative integrity.**

***Index Terms*—Deception Detection, Vision Transformers, Facial Expressions, Temporal Sequence Modeling**

I. Introduction

Deception detection plays a crucial role in high-stakes environments such as police interrogations, security screenings, and forensic investigations. Traditional methods rely heavily on verbal cues, physiological signals, and human intuition, which are often subjective and prone to bias. While polygraph tests and behavioral analysis techniques have been used, they lack consistency and are not always admissible in legal proceedings. Recent advancements in artificial intelligence (AI) and deep learning have facilitated the development of more objective and automated deception detection systems, focusing on subtle facial micro-expressions and behavioral cues.

Facial Action Units (AUs), which represent muscle movements linked to emotions and deception, have been widely studied in deception analysis. Prior research has utilized Long Short-Term Memory (LSTM) networks to capture temporal dependencies in facial expressions over time. However, LSTMs struggle with modeling both short-term spatial relationships within individual frames and long-range temporal dependencies across video sequences. Additionally, they exhibit limitations in generalizing across diverse datasets. To address these challenges, we enhance the deception detection framework by integrating ensemble learning techniques like, bagging, boosting, and stacking as well as newer deep learning architecture like Vision Transformers, specifically TimeSformers. These architectures process video frames as sequences of patches, effectively capturing both spatial and temporal patterns, leading to improved interpretability and robustness.

Our system processes low-resolution video feeds, extracts meaningful AUs, and classifies behaviors as truthful or deceptive. Designed for scalability, it ensures compliance with privacy regulations while mitigating algorithmic biases. By leveraging state-of-the-art deep learning techniques, this research advances AI-driven behavioral analysis, offering a more reliable and objective tool for law enforcement.

II. Related Work

Deception detection in videos is a complex task that involves analyzing both verbal and non-verbal cues. Several methodologies have been explored in recent research, each with its own strengths and limitations. These approaches can be categorized into the following key areas:

A. Feature Extraction Methods

Since raw video data is often noisy and high-dimensional, feature extraction plays a critical role in deception detection. Researchers have used various techniques to extract meaningful information from videos, including:

- **Facial Action Coding System (FACS):** FACS is a standardized framework for analyzing facial expressions by identifying distinct Action Units (AUs)—specific muscle movements that correspond to different emotions and behaviors. In deception detection, FACS helps recognize subtle facial microexpressions and involuntary muscle activations that may indicate deceptive behavior.

[1] employed an automated Action Unit (AU) detection framework to analyze facial expressions for deception detection. Their approach utilized OpenFace, a pretrained AU recognition model to extract facial action descriptors, which were subsequently used to infer deceptive behavior. [7] introduced a fully automatic system for AU detection and temporal analysis, utilizing spatio-temporal features from tracked facial points and SVM classifiers. Their method achieved a 90.2% agreement rate with human coders for 15 AUs

- **Conversational Turn Segmentation:** Conversational Turn Segmentation is the process of dividing a video or audio recording into distinct segments based on shifts in speaker turns. It helps identify when one speaker stops and another begins, allowing for a structured analysis of dialogue patterns, speech dynamics, and behavioral cues, which can be useful in deception detection and other conversational analyses. While [3] discusses conventional voice activity detection and dialog diarization to accomplish this task, [2] concluded that a Markov model combining results of the modulation spectrum analysis and Kullback-Leibler divergence of adjacent signal portions produces the best predictions.
- **Temporal Chunking:** Temporal Chunking is the process of dividing a video into fixed-duration segments to analyze time-dependent patterns more effectively. This technique enhances time-series modeling by capturing sequential changes in facial expressions, speech, and gestures, making it useful for tasks like deception detection and behavioral analysis. The authors in [1] divided their dataset videos into chunks of 30 frames and then provided the AUs of these chunks directly to their LSTM model. [8] Introduces a time series decomposition algorithm for temporal segmentation, which outperforms existing methods in precision and recall. [9] Proposes a dynamic chunking approach that maps sequences of varying lengths into fixed-number chunks, improving recognition accuracy and computational efficiency.
- **Physiological Feature Extraction:** Physiological Feature Extraction involves capturing and analyzing biological signals, such as heart rate and blood oxygen levels, to identify potential indicators of deception. These physiological responses, which can change under stress or anxiety, are often used alongside other behavioral cues to provide a more comprehensive understanding of a person's emotional state or truthfulness. Multiple studies have utilized the Deception Detection and Physiological Monitoring (DDPM) dataset, which includes recordings of visible, near-infrared, and thermal video, along with cardiac pulse and blood oxygenation data [4]. Heart rate estimation from face videos achieved mean absolute errors as low as 2-3 beats per minute [4]. Earlier studies investigated the effectiveness of multiple physiological indices, including breathing patterns, galvanic skin response, and blood pressure [5], [6]. While all measured variables showed significant indications of deception, galvanic skin response and skin potential response consistently demonstrated superior discrimination [5], [6]. Combining multiple physiological measures may provide better deception detection results than individual indices alone [4], [5].

B. Machine Learning Approaches

Early studies applied traditional machine learning classifiers to detect deception based on extracted features:

- **Random Forest (RF) & Support Vector Machines (SVM):** RF is an ensemble learning method that builds multiple decision trees during training and combines their outputs for classification or regression tasks. It reduces overfitting and improves accuracy by averaging predictions from a large number of trees, each trained on a random subset of the data. SVMs on the other hand are supervised learning algorithms used for classification and regression tasks. They work by finding the hyperplane that best separates different classes in the feature space. SVMs aim to maximize the margin between classes, which helps in improving the model's generalization to new data. Both RF and SVMs have shown promising results in deception detection across various domains. These machine learning algorithms have been applied to detect Advanced Fee Fraud in emails [10], identify deceptive messages [11], and detect credit card fraud [12]. While both RF and SVM have shown satisfactory results, SVM has exhibited superior performance in some cases [10]. These algorithms have consistently outperformed conventional methods, increasing overall accuracy and reducing false positives [11].
- **Neural Networks (NNs):** Neural Networks are computational models inspired by the human brain, consisting of layers of interconnected nodes (neurons) that process and learn from data. They are used for various tasks like classification, regression, and pattern recognition by adjusting the weights of connections through training to minimize error. Recent research explores the application of neural networks for deception detection, showing promising results. Convolutional neural networks using linguistic and physiological data have outperformed traditional classification methods [13]. Neural network models, including LSTM and MTL-NN, have achieved up to 62% accuracy in identifying deception using physiological signals, surpassing human performance [14]. In polygraph scoring, neural networks have shown potential for enhancing accuracy, though further research is needed to validate findings and determine optimal integration methods [15]
- **Handcrafted Feature-Based Models:** Handcrafted Feature-Based Models are machine learning models that rely on manually designed features extracted from raw data. These features are selected based on domain knowledge and are used as inputs to traditional algorithms, such as decision trees or support vector machines, for tasks like classification or regression. Recent research has

explored handcrafted feature-based models for deception detection using various non-verbal cues. [16] developed an LSTM-based approach utilizing hand gesture features, achieving state-of-the-art performance in unimodal hand-gesture deception detection. [17] proposed a framework using Facial Action Unit and Gaze signals, offering interpretability through attention mechanisms. [18] investigated fine-grained eye and facial micro-movements, finding that eye movements provide significant clues for automated deception detection. [19] developed a framework based on eye movement changes, using a dynamic Bayesian model to measure deviations from normal behavior during critical points in interrogations. Their approach achieved 82.5% accuracy, suggesting that eye movement parameters effectively capture behavioral changes associated with deception.

C. Deep Learning Models

With the rise of deep learning, more advanced models have been employed to analyze deception in video data:

- **Convolutional Neural Networks (CNNs):** CNNs operate by applying convolutional filters to extract hierarchical spatial features, making them highly effective in image-based tasks. However, deception detection often requires an understanding of subtle facial expressions and microexpressions that evolve over time. CNNs, by design, lack inherent mechanisms to model temporal sequences, as they process frames independently without retaining contextual relationships between them. Several studies have attempted to mitigate this limitation by integrating recurrent architectures, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), with CNNs. Works like [20, 21] explored CNN-LSTM hybrids to capture both spatial and temporal features in deception detection. These models leverage CNNs for spatial representation and LSTMs for sequential learning. However, despite these advancements, challenges such as vanishing gradients and computational inefficiencies persist, affecting the real-time applicability of such models.
- **Long Short-Term Memory Networks (LSTMs):** LSTMs are capable of learning long-term dependencies by using special memory cells that can retain information over extended periods, making them effective for tasks involving sequential data. Recent research has explored LSTM-based approaches for deception detection using various modalities. [16] proposed an LSTM system that analyzes hand movements in RGB videos, achieving state-of-the-art performance in unimodal hand-gesture deception detection. [22] developed a multimodal attention LSTM framework combining gaze and speech features, outperforming single-modal systems. [23] introduced a convolutional bidirectional LSTM model using acoustic features, achieving 70.3% accuracy on the Columbia-SRI-Colorado corpus. LSTMs have been widely used but they often face challenges in long-range sequence modeling.

D. Multimodal Approaches

Recent work highlights the effectiveness of combining multiple modalities for deception detection:

- **Speech & Facial Expressions Analysis:** Multimodal systems systems combine features from multiple modalities, including visual, acoustic, and linguistic, to improve accuracy in detecting deception. [24] proposed a framework using audio, text, and non-verbal features, achieving 97% accuracy on a real-life deception video dataset. [25] developed a deep learning approach incorporating micro-expression features, reaching 96.14% accuracy. [26] introduced an emotional state-based feature and achieved 91.67% accuracy, addressing data scarcity and preprocessing challenges. [27] applied multimodal deception detection to real-life trial data, achieving 83.05% accuracy and outperforming non-expert human capability.
- **Physiological Signals:** Recent research has focused on developing multimodal deception detection systems that incorporate multiple physiological signals, aiming to improve accuracy over traditional polygraph tests. These systems integrate features from various modalities, including physiological sensors, thermal imaging, speech, and linguistic analysis [28]. Studies have shown that combining features from different modalities significantly enhances deception detection compared to single-modality approaches [28,29]. Researchers have created multimodal datasets containing physiological, thermal, and visual responses under deceptive scenarios to facilitate this work [28]. Non-contact modalities, such as thermal imaging, have demonstrated comparable or superior performance to contact-based methods [28].

Multimodal fusion methods, such as those tested on datasets like Bag-of-Lies, show promise in capturing deception more accurately by leveraging both verbal and non-verbal indicators.

III. Methodology

A. Overview

To systematically analyze deception detection, we explore four distinct architectural approaches: (1) a simple LSTM baseline, (2) an ensemble of LSTMs using Bagging, (3) an ensemble of LSTMs using Boosting, (4) an ensemble of LSTMs using Stacking, and (4) a Vision Transformer (TimeSformer) model. Each model is trained and evaluated independently to assess its effectiveness in capturing deception-related facial dynamics.

B. LSTM-Based Approaches

*1) **Baseline LSTM Model**:* The base model follows the architecture proposed in prior research, leveraging a single Long Short-Term Memory (LSTM) network to model temporal dependencies in facial expressions. The input consists of sequential Action Units (extracted in chunks, from the CLNF), and the model outputs a binary classification (deceptive or truthful) for each different frame vector of AUs. Despite its ability to capture short-term dependencies, this approach

struggles with long-range temporal dependencies (that may lie hundreds of frames ahead) and generalization.

*2) **Ensemble LSTM Approaches**:* To improve the baseline model's robustness, we introduce ensemble learning techniques:

*a) **Bagging** (Bootstrap Aggregating):* We train multiple LSTM models on different subsets of the dataset, each learning independent patterns in deception cues. The final prediction is determined via majority voting:

$$\hat{y} = \text{mode}\{f_1(X), f_2(X), ..., f_n(X)\} \tag{1}$$

where $f_i(X)$ represents the prediction from the i-th LSTM model.

*b) **Boosting**:* Boosting is employed by training LSTMs sequentially, where each model focuses on correcting the errors made by its predecessor. The process involves:

1) Training multiple LSTM classifiers sequentially.
2) Assigning higher weights to misclassified samples after each iteration.
3) Using a weighted sum of predictions from all models to make the final classification.

The final classification score is computed as a weighted sum of individual model outputs:

$$\hat{y} = \sum_{i=1}^{N} \alpha_i f_i(X) \tag{2}$$

where α_i represents the weight assigned to model i based on its accuracy.

*c) **Stacking**:* Stacking combines the predictions of multiple LSTMs by training a meta-classifier on their outputs. Unlike Bagging and Boosting, where models work independently or sequentially, Stacking learns how to optimally combine multiple models' outputs. The process involves:

- Training multiple LSTM classifiers ($f_1, f_2, ..., f_N$) on the dataset.
- Collecting their predictions as input features for a secondary model.
- Using a **meta-classifier** (an LSTM in our case) to learn the best combination of predictions.

The final classification is given by:

$$\hat{y} = g(f_1(X), f_2(X), ..., f_N(X)) \tag{3}$$

where g represents the meta-classifier that learns from the base models' outputs.

C. Vision Transformer Approach

1) TimeSformer (ViT-Based Model): In contrast to LSTM-based models, we employ a TimeSformer architecture (as shown in Figure.1) to leverage the power of self-attention in both spatial and temporal dimensions. The model processes video frames as sequences of non-overlapping patches, capturing both short-term spatial dependencies and long-range temporal relationships. The pretrained TimeSformer is fine-tuned on our dataset with modifications to its temporal embeddings for shorter video sequences.

a) Pretraining and Fine-Tuning: We use a pretrained TimeSformer from Kinetics-600 and adapt it for deception detection:

- Adjusting temporal embedding layers to accommodate **8-frame** video sequences. (Pre-Trained ViT had 96-Frame sequence)
- Replacing the classification head for binary classification (deceptive vs. truthful).
- Fine-tuning the model to optimize deception-related feature extraction.

D. Transformer approach

In our proposed system, we replace traditional recurrent neural architectures such as LSTMs with a transformer-based model to enhance the learning of temporal dependencies in facial behavior. Transformers, originally introduced in [30], rely entirely on the self-attention mechanism to model relationships between sequential inputs without requiring recurrence. This architecture is particularly beneficial when working with sequences like Facial Action Units (FAUs), as it enables the model to focus on the most salient expressions across time steps regardless of their position in the sequence.

The core of the transformer lies in the self-attention mechanism, which computes a weighted representation of the entire sequence for each position. For a given input sequence $X = [x_1, x_2, ..., x_n]$, self-attention is computed using learned query (Q), key (K), and value (V) matrices as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

Here, d_k is the dimensionality of the key vectors, and the dot product QK^T measures similarity between tokens. The softmax normalizes these scores to compute a weighted sum of values. This mechanism allows the model to capture complex interactions among FAUs over time, such as subtle shifts in facial expressions that may indicate deceptive behavior.

E. Evaluation Metrics

To compare the effectiveness of these four architectures, we evaluate them using:

- **Accuracy**: Measures correct deception/truthful classifications.
- **F1-Score**: Balances precision and recall for imbalanced data.
- **Inference Efficiency**: Assesses real-time applicability in interrogations.

The model is fine-tuned over **15 epochs**, iterating over batches of video data to minimize classification loss using CrossEntropy Loss:

$$L = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}) \tag{5}$$

where:

- N is the number of training samples,
- $C = 2$ (for binary classification),

- y_{ij} is the true label for sample i and class j,
- $\hat{y}_{ij}$ is the predicted probability for class j.

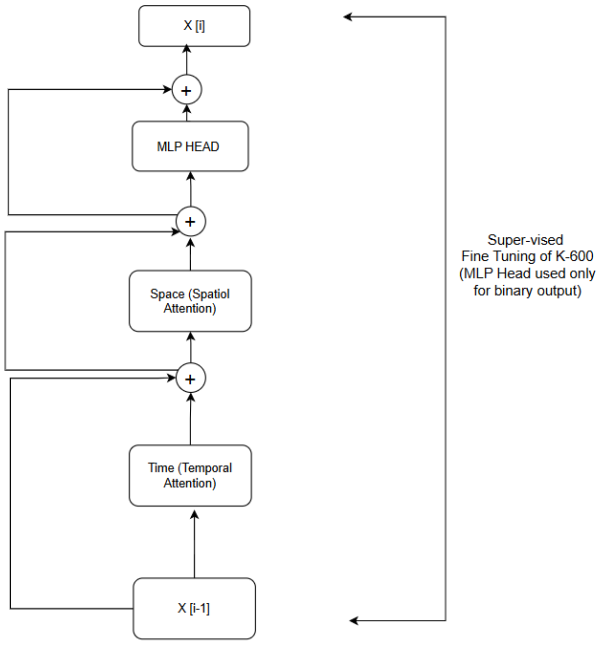

Fig. 1. Divided Space Time Transformer Architecture

IV. Datasets & Pre-Processing

A. Real-Life Court Trials

The **Real-Life Court Trials Dataset** comprises real-world courtroom trial recordings, scraped from publicly available sources. Each video is labeled as **truthful (Not Guilty)** or **deceptive (Guilty)**, representing a **high-stakes, low-quality** environment due to uncontrolled recording conditions.

To ensure dataset quality, we manually preprocessed the videos, retaining only segments where the **defendant** was speaking, as these moments provide the most relevant cues for deception detection. Segments dominated by attorneys, judges, or other courtroom participants were removed. Additionally, we discarded cases with:

- **Poor video resolution**, which could hinder facial microexpression analysis.
- **Excessive background noise**, which could introduce distractions in potential verbal cues.

This manual curation ensured that the dataset retained high-quality deception-related instances while minimizing noise and irrelevant data.

B. Silesian Dataset

The **Silesian Dataset** consists of recorded confessions where students express **truthful or deceptive opinions** about specific objects or images. Unlike the courtroom dataset, this dataset was collected in a **controlled laboratory setting**, using high-resolution cameras, making it a **low-stakes, high-quality** dataset.

To standardize preprocessing, we employed a **MATLAB script** to automatically trim video clips based on time-stamped annotations. These annotations indicated the precise start and end times of **truthful and deceptive** statements. Our script performed the following steps:

- **Extracted only the annotated deception/truth segments**, removing off-topic discussions and silent intervals.
- **Standardized video lengths**, ensuring uniform input dimensions for training.
- **Discarded redundant frames**, optimizing storage and processing efficiency.

Automating this step minimized **m**anual intervention and bias, ensuring consistency across samples while preserving the dataset's integrity for model training.

V. Experiments & Results

A. Setup

1) Libraries & Frameworks Used:

- **PyTorch** - Main deep learning framework used for model building, training, and evaluation.
- **scikit-learn** - Used for splitting data and measuring performance
- **Timm (`timm`)** - A library for pretrained vision models, though it is imported but not actively used.
- **Pandas (`pandas`)** - Used for reading CSV files containing labels.
- **OS (`os`)** - Used for file path operations.
- **TimeSformer** - The main model used for video classification.
- **TensorFlow, Keras** - Used for building & training LSTM models.

2) Harware Used For Training:

- **Processor:** Intel Core i5 (7th Generation)
- **RAM:** 8GB

B. Results

- **Ensemble LSTM Models** We developed Ensemble LSTMs comprising of three and five LSTM models, combined via three different techniques - Bagging, Boosting & Stacking. Table 1 summarizes the Ensemble model comprising of **Three** LSTM models. Table 2 summarizes the Ensemble model comprising of **Five** LSTM models.

Dataset	Bagging	Boosting	Stacking
Silesian Deception Detection	80.3%	60.86%	85.42%
Real Life Court Trails	89.12%	63.32%	90.83%

TABLE I

- **Vision Transformer Model** Table 3 summarizes the accuracies we obtained, while testing our ViT model.
- **Transfomer model** Table 4 summarizes the accuracies we obtained, while testing our Transformer based model, on both the datasets

Dataset	Bagging	Boosting	Stacking
Silesian Deception Detection	89.9%	53.0%	91.9%
Real Life Court Trails	92.11%	50.10%	91.04%

TABLE II

Dataset	Accuracy
Silesian Deception Detection	81.2%
Real Life Court Trails	80.17%

TABLE III

Dataset	Accuracy
Silesian Deception Detection	59.69%
Real Life Court Trails	88.71%

TABLE IV

Figure 2, Figure 3, Figure 4 are the graphs for training & validation loss and accuracies, for Bagging, Boosting & Stacking Ensemble strategies respectively, where the Ensembled model comprises of **Three** LSTM models. Figure 5, Figure 6, Figure 7 are the graphs for training & validation loss and accuracies, for Bagging, Boosting & Stacking Ensemble strategies respectively, where the Ensembled model comprises of **Five** LSTM models. Figure 8 and Figure 9 are the graphs for training & validation loss and accuracies, for our transformer model, on Silesian Deception detection and RLCT respectively.

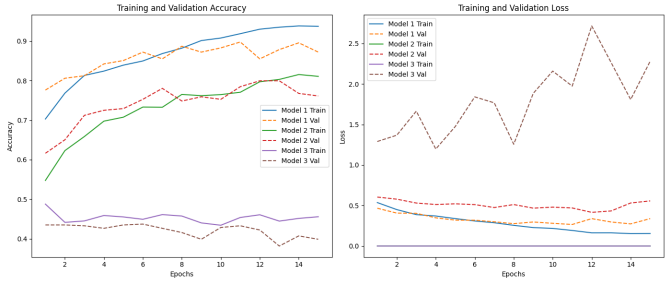

Fig. 2. Training & Validation Accuracy and Loss With Boosting (3 Models)

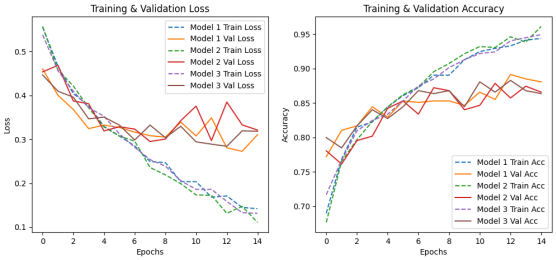

Fig. 3. Training & Validation Accuracy and Loss With Stacking (3 Models)

VI. Discussions & Future Work

A. Analysis & Results

The performance variation observed across different architectures and ensemble strategies can be attributed to their

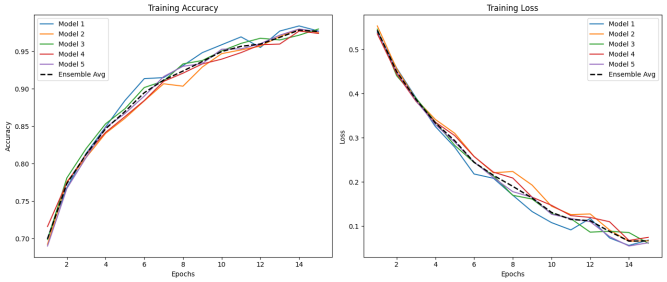

Fig. 4. Training & Validation Accuracy and Loss With Bagging (5 Models)

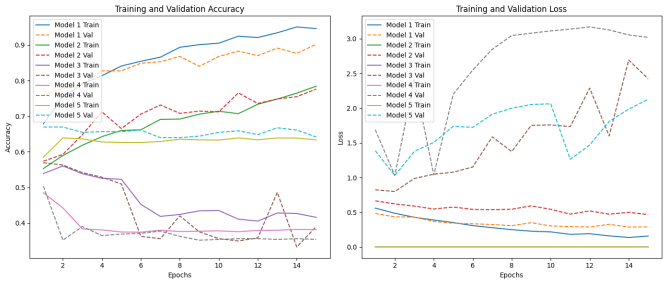

Fig. 5. Training & Validation Accuracy and Loss With Boosting (5 Models)

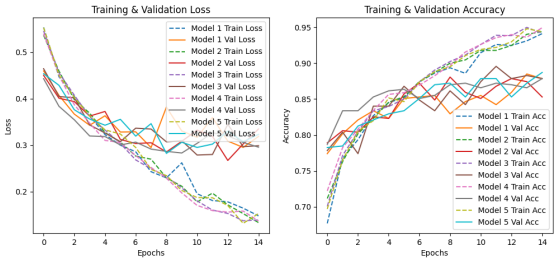

Fig. 6. Training & Validation Accuracy and Loss With Stacking (5 Models)

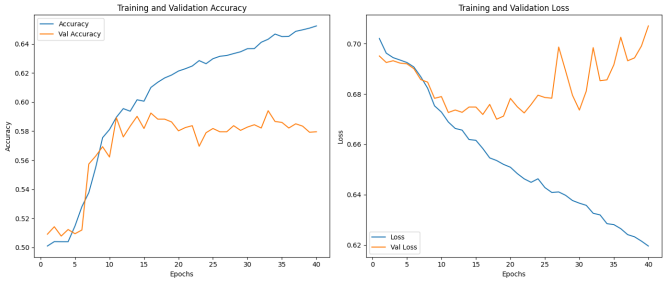

Fig. 7. Training & Validation Accuracy and Loss With Transformer (Silesian)

inherent ability to model temporal dynamics, capture subtle facial patterns, and generalize across varying data conditions.

Bagging-based LSTM ensembles consistently outperformed other configurations, achieving the highest accuracy of **92.11%** on the Real-Life Court Trials dataset and **89.9%** on the Silesian dataset. This can be explained by the model's robustness to overfitting and variance reduction, as Bagging trains multiple LSTMs on different random subsets of the data

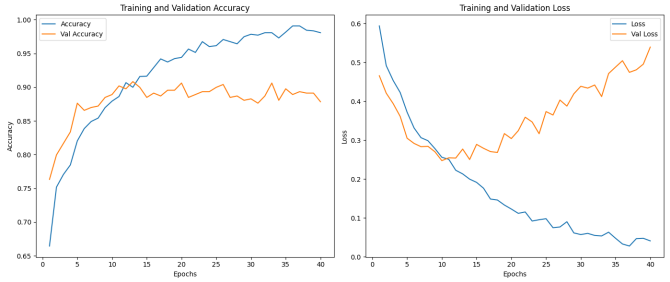

Fig. 8. Training & Validation Accuracy and Loss With Transformer (RLCT)

and aggregates their predictions through majority voting. This ensemble diversity makes Bagging particularly effective for real-world data, where noise and variability are common.

Stacking also showed strong performance, especially on the Silesian dataset, due to its ability to combine complementary strengths of multiple base LSTM models using a meta-learner. By learning how to optimally weigh each model's output, Stacking effectively captures complex deception cues present in structured, controlled environments.

On the other hand, Boosting underperformed in both datasets. This can be attributed to its sequential dependency on correcting previous errors, which can lead to overfitting in noisy or imbalanced datasets, especially evident in the RLCT dataset where Boosting yielded only **50.10%** accuracy with five models.

Transformer-based models demonstrated mixed results. The standard Transformer achieved relatively high accuracy on RLCT (**88.71%**) but struggled with the Silesian dataset (**59.69%**). This suggests that Transformers are better suited to handling real-world data with longer and more complex temporal dependencies but may require more data or better tuning for structured settings. Meanwhile, the Vision Transformer (TimeSformer) showed decent performance across both datasets, leveraging spatio-temporal attention mechanisms. However, its high memory consumption limited the batch size during training, possibly affecting generalization.

In summary, the superior performance of Bagging and Stacking LSTM ensembles is due to their resilience to data noise and their ability to leverage multiple perspectives of the same task. In contrast, Boosting's sensitivity to noisy samples and Transformers' data and compute demands explain their relatively lower performance under certain conditions.

B. Ethical Considerations

In developing our AI model for deception detection using facial expressions, we have carefully considered the ethical challenges involved. Privacy is a major concern, as collecting and analyzing facial data must follow legal guidelines like GDPR, ensuring that individuals give clear consent and their data is handled securely. We also recognize the risk of bias in AI models, which can lead to inaccurate results, especially for people from different cultural or ethnic backgrounds. Since facial expressions alone do not always indicate deception, relying only on them can result in false positives, leading to wrongful judgments. Additionally, constant monitoring of facial expressions can cause stress and make people overly conscious of their behavior. There is also a risk of misuse, such as using this technology for mass surveillance, unfair hiring decisions, or legal cases without proper human oversight. To ensure fairness and responsible use, we encourage future researchers and developers in this field to focus on transparency, accuracy, and ethical considerations when advancing deception detection technology.

C. Findings

To evaluate the effectiveness of different deep learning architectures in deception detection, we conducted extensive experiments across two datasets: the Silesian Deception Detection dataset and the Real-Life Court Trials (RLCT) dataset. Our experiments involved multiple LSTM-based ensemble techniques (Bagging, Boosting, and Stacking), a standard Transformer model, and a Vision Transformer (ViT) based TimeSformer architecture.

Tables I and II present the performance of LSTM ensembles comprising three and five models, respectively. Among the ensemble techniques, Stacking and Bagging consistently outperformed Boosting across both datasets. Notably, Bagging with five LSTM models achieved the highest accuracy overall,**92.11**% on RLCT and **89.9**% on the Silesian dataset. This is to the best of our knowledge, the highest accuracy any other paper has achieved with a mono-modal approach. This also highlights the effectiveness of model variance introduced through bootstrap sampling in capturing subtle deception cues.

Vision transformers (Table III) demonstrated promising results, achieving **81.2**% accuracy on the Silesian dataset and **80. 17**% accuracy on RLCT. This marks a novel approach to this problem which hasn't been explored by any of the previous paper, which includes fine tuning a spatio-temporal ViT. These results suggest that ViTs can generalize well to deception-related facial patterns, although their performance did not surpass the best LSTM ensemble configurations. The standard Transformer model (Table IV) achieved **59.69**% on the Silesian dataset and **88.71**% on RLCT,showing better performance on real-world, noisy data but relatively lower accuracy on controlled data.

In conclusion, while Transformer-based models offer scalability and modeling power for temporal sequences, the Bagging LSTM ensemble remains the most effective strategy in our setup for deception detection using facial action units, particularly in both high-stakes and controlled environments.

D. Future Work

While our implementation of Vision Transformers (ViTs) for deception detection using facial expressions has demonstrated promising results, there remain several challenges and areas for future research. One key limitation encountered during our experiments was the constraint on batch size due to the high memory requirements of ViTs. ViTs process images as sequences of patches, and their self-attention mechanism

scales quadratically with the number of patches. As a result, increasing image resolution or batch size significantly raises memory consumption, limiting training efficiency on standard GPU hardware. This limitation affects both model convergence speed and generalization, as smaller batch sizes may lead to higher variance in gradient updates. Future work could explore strategies such as implementing gradient checkpointing, mixed precision training, and efficient memory management to reduce memory overhead while maintaining performance. Investigating alternative transformer-based models such as Swin Transformer, Linformer, or Performer could optimize self-attention mechanisms to reduce computational complexity. Hybrid architectures that combine convolutional layers with ViTs could leverage CNNs' spatial efficiency while benefiting from the ViT's global attention. Additionally, leveraging multi-GPU or TPU-based training setups could enable larger batch sizes and improve generalization. Applying tensor decomposition or low-rank matrix factorization techniques to approximate self-attention computations may further reduce memory footprint. By addressing these limitations, future research can enhance the scalability and efficiency of ViTs for deception detection, making them more accessible for deployment in real-world applications.

In our implementation of ViTs, we overcame the hurdle of hefty pre-processing on videos, particularly the computation of Facial Action Units (AUs) for each frame. However, despite these optimizations, there is room for improvement in the obtained accuracies. Future research could focus on alternative approaches that enhance accuracy without relying on pre-calculated AUs. Exploring end-to-end learning techniques, where the model directly learns deception-related features from raw video frames, may lead to better performance. The use of spatiotemporal transformers or hybrid deep learning models that incorporate temporal dynamics more effectively could provide valuable insights. Additionally, self-supervised learning techniques and attention-based feature extraction methods might enable ViTs to capture subtle deception cues more effectively. By moving beyond pre-computed AUs and leveraging more advanced architectures, future studies can push the boundaries of deception detection and achieve higher accuracy in real-world applications.

REFERENCES

[1] H. U. D. Ahmed, U. I. Bajwa, F. Zhang, and M. W. Anwar, "Deception Detection in Videos using the Facial Action Coding System," arXiv preprint arXiv:2105.13659, May 2021. (unpublished)

[2] A. V. Ivanov and G. Riccardi, "Automatic Turn Segmentation in Spoken Conversations," in Proceedings of Interspeech 2010, Makuhari, Japan, Sep. 2010, pp. 1858–1861.

[3] F. Soldner, V. Pérez-Rosas, and R. Mihalcea, "Box of Lies: Multimodal Deception Detection in Dialogues," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, Jun. 2019, pp. 21–31.

[4] J. Speth, N. Vance, A. Czajka, K. W. Bowyer, D. Wright, and P. Flynn, "Deception Detection and Remote Physiological Monitoring: A Dataset and Baseline Experimental Results," in Proceedings of the 2021 IEEE International Joint Conference on Biometrics (IJCB), 2021, pp. 1-8.

[5] Cutrow RJ, Parks A, Lucas N, Thomas K. The objective use of multiple physiological indices in the detection of deception. Psychophysiology. 1972 Nov;9(6):578-88.

[6] Thackray RI, Orne MT. A comparison of physiological indices in detection of deception. Psychophysiology. 1968 Jan;4(3):329-39

[7] M. Valstar and M. Pantic, "Fully Automatic Facial Action Unit Detection and Temporal Analysis," 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), New York, NY, USA, 2006, pp. 149-149

[8] J. Zhao and L. Itti, "Decomposing time series with application to temporal segmentation," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 2016, pp. 1-9

[9] W. -C. Lin and C. Busso, "Chunk-Level Speech Emotion Recognition: A General Framework of Sequence-to-One Dynamic Temporal Modeling," in IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1215-1227, 1 April-June 2023

[10] A. Modupe, O. O. Olugbara and S. O. Ojo, "Exploring Support Vector Machines and Random Forests to Detect Advanced Fee Fraud Activities on Internet," 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, 2011, pp. 331-335

[11] More, Sujeet & Kalkundri, Ravi. (2015). Evaluation of deceptive mails using filtering & WEKA. 1-4. 10.1109/ICIIECS.2015.7193262.

[12] Sundari, M.Shanmuga & Nayak, Rudra. (2020). Master Card Anomaly Detection using Random Forest and Support Vector Machine Algorithms. Journal of Critical Reviews. 7. 2020.

[13] P. Li, M. Abouelenien, R. Mihalcea, Z. Ding, Q. Yang and Y. Zhou, "Deception Detection from Linguistic and Physiological Data Streams Using Bimodal Convolutional Neural Networks," 2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPDS), Guangzhou, China, 2024, pp. 263-267

[14] Zhang, X., Zhu, X. (2021). Explaining Neural Network Results by Sensitivity Analysis for Deception Detection. In: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (eds) Neural Information Processing. ICONIP 2021. Communications in Computer and Information Science, vol 1517. Springer, Cham.

[15] Rad, D.; Paraschiv, N.; Kiss, C. Neural Network Applications in Polygraph Scoring—A Scoping Review. Information 2023, 14, 564.

[16] Avola, D. et al. (2023). LieToMe: An LSTM-Based Method for Deception Detection by Hand Movements. In: Foresti, G.L., Fusiello, A., Hancock, E. (eds) Image Analysis and Processing – ICIAP 2023. ICIAP 2023. Lecture Notes in Computer Science, vol 14233

[17] Stathopoulos, A., Han, L., Dunbar, N., Burgoon, J.K., Metaxas, D. (2021). Deception Detection in Videos Using Robust Facial Features. In: Arai, K., Kapoor, S., Bhatia, R. (eds) Proceedings of the Future Technologies Conference (FTC) 2020, Volume 3. FTC 2020. Advances in Intelligent Systems and Computing, vol 1290.

[18] W. Khan, K. Crockett, J. O'Shea, A. Hussain, and B. M. Khan, "Deception in the eyes of deceiver: A computer vision and machine learning based automated deception detection," Expert Systems with Applications, vol. 169, p. 114341, 2021.

[19] N. Bhaskaran, I. Nwogu, M. G. Frank and V. Govindaraju, "Lie to Me: Deceit detection via online behavioral learning," 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), Santa Barbara, CA, USA, 2011, pp. 24-29

[20] Ebrahimi Kahou, Samira, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. "Recurrent neural networks for emotion recognition in video." In Proceedings of the 2015 ACM on international conference on multimodal interaction, pp. 467-474. 2015.

[21] A. Salah, N. Ibrahim and M. Ghantous, "Truth Revealed: Enhancing Deception Detection Using Long-Term Recurrent Convolutional Networks," 2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 2024, pp. 48-52

[22] 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.

[23] Y. Xie, R. Liang, H. Tao, Y. Zhu and L. Zhao, "Convolutional Bidirectional Long Short-Term Memory for Deception Detection With Acoustic Features," in IEEE Access, vol. 6, pp. 76527-76534, 2018

[24] Venkatesh, Sushma & Ramachandra, Raghavendra & Bours, Patrick. (2019). Robust Algorithm for Multimodal Deception Detection. 10.1109/MIPR.2019.00108.

[25] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, "A deep learning approach for multimodal deception detection," in Computa-

tional Linguistics and Intelligent Text Processing: 19th International Conference, CICLing 2018, Hanoi, Vietnam, Mar. 2018, pp. 87–96

[26] Yang, Jun-Teng & Liu, Guei-Ming & Huang, Scott. (2021). Multimodal Deception Detection in Videos via Analyzing Emotional State-based Feature. 10.48550/arXiv.2104.08373.

[27] M. U. Şen, V. Pérez-Rosas, B. Yanikoglu, M. Abouelenien, M. Burzo and R. Mihalcea, "Multimodal Deception Detection Using Real-Life Trial Data," in IEEE Transactions on Affective Computing, vol. 13, no. 1, pp. 306-319, 1 Jan.-March 2022.

[28] Mihai Burzo, Mohamed Abouelenien, Veronica Perez-Rosas, and Rada Mihalcea. 2018. Multimodal deception detection. The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition - Volume 2. Association for Computing Machinery and Morgan & Claypool, 419–453.

[29] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea and M. Burzo, "Detecting Deceptive Behavior via Integration of Discriminative Features From Multiple Modalities," in IEEE Transactions on Information Forensics and Security, vol. 12, no. 5, pp. 1042-1055, May 2017

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.