University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)

ISSN: 2714-3627

A Journal of the Faculty of Computing, University of Ibadan, Ibadan, Nigeria

Volume 17 No. 1, June 2026

journals.ui.edu.ng/uijslictr
http://uijslictr.org.ng/
uijslictr@gmail.com

University of Ibadan
Journal of Science and Logics in ICT Research

Ensemble Machine Learning Approach to Multimodal Deception Detection

[1]Adetoye Oluwatoyin ADEDOKUN and [2]Adebola K. OJO

[1,2]Faculty of Computing, University of Ibadan, Ibadan, Nigeria.
[1]adedokunadetoye@gmail.com, [2]adebola.ojo@gmail.com

Abstract
Demand for reliable deception detection systems has stimulated the application of machine learning techniques to behavioural analysis. Conventional approaches either rely on single or multiple modalities, including language, facial expression, or gestures. The single modality may not adequately capture the complex nature of deceptive behaviour, while the multimodal approaches also suffer from inaccuracies due the inability of a single algorithms to be able to effectively capture different aspects of the human behavioural indicators. To address this limitation, an ensemble machine learning framework that integrates linguistic and facial cues for multimodal deception detection was developed. The framework combines psycholinguistic features extracted from real life dataset obtained from YouTube, using facial cues derived from the Facial Action Coding System (FACS) and linguistic cues gotten from the Linguistic Inquiry and Word Count (LIWC). An ensemble machine learning approach was employed, using the Random Forest classifier to learn facial deceptive traits, Support Vector Machine (SVM) to learn deceptive linguistic patterns, and fusing the predicted outputs by using the Extreme Gradient Boosting (XGBoost) as the meta-classifier for producing the final classification decision. Experimental evaluation showed that the ensemble multimodal framework outperformed the individual classifiers across multiple performance metrics. The proposed system achieved an overall classification accuracy of 89.8%, an F1-score of 89.8%, and an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.93, indicating strong discriminative capability between deceptive and truthful instances. This study affirms the effectiveness of ensemble strategies for improving deception detection accuracy by using modality-specific classifiers in multimodal feature analysis.
Keywords: *Deception Detection, Multimodal Learning, Ensemble Learning, Facial Action Units*

1. Introduction

Deception has become a prevalent phenomenon and has become a popular area of research in machine learning, artificial intelligence, psychology and human computer interaction. Deception and deceptive behaviours can be found in many domains including border security, court room trials, criminal investigations and job interviews. Traditional deception detection techniques such as polygraphs have been flawed due to a few reasons: it is contact based and intrusive; the measurement of physiological conditions can be unreliable and can be manipulated by subjects [1].

Advancement in machine learning has enabled a significant transformation of manual deception detection into an automatic or automated deception detection using various human behavioural cues such as facial expressions, eye gaze, speech patterns, voice pitch, body movements, physiological signals, and textual semantics [2]. Initial works in automated deception detection carried out a unimodal deception detection system, using a single modality to train their models. However, the drawbacks of such unimodal systems were that they could not effectively detect deception due to its limitation to variety of modality. More recent researches performed a multimodal deception detection, combining multiple modalities such as facial expression, voice pitch, hand gestures, body movement and textual patterns to detect deception.

Furthermore, an effective deception detection system does not only rely on the modality, it also largely depends on the learning model applied. Although there has been a remarkable progress achieved by machine learning and deep learning models such as Random Forest, SVM, Convolutional Neural Networks (CNNs),

Adetoye Oluwatoyin Adedokun and Adebola K. Ojo (2026). Ensemble Machine Learning Approach to Multimodal Deception Detection. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 17 No. 1, pp. 91 - 98

Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), and transformer-based models; Deception detection systems still suffer from a few challenges such as overfitting, noisy data, limited dataset generalization, imbalanced modality, and cultural bias. Individual machine learning classifiers exhibit limitations such as overfitting, sensitivity to data imbalance, and inadequate recall performance [3][4]. Studies have also shown that deceptive cues vary across contexts, languages, demographics, and communication environments, making robust generalization difficult [5].

To address these drawbacks, the ensemble machine learning approach has become a viable option for improving predictive performance and system robustness. The ensemble learning can combine multiple learners to generate a stronger predictive model [4] and in turn reduce variance, minimize bias, and improve classification stability. The ensemble learning model can be used to integrate multiple features and improve the system's ability to capture complex deception patterns that may otherwise be difficult to capture by a single classifier. An ensemble machine learning approach for multimodal deception detection was developed by integrating features obtained from Facial Action Units (AU) and Linguistic features. The aim of this study is to investigate whether ensemble techniques can improve deception detection accuracy, reduce classification errors, and enhance model generalization compared to single classifiers.

2. Related Works

Recent advancement in machine learning has stimulated automated deception detection techniques with the use of machine learning and multimodal analysis. Initial unimodal systems focused on data from single modality such as text, speech or facial expressions. Studies have further shown that deception cues vary across contexts, languages, demographics, and communication environments, making robust generalization difficult [2].

Mihalcea *et al.* [6] showed that deceptive text may contain emotional expressions that were exaggerated and self-reference can be reduced and text may contain irregular syntactic patterns. Pérez-Rosas *et al.* [7] also carried out a speech-based deception detection by exploring the acoustic and linguistic features for deceptive speech analysis. They reported that vocal pitch variations, speech hesitations, pauses, and articulation irregularities could be used as deception indicators. Although unimodal speech systems achieved moderate success, their performance was constrained by environmental noise and speaker variability.

Zhang et al. [8] developed an ensemble learning framework for video-based deception detection by combining facial micro-expression features and speech characteristics. Multiple base classifiers were trained independently, and their outputs were aggregated through a voting mechanism, demonstrating that ensemble models outperformed standalone classifiers in identifying deceptive behaviour from multimodal data. Levitan et al. [9] investigated deception detection in spoken dialogue using linguistic, acoustic, and interactional features. Their study employed multiple classifiers and combined their outputs to improve classification performance. They observed that integrating predictions from different models produced more stable results than relying on a single classifier.

Recent studies such as Baraa *et al.* [10] have further emphasized the relevance of multimodal learning in deception detection systems. They reviewed non-invasive multimodal deception detection techniques and concluded that integrating heterogeneous behavioural features enhances model robustness and reliability.

Several researchers have also demonstrated that multimodal approaches outperform unimodal systems because deceptive cues are often subtle, inconsistent, and distributed across different behavioural channels. Dinges *et al.* [11] emphasized that facial micro-expressions, emotional variations, and nonverbal signals provide valuable indicators for automated deception analysis. Similarly, Bahaa *et al.* [2] reported that combining visual, audio, and textual features significantly improved deception detection performance compared to individual modality models.

The reviewed studies show that deception is a complex human behavior that cannot be fully explained through a single modality. Also, existing multimodal methods often depend on single learning algorithms to learn multiple

behavioural signals, which may limit their ability to effectively model the unique characteristics of each modality thereby reducing overall detection performance. Hence, the need for a more robust approach that can exploit the strengths of different machine learning techniques. This study therefore proposes an ensemble-based multimodal deception detection framework that integrates linguistic and facial cues and allows modality-specific classifiers to learn the distinctive patterns within each feature set before combining their predictions to produce a more accurate and reliable classification.

3. Methodology

This research presents an ensemble machine learning approach that combines linguistic and facial cues for a multimodal detection framework. The system was designed to combine facial action units and linguistic features extracted from real life data set gotten from forensic interviews of criminal perpetrators, suspected individuals and persons of interest in an ongoing investigation.

The framework's objective was to improve deception detection accuracy and performance by segmenting the modalities such that specialized machine learning models can leverage on their strengths to learn patterns from the modalities using an ensemble approach. The system framework was made up of four stages that follows a sequential order: data acquisition and preprocessing, feature extraction, base model training, meta-learning and classification. Figure 1 shows the system framework of the ensemble deception detection system.

3.1 Data Acquisition and Preprocessing

As shown in the system framework diagram in figure 1, the first part of the framework contains the audio-visual data. These data were extracted from real life YouTube videos of criminal perpetrators who were part of a crime or whose loved ones were murdered, missing or took part in a crime or persons of interest in an ongoing investigation. The real-life videos comprised of 80 participants. The choice of dataset was motivated by the need to get high quality data with minimal noise to improve the overall system accuracy instead of simulated dataset where volunteers were asked to act out a lying scenario or intentionally show deceptive traits.

The video data was originally intended for forensic and media use, so there were parts of the video containing advertisements and other irrelevancies to the research, these parts were removed to remain only the needed parts of the video data. The ensemble deception detection system uses two modalities: textual data and facial action unit data. The video data was split into video only and audio only so that the data can be further processed for their respective training models. The video only data were further pre-processed to ensure consistency. Blurry frames and frames with poor visibility were removed to reduce noise and improve extracted facial feature quality. The audio only data were also transcribed into text. The textual data went through some preprocessing procedures including text normalization, tokenization, stop-word removal and punctuation filtering. This was done to maintain consistency, eliminate irrelevancies and reduce noise.

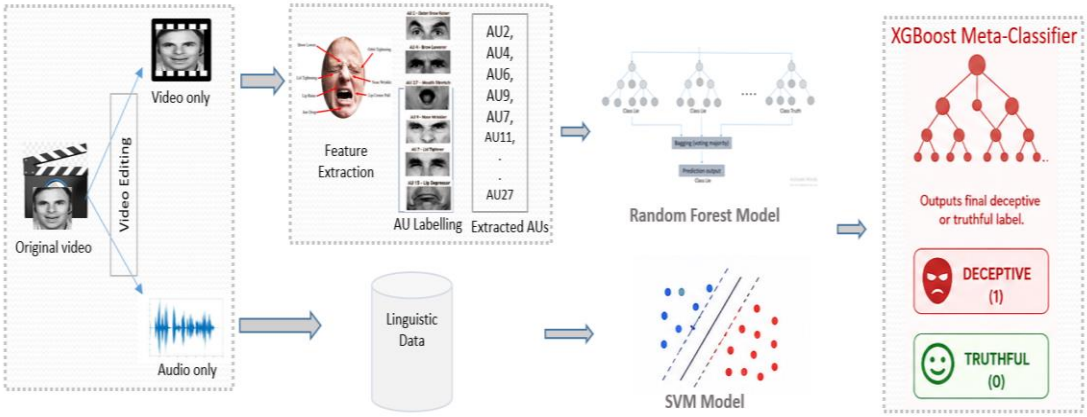

Figure 1 Ensemble Deception Detection System Framework

3.2 Feature Extraction

After splitting the audio-visual data into video and audio data. The image and textual features were extracted separately since deceptive traits differ in these two modalities. The facial features were extracted using the Facial Action Coding System (FACS) while the textual features were extracted using the Linguistic Inquiry and Word Count (LIWC) framework.

3.2.1 Facial Feature Extraction using Facial Action Coding System

Facial Action Coding System is a standardized framework for describing facial muscle movement. These muscle movements are known as Action Units (AU). Each of the AUs correspond to specific facial muscle movement and can be used to measure facial movements that are associated with cognitive load, emotion, discomfort or behaviours associated to deception. The video only data was broken down into video frames. Using the Facial Action Coding system, the image frames went through a process of feature extraction and AU labelling. The video only data was broken down into a frame level AU feature of 3,340 instances which became an input to the facial classification model. Figure 2 shows some sample AU labelling.

3.2.2 Linguistic Feature Extraction using LIWC

Textual features were extracted using the Linguistic Inquiry and Word Count (LIWC) framework. The LIWC is a psycholinguistic analysis tool that categorizes words into psychologically meaningful groups by reading the content of a text file. The text was then analysed by matching the words in the text to the words in its internal dictionaries. The words were categorized into predefined psychological, linguistic and social dimensions for proper analysis of the text. The categories include: cognitive process words, social process, personal pronouns, emotional expressions, negation terms, social process, affective language. These linguistic indicators can be used to detect the psychological and cognitive states of people during communication. The extracted LIWC feature vector becomes an input to the linguistic classification model.

3.3 Base Model Training

After the feature extraction, separate machine learning model were trained for each of the modalities. This concept allows each model to uniquely learn the patterns in each feature space.

3.3.1 Random Forest for Facial Action Unit Classification

Randon Forest is an ensemble model that is robust to overfitting, can handle nonlinear relationships effectively and can handle complex interactions among the AUs. It can also be used to train the features so that they can generalize across different contexts and scenarios. The AUs earlier extracted were trained using the Random Forest classifier. Training the base classifiers allows each model to independently process future unseen features to generate prediction outputs. After the AU features have been classified using the Random Forest classification, there is deception probability output.

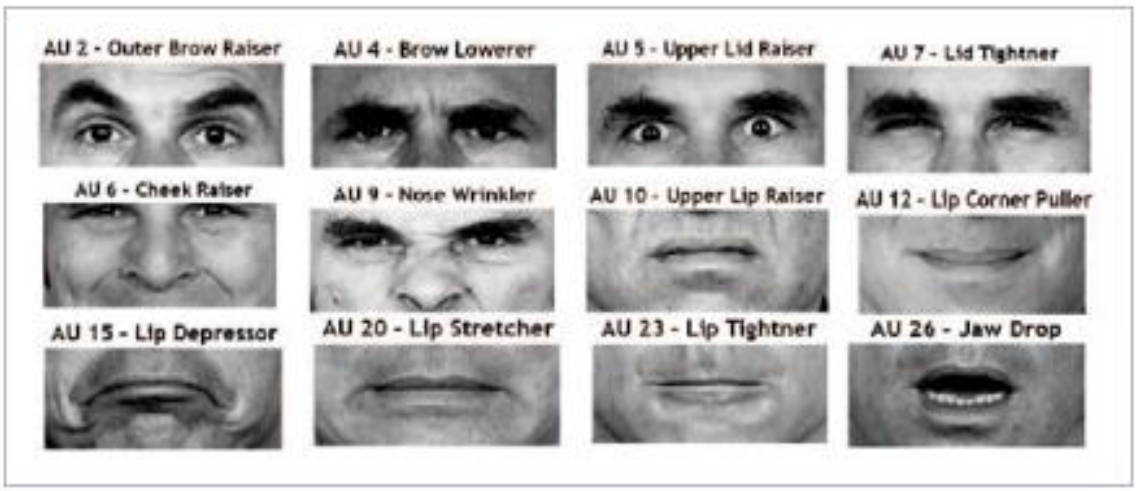

Figure 2: FACS AU labelling

The probability output is generated for the Random Forest facial AU classification and retained for subsequent fusion with the linguistic deception probability output. This allows for a more robust learning process by combining the deception probability output by each base model.

3.3.2 Support Vector Machines for Linguistic Classification

The extracted linguistic feature vectors were trained using the Support Vector Machine (SVM) classifier. The choice of SVM was influenced by its ability to handle high-dimensional textual data and being able to find the optimal decision boundaries between truthful and deceptive classes. The linguistic features and their class labels were fed into the SVM during training. The SVM learns a classification function that is capable of predicting deception from linguistic input. After training, the SVM can independently process future unseen features to generate prediction outputs for linguistic data to produce a deception probability using linguistic features. The deception probabilities from the two base models (i.e. Random Forest and SVM) can then be fed into the meta-classifier.

3.4 Meta-Learning and Classification

The Extreme Gradient Boost (XGBoost) was used as the meta classifier because of it has a strong predictive capability, ability to model complex nonlinear relationships among input variables and robustness to overfitting. After the base classifiers (Random Forest and SVM) have been trained independently using AU and Linguistic data, each model individually generates prediction outputs of deception probability based on their input feature vectors. Each output probability generated from each classifier was concatenated to form a new feature representation. These new concatenated features were fed in as input to the meta-classifier. The Extreme Gradient Boost (XGBoost) was used as the meta classifier because of it has a strong predictive capability, ability to model complex nonlinear relationships among input variables and robustness to overfitting. Using the meta-learning concept enables us to identify situations in which facial cues can provide stronger evidence of detection and vice versa. This can help us leverage on these modality strengths to compensate for the other modality.

The XGBoost receives two inputs: the Random Forest deception probability and the SVM deception probability. These prediction scores were learned by the XGBoost as an optimized decision function to generate a final classification output. The final result was a binary classification output of either “Truth” or “Lie”.

3.5 Performance Evaluation

After the training of the ensemble model, an assessment of its performance was done using the standard performance metrics such as accuracy, precision, recall, F1-score, confusion matrix and Area Under the ROC Curve (AUC). Accuracy was used to determine the overall accuracy performance, while precision and recall were used to investigate the model’s ability to correctly identify deception. F1 score helped us to assess the precision and recall while the AUC showed the ability of the classifier to discriminate across varying decision thresholds. These metrics are defined as shown below:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 Score} = \frac{2\,.\,Recall\,.\,Precision}{Recall + Precision} \quad (4)$$

4. Results

This section discusses in detail the results of the integration of facial and linguistic features in an ensemble learning framework, having classified facial AUs using a Random Forest classifier and linguistic features using an SVM model. The prediction outputs were fused using the XGBoost as the meta-classifier to learn the optimum combination strategy and generate the final deception result. This experiment was carried out with the objective of finding out if an ensemble learning strategy will improve deception detection compared to single machine learning models.

Firstly, an evaluation of Random Forest and SVM was done and their outputs were subsequently integrated using the XGBoost meta-classifier. The performance of the ensemble model was assessed using metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristics (AUC) curve.

4.1 Base Model Performance

The evaluation was carried out in phases, beginning with the assessment of the performance of individual model and modality. The Random Forest classifier was used to train the facial action units that were extracted using the FACS. The SVM was also used to train the linguistic features. Table 1 contains the summary of the performance of the two models. As shown in Table 1, both classifiers had a fairly good accuracy although both classifiers still had a few misclassifications. The AUs classified using Random Forest showed a more superior performance compared to the SVM, indicating that facial cues provide a better indicator of deception compared to linguistic cues.

4.2 Meta-Classifier Performance

After the base models were trained, their prediction outputs were integrated (Random Forest and SVM) using the XGBoost meta-classifier. Table 2 shows the performance of the ensemble model. As shown in Table 2, the ensemble model achieved an overall accuracy of 89.8%, Precision of 89.2%, Recall of 90.5%, F1-Score of 89.8% and AUC of 0.93, clearly outperforming the individual classifiers. When compared to the individual classifiers, the ensemble model outperformed the facial classifier with a difference of 5.7% while it outperformed the linguistic classifier by a difference of 8.1%, suggesting that the ensemble architecture leveraged on the strengths of the two individual classifiers.

Table 3 shows a comparative analysis of the individual classifiers and the ensemble model. The XGBoost meta-classifier successfully learned how the two data sources interact and combined their strengths to act as a meta-classifier.

Table 1 Individual Classifier Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
RF	84.1	83.7	84.8	84.2	0.87
SVM	81.7	80.4	82.6	81.5	0.84

Table 2 Ensemble Model Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Ensemble (SVM + RF + XGBoost)	89.8	89.2	90.5	89.8	0.93

Table 3 Comparative Performance Analysis

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
RF	84.1	83.7	84.8	84.2	0.87
SVM	81.7	80.4	82.6	81.5	0.84
Ensemble (SVM + RF + XGBoost)	89.8	89.2	90.5	89.8	0.93

Figure 3 shows the Receiver Operating Characteristic (ROC) curve of the ensemble meta-classifier, used to evaluate the ensemble model across different classification thresholds. The ensemble model achieved an AUC value of 0.93, confirming a good discrimination between the deceptive and truthful instances and the effectiveness of the multimodal data in ensemble learning.

From the results, the integration of linguistic and facial features for an ensemble learning produced an improved performance over single model approaches and emphasizes the need to further explore the use of multiple models in deception analysis and learning.

4.3 Further Discussions

The primary objective of this study was to determine if an ensemble machine learning approach for multimodal deception detection can improve deception detection performance. Based on the experimental results, this approach provided a substantial improvement over unimodal and single model approaches. Each algorithm was able to leverage on its strength to learn the data more accurately, reducing false classifications. This becomes a complementary relationship between the algorithms, enabling the use of multiple behavioural channels concurrently. This result therefore confirms that the combination of modality-specific classifiers can improve predictive performance beyond what individual models can achieve. Apart from the ensemble strategy, the improvement in performance of this study can be attributed to some other factors.

First, the use of LIWC enabled the extraction of meaningful linguistic cues associated with deception. Secondly, FACS enabled an objective measurement of facial cues associated with deception. Thirdly, XGBoost metaclassifier learned how to combine both modalities while reducing classification errors. The ensemble framework clearly outperformed SVM and Random Forest across all metrics.

One of the findings of this study is that deception is a multidimensional behavioural activity and cannot be solved by isolated cues. The linguistic model captured variations in language use whereas the facial model identified subtle facial muscle movements that occurs during cognitive load during deceptive communication. The ensemble framework exploited the complementary strengths of these models, reducing individual classifier weakness.

Academically, this study contributes to the growing knowledge in deception detection by demonstrating how ensemble learning can be integrated with multimodal behavioural analysis. Apart from academics, the framework can be practically applied in the society in arears such as forensic investigations, border security, law enforcement and intelligence gathering. The system can also be applied in industries, including financial institutions during fraud investigations, online recruitment platforms during remote interviews.

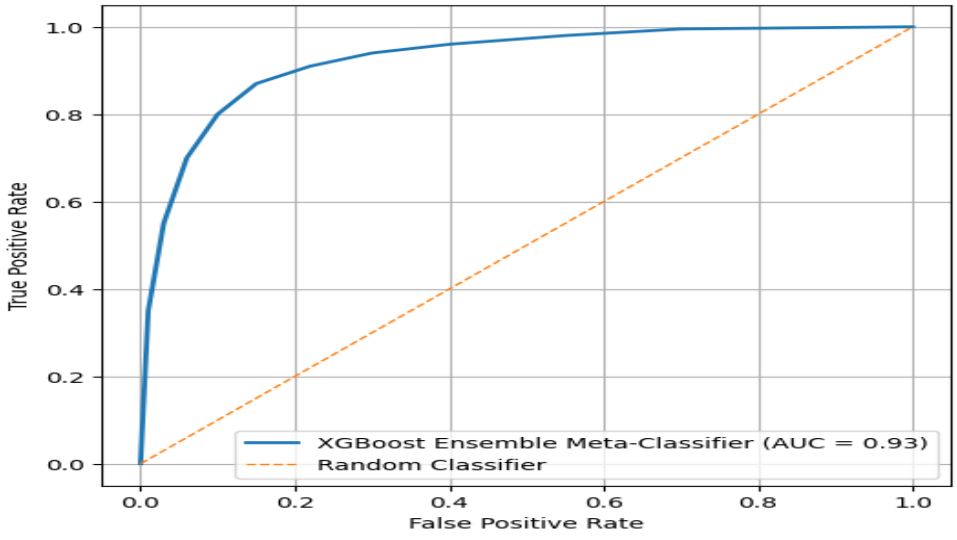

Figure 3: Receiver Operating Characteristic (ROC) curve of the Ensemble classifier

The results obtained in this study is also consistent with previous studies such as L. F. M. and Dinges [9] who reported that facial behavioural analysis using artificial intelligence can provide valuable indicators for deception detection when integrated with other information sources. Bahaa *et. al.* [2] also pointed that multimodal machine learning approaches generally achieve greater robustness than unimodal systems because deception cues are distributed across different behavioural modalities. The findings of this research therefore confirm that the use of ensemble learning for multimodal deception detection can practically improve deception detection performance.

5. Conclusion

An ensemble machine learning framework for multimodal deception detection was developed, combining facial and linguistic cues. The architecture made use of Random Forest classifier for the AU analysis and SVM to model linguistic patterns. The output of these classifiers was integrated by employing a stacking-based ensemble strategy using XGBoost as the meta-classifier. The proposed framework showed a better performance compared to the individual classifiers. This result reveals that using an ensemble strategy with XGBoost as the meta classifier enabled the model to capture characteristics that may be impossible to detect using a single model. Further findings also suggests that multimodal ensemble learning has a promising path for the development of reliable deception detection systems which can be deployed in real life communications.Although this study made use of only two modalities for deception detection, which proved to be effective in the accuracy of deception detection, other behavioural features such as eye movement, body gestures, vocal characteristics and other physiological traits can be explored in the future. The inclusion of these additional modalities may lead to further improvements in deception detection. Future works may also explore the use of advanced deep learning and transformer-based architectures for deception detection.

References

[1] Adedokun, A.O., Ojo, A. K. (2026). Development of an Enhanced Machine Learning Model for Deception Detection Leveraging Facial Action Units and Linguistic Features. *IndabaX Nigeria conference (*pp. 62 - 73*).* PMLR.

[2] Bahaa, M., Hany, M., & Zakaria, E. E. (2024). Advancing automated deception detection: A multimodal approach to feature extraction and analysis. In *International Conference on Intelligent Systems, Blockchain, and Communication Technologies* (pp. 727-738).

[3] Oguntunde, T., & Momoh, M. S. (2023). Development of an Hybrid Machine Learning Model to Detect Email-based Social Engineering Attacks. *Social Informatics Business, Politics, Law, Environmental Sciences & Technology*, vol. 9.

[4] Oyeniran, B. G., Oguntunde, T., & Akinola, S. O. (2020). Development of an Anomaly Detection System for Zero-Day Attacks in Network Traffic Using Ensemble Machine Learning Algorithms, *Computing, Information Systems & Development Informatics Journal.* , vol. 11.

[5] Boumber, D., Verma, R. M., & Qachfar, F. Z. (2024). Blue Sky: Multilingual, multimodal domain independent deception detection. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM) (pp. 396-399).*

[6] Mihalcea, R., & Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP*, 2009.

[7] Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Xiao, Y., Linton, C. J., & Burzo, M. (2015). Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*.

[8] Zhang, Y., Liu, X., Wang, H., & Zhao, Y. (2021). Multimodal deception detection using ensemble learning and feature fusion. In *Expert Systems with Applications*.

[9] Levitan, S. I., Maredia, A., & Hirschberg, J. (2018). "Acoustic-prosodic correlates of deception and trust in interview dialogues. In *Proceedings of Interspeech*, 2018.

[10] Baraa, F. A., & Albaker, B. M. (2024). A review on non-invasive multimodal approaches to detect deception based on machine learning techniques.

[11] Dinges, L., Fiedler, M. A., Al-Hamadi, A., Hempel, T., Abdelrahman, A., Weimann, J., & Bershadskyy, D. (2024). Exploring facial cues: Automated deception detection using artificial intelligence.," *Neural Computing and Applications, 36*(24), 14857-14883.