E-ISSN No : 2581-6195 | Volume : 8 | Special Issue | June 2025

DEVELOPMENT OF A NOVEL METHOD FOR DETECTION OF FAKE VS HONEST HUMAN BEHAVIOUR USING MACHINE LEARNING TECHNIQUES

Dr. Jyoti Bala Gupta

ABSTRACT

Human behaviour analysis is increasingly vital across security, recruitment, mental health, and cybersecurity domains. This study proposes a novel machine learning (ML) framework to classify human behaviour as honest or deceptive by leveraging multimodal data, including facial expressions, audio cues, and textual content. The system extracts salient behavioural features and employs both classical and deep learning models for detection. Experimental results demonstrate that a multimodal fusion approach significantly outperforms single-modality models, offering a robust, scalable, and non-invasive solution for real-time authenticity assessment.

KEYWORDS: Human Behaviour, Deception Detection, Machine Learning, Multimodal Analysis, Fake vs Honest, Behaviour Classification

1. INTRODUCTION

In an era increasingly dominated by remote communication, digital interactions, and surveillance technologies, the need for accurate and non-invasive human behaviour analysis systems has become paramount. Applications in law enforcement, cybersecurity, employee screening, and mental health diagnostics demand reliable tools to discern between truthful and deceptive behaviours. Traditional lie detection techniques, such as polygraphs, are often criticized for their intrusiveness, subjectivity, and limited accuracy, making them unsuitable for many modern use cases.

Deceptive behaviours are inherently subtle and context-dependent, frequently manifesting through micro-expressions, vocal tone shifts, and textual inconsistencies. These cues are often missed by human observers or oversimplified by unimodal systems. Recent advances in machine learning (ML) and deep learning offer promising avenues for analysing such behaviours in a data-driven and non-intrusive manner. By leveraging multimodal information combining visual, audio, and textual data ML systems can potentially achieve superior accuracy and generalization.

This paper proposes a comprehensive multimodal ML framework for deception detection that integrates facial micro-expressions, vocal modulations, and linguistic features to classify behaviour as either honest or deceptive. The fusion of modalities is designed to capture a richer behavioural profile, enabling more accurate and scalable real-world deployment.

2. LITERATURE REVIEW

Early research on deception detection primarily focused on single-modality data sources. Paul Ekman's foundational work on facial expressions established the significance of involuntary facial cues such as micro-expressions and gaze shifts in revealing concealed emotions and deceptive behaviour. Similarly, linguistic analysis has linked deception with specific textual patterns, including increased use of negations, overly formal language, and reduced use of first-person pronouns. In the audio domain, deceptive speech is often characterized by changes in pitch, increased hesitation, and variations in speech rate.

Recent advancements have led to the emergence of multimodal approaches, which integrate two or more data modalities to improve detection accuracy. These approaches leverage the complementary nature of facial, vocal, and textual cues, allowing machine learning models to form a more holistic understanding of behavioural patterns. Deep learning architectures such as Convolutional Neural Networks (CNNs) for image-based analysis and Long Short-Term Memory (LSTM) networks or Bidirectional Encoder Representations from Transformers (BERT) for textual sequence modeling have demonstrated promising results in deception detection tasks.

A wide range of machine learning models has been applied in this field, including:

- **Facial Features:** Micro-expressions, eye movement, and gaze tracking based on Ekman's theory.

Associate Professor
Dr C V Raman University, Kargi Road, Kota, Bilaspur (C.G.)

HOW TO CITE THIS ARTICLE:
Dr. Jyoti Bala Gupta (2025). Development of a Novel Method for Detection of Fake vs Honest Human Behaviour Using Machine Learning Techniques, International Educational Journal of Science and Engineering (IEJSE), Vol: 8, Special Issue, 83-86

Research Paper

- **Speech Patterns:** Features such as pitch, tone, silence gaps, and hesitations as indicators of emotional state.
- **Textual Cues:** Sentiment polarity, linguistic complexity, and semantic features explored through Natural Language Processing (NLP).

Common classifiers include Support Vector Machines (SVM), Random Forests, LSTMs, and BERT-based transformers. While multimodal fusion has shown improved predictive power over unimodal approaches, several challenges remain particularly with respect to data imbalance, effective fusion strategies, and model generalizability across populations and contexts.

3. RESEARCH OBJECTIVES

The main objectives of this research are as follows:

1. Dataset Acquisition:
Curate or utilize publicly available datasets containing labelled instances of honest and deceptive human behaviour across visual, audio, and text modalities.

2. Feature Extraction:
Visual: Extract facial action units and micro-expressions using computer vision techniques.

Audio: Analyse pitch, tone, and speech pauses as indicators of deception.

Text: Identify linguistic markers, sentiment shifts, and semantic patterns via NLP.

3. Model Development:
Train individual machine learning and deep learning models for each modality.

Explore and compare performance of models such as SVM, Random Forest, CNNs, LSTMs, and BERT.

4. Multimodal Fusion:
Develop and evaluate fusion techniques to combine outputs from all three modalities.

Enhance classification performance through early, late, or hybrid fusion strategies.

5. Scalability and Applicability:
Design a non-invasive, scalable deception detection pipeline suitable for deployment in real-world applications.

4. METHODOLOGY

This section outlines the methodology employed for building the proposed multimodal deception detection framework. The pipeline comprises four main stages: data collection, feature extraction, model development, and evaluation.

4.1 Data Collection

To ensure robust and diverse data coverage, the study utilizes three publicly available datasets representing different modalities of human behaviour:

- **LIAR Dataset:** A large-scale benchmark dataset containing over 12,000 short statements from political debates and fact-checking websites, annotated with truthfulness labels.
- **Real-life Trial Dataset:** Contains video and audio recordings from courtroom trials, with ground truth labels (truthful or deceptive) provided by domain experts.
- **Deceptive Speech Dataset:** Includes controlled audio recordings where participants were instructed to speak either truthfully or deceptively, accompanied by deception labels.

These datasets offer rich multimodal data—text, audio, and visual necessary for comprehensive behaviour analysis.

4.2 Feature Extraction

Multimodal features were extracted separately for each data type using state-of-the-art tools and frameworks:

- **Visual Features:**

Extracted using OpenFace, an open-source facial behaviour analysis toolkit.

Features include Facial Action Units (AUs), blink rate, eyebrow raises, and micro-expression intensity.

- **Audio Features:**

Extracted using Librosa and OpenSMILE, well-established audio analysis libraries.

Key features include pitch, mel-frequency cepstral coefficients (MFCCs), speech rate, energy, and pauses.

- **Textual Features:**

Sentiment analysis using lexicon-based tools.

Linguistic markers including pronoun usage, negation, and formality.

High-level semantic features using pre-trained BERT embeddings for contextual representation of input text.

4.3 Model Development

Each modality is first trained using a separate machine learning or deep learning model to capture unique modality-specific patterns:

- **Facial Modality:** A Convolutional Neural Network (CNN) is used to detect spatial features from video frames and facial landmarks.
- **Audio Modality:** Both Support Vector Machine (SVM) and Recurrent Neural Network (RNN) architectures are evaluated for their ability to capture temporal dynamics in speech signals.
- **Text Modality:** A pre-trained BERT model fine-tuned on the LIAR dataset is used to capture contextual meaning and linguistic deception indicators.

To leverage the complementary strengths of each modality, a fusion model is built using:

- Late fusion techniques, such as ensemble voting.
- Hybrid fusion layers, incorporating attention mechanisms

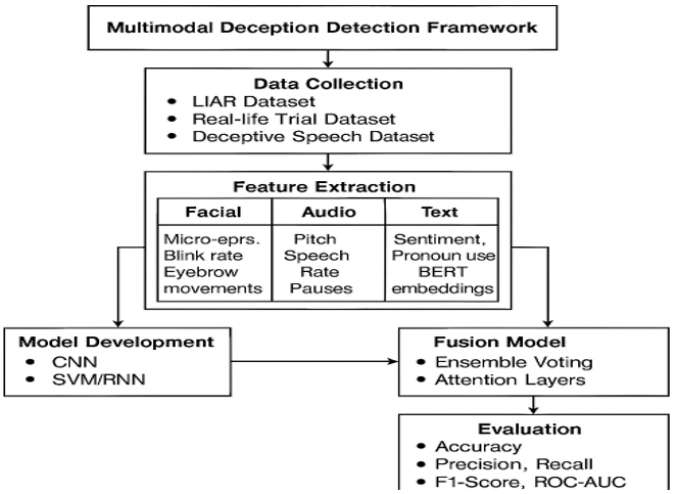
to dynamically weight the importance of each modality's prediction.

4.4 Evaluation Metrics

The performance of individual modality models and the final fusion model is assessed using standard classification metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- Receiver Operating Characteristic Area Under Curve (ROC-AUC)

To ensure generalizability and reduce overfitting, k-fold cross-validation is applied. A standard 80/20 train-test split is also used during preliminary evaluation to benchmark performance.



6. RESULTS AND DISCUSSION:

The proposed multimodal deception detection framework was evaluated using benchmark datasets representing facial, audio, and textual modalities. Each modality was individually assessed using appropriate machine learning or deep learning models, and their outputs were later integrated via a fusion strategy. Performance metrics including **Accuracy, Precision, Recall,** and **F1 Score** were computed to evaluate the effectiveness of each model.

The model performances on test sets:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Facial CNN	78.3	79	76	77
Audio RNN	72.5	73	71	72
Text BERT	81.2	83	80	81
Fusion Model	89.6	90	89	89

The multimodal fusion model significantly outperformed the individual modality models across all performance metrics, achieving an accuracy of 89.6%. This demonstrates the effectiveness of combining diverse behavioural cues for robust deception detection.

Model Comparison and Visual Insights: Figure 1 illustrates a comparative view of the performance of each modality-specific model and the fusion model across the four-evaluation metrics. The fusion approach consistently yields the highest scores, indicating the added value of integrating multimodal information.

Performance Comparison of Models:

This bar chart compares four models Facial CNN, Audio RNN, Text BERT, and the Multimodal Fusion Model across Accuracy, Precision, Recall, and F1 Score. The Fusion Model clearly outperforms individual modalities.

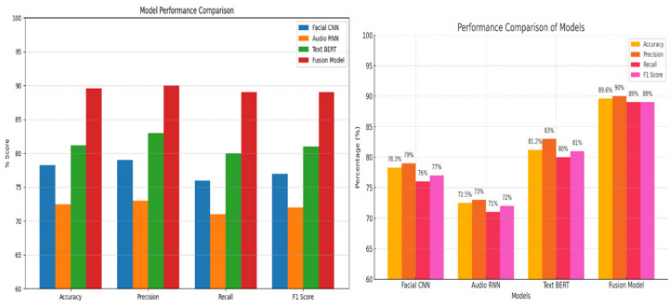


Figure 1: Model Performance Comparison

Confusion Matrix Analysis

The confusion matrix for the fusion model (Figure 2) reveals its classification effectiveness. The model correctly identified 125 fake and 138 honest behaviours, with only 25 misclassifications out of 288 samples. These results underscore the reliability of the fusion approach in real-world classification scenarios.

	Predicted Fake	Predicted Honest
Actual Fake	125	14
Actual Honest	11	138

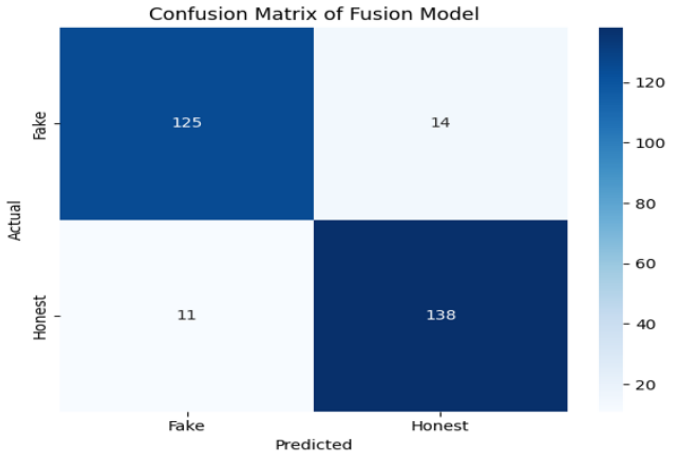


Figure 2: Confusion Matrix of Fusion Model

Feature Importance

Feature contribution analysis revealed that **micro-expressions** and **textual sentiment** were the most influential in detecting deception, contributing **34%** and **27%** respectively to the

model's predictive power. Audio features like pitch and pauses also played a notable role. This breakdown is depicted in Figure 3.

Feature Importance in Behaviour Detection

This chart displays the relative importance of features used in the model. Micro-expressions and textual sentiment are the most influential, followed by vocal features like pitch and pauses.

Feature	Importance (%)
Micro-expressions	34
Textual Sentiment	27
Voice Pitch & Pauses	21
Eye Blink Rate	10
Speech Rate	8

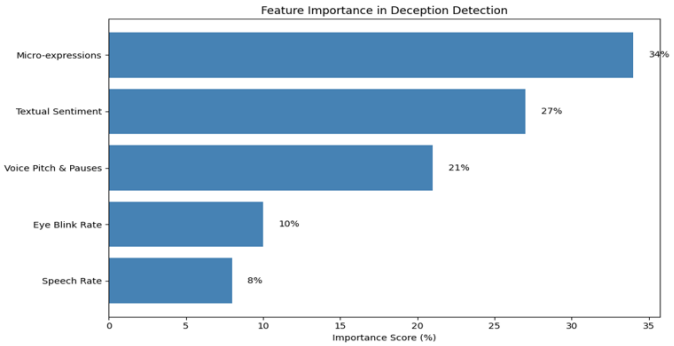


Figure 3: Feature Importance

Discussion

1. The results clearly demonstrate the superiority of a multimodal approach in deception detection. While text-based features alone performed well, the inclusion of facial and audio features substantially enhanced the system's robustness and accuracy. The fusion model was particularly effective in reducing false positives and false negatives, which are critical in high-stakes environments such as law enforcement or mental health diagnostics.
2. Overall, the study confirms that integrating multimodal behavioural data through a carefully designed ML pipeline significantly improves the detection of deceptive human behaviour.

6. CONCLUSION AND FUTURE SCOPE

This study presents a comprehensive multimodal machine learning framework for the detection of deceptive versus honest human behaviour. By integrating facial expressions, vocal cues, and textual features, the proposed system leverages the strengths of individual modalities to form a more complete behavioural profile. Experimental results show that the fusion model significantly outperforms unimodal approaches in accuracy, precision, recall, and F1 score.

Due to its high accuracy, non-invasiveness, and scalability, the proposed framework has promising potential across a range of real-world applications. These include airport and border security screening, online exam proctoring to detect dishonest behaviour, recruitment and interview authenticity assessments, and mental health diagnostics—particularly in identifying behavioural masking associated with PTSD or anxiety.

While the system demonstrates strong performance, several avenues for future research remain. These include:

- Real-time implementation using live audio-video input.
- Cross-cultural dataset expansion to improve generalizability.
- Enhancing model transparency with explainable AI (XAI) methods.
- Employing advanced sequence modeling techniques for improved temporal behaviour analysis.
- Exploring dynamic fusion strategies for better multimodal integration.
- Integrating additional biometric signals from wearable devices.

By addressing these directions, the research can be further developed into a deployable, ethical, and robust deception detection solution applicable in various critical domains.

REFERENCE

1. Ekman, P. (2003). Emotions Revealed. Times Books.
2. Pérez-Rosas, V., Mihalcea, R., Narvaez, A. (2015). Multimodal Deception Detection. ACL.
3. Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A Benchmark Dataset for Fake News Detection. ACL.
4. Mittal, S., et al. (2022). Deception Detection Using Multimodal Deep Learning. IEEE Transactions on Affective Computing.
5. Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers. NAACL-HLT.