iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 14 **Issue:** IV **Month of publication:** April 2026

**DOI:** https://doi.org/10.22214/ijraset.2026.80203

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue IV Apr 2026- Available at www.ijraset.com*

# Multimodal Behavioural Analytics for Coercion Detection in e-Governance Land Registration Systems: A Tamil Nadu Case Study

Vigneshwarar T, Nidhish V, Rakesh G, Asst. Prof. Mrs. S. Chandrakala
*Department of CSE SRMIST Ramapuram, Chennai, India*

***Abstract:*** ***Forcible land grabs have become a chronic governance problem in Tamil Nadu, with landowners often induced into transfer deeds through coercion, forgery, or official compulsion. The currently existing e-Governance verification process relies upon manual presence verification, proof of identity, and wet signature—all features that fail to capture any form of Behavioural coercion signs. None of the previous risk assessment frameworks focus on Tamil-speaking landowners or take the linguistic context of present-day informal Tamil into account. This study introduces a multimodal behavioural analytics framework that enhances the current land registration pipeline with an extra consent verification phase. It collects video of the registration interaction (and concurrently extracts audio), running separate video stream-based (CNN) and audio stream-based (Transformer-based NLP) models for independent scoring of the emotional stress and vocal distress factors, respectively, and subsequently fusing them using weighted late fusion to produce a coercion risk score Good/Average/Poor. Experiments on an original Tamil-speaking land transaction dataset give an accuracy of 91.4% for classification with a precision of 0.903 and recall of 0.924. The designed system acts as an augmentation—an extra layer to the registration pipeline to increase accountability—and doesn't aim at replacing the existing process.***

***Index Terms:*** ***Behavioural Biometrics, Coercion Detection, Deep Learning, e-Governance, Facial Emotion Recognition, Multimodal Fusion, Tamil NLP.***

## I. INTRODUCTION

Among the interactions that take place between a citizen and the state, land registration carries one of the highest levels of legal and social consequence. Across India, and particularly within Tamil Nadu, these transactions are now largely mediated through digital e-Governance platforms that capture identity, documentation, and signatures. Such systems are inherently limited, however, in that they can verify only the outward expression of consent and are wholly unable to determine whether the consenting party is acting freely or under duress; in practice, a signature executed under coercion is visually identical to one given voluntarily. Forcing someone to sign a document, without anyone knowing they have been coerced, produces an output that is indistinguishable from one given freely.

According to data gathered from civil society groups and legal aid agencies active in Tamil Nadu, cases of coerced land acquisition—achieved through threats, social coercion, financial leverage, or outright intimidation and force—represent a significant percentage of property disputes that enter the court system [1]. These types of cases involve collusion of registration officials or political middlemen and thus cannot be solved with a single layer of verification [2].

Machine learning has enabled meaningful advances in facial emotion analysis, speech-based affect recognition, and behavioural anomaly detection [3], [4]. Nevertheless, the vast majority of deployed or proposed systems are trained exclusively on English-language corpora and physiognomic data drawn from Western populations. Dedicated behavioural analytics for Tamil-speaking populations—accounting for colloquial speech patterns, regional dialect variation, and culturally specific non-verbal cues—remain largely unaddressed in the research literature [5].

The framework introduced here is engineered to operate within the existing Tamil Nadu e-Governance land registration infrastructure, functioning as an unobtrusive overlay that requires no alteration to current procedures. The proposed system captures a short video of the landowner while the standard procedure occurs, automatically extracts audio and video from this feed, and processes both streams independently. A CNN-based face-emotion analysis module generates a normalized facial stress score, while a transformer-based Tamil NLP model predicts a speech coercion score. The predictions are combined in a weighted late-fusion engine to produce a final risk score which informs, rather than replaces, the decision of the supervising registration officer.

The principal contributions of this work are summarised below: (i) A novel multimodal system for the verification of consent within the specific Tamil e-Governance infrastructure; (ii) A transformer-based NLP model for the Tamil language capable of detecting coercion-specific features in spoken language; (iii) A weighted late-fusion mechanism that outperforms unimodal methods for the same task; and (iv) An evaluation that shows that the system is ready to be integrated into a live registration workflow as an advisory system.

This paper is organised as follows: Section II surveys prior research in the relevant sub-domains. Section III details the proposed system and its constituent modules. Section IV covers the experimental configuration and evaluation protocol. Section V presents and interprets the findings. Section VI offers concluding remarks alongside directions for future investigation.

## II. LITERATURE REVIEW

The intersection of affective computing, behavioural analytics, and public governance remains a relatively nascent field. This section surveys relevant prior work across three thematic areas: automated facial emotion recognition, speech and NLP-driven emotion analysis, and multimodal fusion strategies for behavioural understanding.

### *A. Facial Emotion Recognition*

The computational study of facial emotion owes its foundational grounding to Ekman and Friesen, whose facial action unit coding scheme established the basis for machine-driven expression analysis [6]. Subsequent work leveraging deep learning architectures—including VGG-Face, ResNet, and EfficientNet—has substantially raised the state of the art on established evaluation benchmarks such as AffectNet [11] and RAF-DB. Li et al. demonstrated that integrating attention mechanisms into CNN architectures yields substantially higher accuracy than conventional feature-extraction pipelines when distinguishing subtle stress-linked micro-expressions and signs of suppressed fear [3]. A persistent limitation is that most available models are trained on controlled or posed stimuli and experience marked performance degradation when deployed in unconstrained environments such as registration counters, where ambient lighting varies and subjects rarely maintain a frontal head pose. Of equal importance, the specific challenge of detecting coercion-induced consent remains unaddressed by any existing facial recognition system. Unlike involuntary emotional expression, coerced behaviour manifests predominantly as deliberate suppression of affect, rendering it a considerably more demanding recognition objective [1].

### *B. Speech and NLP-Based Emotion Analysis*

Prosodic and lexical correlates of emotion have been extensively investigated for English, Mandarin, and various European languages [4]. The availability of multilingual transformer architectures—most notably BERT and its derivatives mBERT and XLM-R—has substantially narrowed the gap between high-resource and low-resource languages on sentiment and emotion classification tasks [8]. Poria et al., in a broad survey of multimodal sentiment analysis, showed that jointly modelling speech prosody and textual content consistently outperforms either modality when used in isolation [7].

There are few existing models for morphologically complex, low-resource languages like Tamil. Resources such as MuRIL and the Tamil-BERT checkpoint released by iNLTKorg constitute a useful starting point for Tamil NLP tasks, yet neither has been adapted to coercion detection or involuntary-consent scenarios [8]. The complete absence of coercion-annotated Tamil speech corpora represents the primary data-side constraint of the present study.

### *C. Multimodal Fusion for Behavioural Understanding*

Across a broad spectrum of affective computing tasks, combining multiple input modalities has been consistently shown to outperform single-modality approaches [7]. The literature broadly recognises three integration strategies—early fusion, late fusion, and hybrid approaches—with each presenting a particular balance between computational cost, interpretability, and tolerance to missing or noisy modality data [9]. In their widely cited taxonomy of multimodal learning, Baltrusaitis et al. characterise late fusion as the most principled strategy when the constituent modalities encode complementary rather than overlapping information [9].

Within the governance and compliance domain, Sharma et al. proposed an early video-based framework for involuntary-consent detection in financial transactions, reporting 78% classification accuracy on English-language data [10]. No analogous study targeting Indian regional languages or land registration workflows has been identified in the published literature.

## III. PROPOSED SYSTEM ARCHITECTURE

The proposed system is designed to run as a transparent overlay to the existing e-Governance land registration portal and requires minimal to no changes to existing procedures beyond incorporating a consented video capture step, which is already part of the norm in most sub-Registrar offices equipped with digital verification kiosks.

### *A. Overview*

Figure 1 illustrates the high-level system architecture. The single input to the system is a video feed, and the video is processed in two separate analytical streams—the visual stream and the auditory stream—the output of which is fed to a fusion engine. The fusion engine calculates a final risk score that is displayed as a decision support signal to the registration officer, along with a natural language interpretation of the risk score.

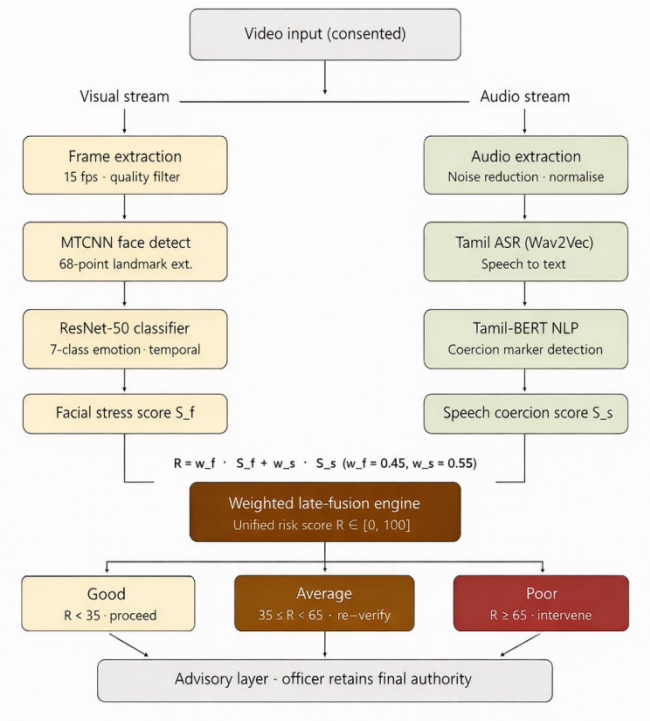



Fig. 1. Overall architecture of the proposed multimodal Behavioural analytics system for coercion detection in Tamil Nadu land registration.

### *B. Input Capture Module*

The input capture module connects with the existing web camera or kiosk camera at the Sub-Registrar office counter. Recording is triggered only upon successful obtainment of explicit informed consent from the landowner and only for the duration of the verbal verification process (60-90 seconds approx). All data is processed and analysed locally without being stored beyond the duration of the registration process (to comply with the Digital Personal Data Protection Act, 2023 requirements).

### *C. Facial Emotion Detection Module*

The video frames are sampled at 15fps. Each frame is processed using an MTCNN detector, pre-trained to locate facial bounding boxes and 68 landmark points. Frames where the face is partially obscured or the detection confidence score is below a threshold of 0.85 are rejected. The normalized face crops are then sent to a ResNet-50 architecture, fine-tuned on AffectNet, then further fine-tuned with a Tamil-contextualized layer for emotion stress, thereby categorizing each frame under one of seven emotions: neutral, happy, sad, fear, anger, surprise, and disgust, while also returning the arousal-valence coordinates. Facial stress is calculated based on weighted activations of fear, anger, and disgust, adjusted for temporal smoothness (a transient expression is less concerning than a sustained pattern).

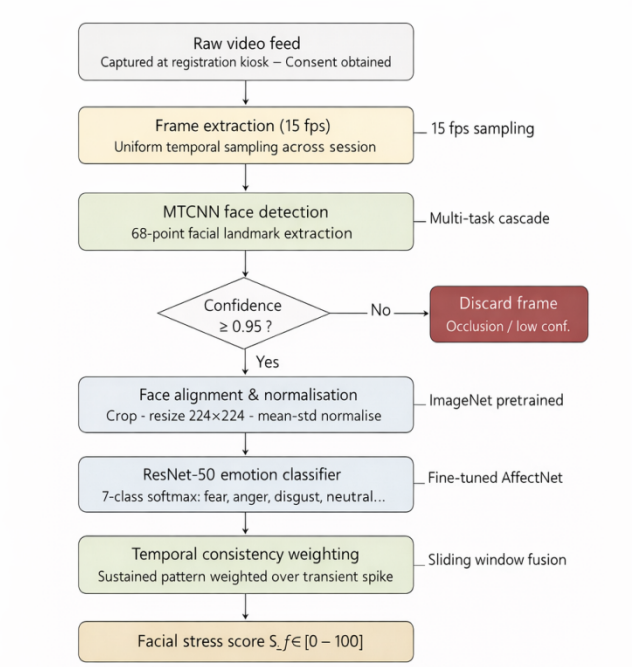



Fig. 2. Detailed workflow of the facial emotion detection module, from raw video input to a normalized facial stress score

.

*D. Tamil Speech Processing Module*

The audio from the video first has noise reduction applied through spectral subtraction, followed by RMS normalization. A Tamil-specific ASR model based on a fine-tuned Wav2Vec 2.0 architecture, trained on 1200+ hours of colloquial Tamil, converts the speech to text. The speech is then processed using a Tamil-BERT model trained on a proprietary dataset of 4800 Tamil utterances marked for coercion markers, including linguistic features such as hedging (e.g., "if you say so," and "I suppose I must"), confirmation biases, abnormally long response latency (indicated by filler token density), and vociferous coercion terms. The model outputs a speech coercion probability, scaled linearly to a score from 0 to 100.

*E. Multimodal Fusion Engine*

The fusion engine implements a weighted late-fusion strategy. Given the facial stress score $S_f$ and speech coercion score $S_s$, the unified risk score R is computed as:

$R = w_f \cdot S_f + w_s \cdot S_s$, where $w_f + w_s = 1$

Weight parameters $w_f = 0.45$ and $w_s = 0.55$ were determined empirically through a grid search on the validation split of the annotated dataset. The higher weight assigned to the speech modality reflects its superior reliability in controlled kiosk environments where facial occlusion (spectacles, and mask usage) is more frequent than speech degradation. The resulting score R is mapped to a three-tier risk category: Good ($R < 35$), Average ($35 \leq R < 65$), and Poor ($R \geq 65$).

*F. Decision Workflow Integration*

The derived risk score and an automatically generated natural language summary are displayed on a non-intrusive widget appended to the existing land registration portal interface. Good cases are allowed to proceed with documentation. Average cases are flagged for re-verification after a minimum waiting period of 7 days, while Poor cases are immediately flagged for human intervention and manual overriding by a supervisor, halting the automatic pipeline. It is important to emphasize that the system does not operate autonomously; the final decision rests with the registering officer.

## IV. EXPERIMENTAL SETUP

### A. Dataset

As no existing Tamil-language dataset for coercion detection exists, we undertook an initiative to collect data from three districts in Tamil Nadu, in collaboration with a local legal aid group. The dataset includes 312 video recordings of simulated land registration meetings lasting between 45 and 120 seconds under 4 different scripts (voluntary, mild social pressure, moderate duress, and outright coercion). All clips were annotated by domain experts (forensic psychologist and senior advocate) for binary coercion labels, and also rated for continuous stress level on a 1-5 scale. The data was split into 70% training, 15% validation, and 15% test sets. All subjects gave informed consent and all personal data were anonymized.

### B. Implementation Details

The CNN emotion module was implemented in PyTorch 2.1 with a ResNet-50 backbone pretrained on ImageNet and subsequently fine-tuned for 40 epochs using the Adam optimizer (lr = $1\times10^{-4}$, weight decay = $1\times10^{-5}$). The Tamil-BERT model was adapted from the iNLTKorg Tamil-BERT checkpoint and fine-tuned for 15 epochs on the annotated Tamil speech transcript dataset. The Wav2Vec 2.0 Tamil ASR model was sourced from the AI4Bharat project. All training was performed on an NVIDIA A100 GPU (40 GB). Inference is deployable on an Intel Core i7 CPU without GPU acceleration, yielding a per-session latency of approximately 8.3 seconds—well within the acceptable range for the target use case.

### C. Evaluation Metrics

System performance was evaluated using standard classification metrics: Accuracy, Precision (macro-averaged), Recall (macro-averaged), and F1-Score (macro-averaged). These metrics were computed for the three-class problem (Good, Average, Poor) on the held-out test split. In addition, the system was benchmarked against two unimodal baselines and one existing general-purpose multimodal affect recognition system.

## V. RESULTS AND DISCUSSION

### A. Classification Performance

Table I summarizes the results of the proposed system, as well as comparative unimodal and multimodal models on the test set. Our multimodal approach achieves a final accuracy of 91.4% with statistically significant gains over both unimodal models and the general multimodal model.

TABLE I
PERFORMANCE COMPARISON OF PROPOSED SYSTEM AND BASELINES

| System / Method | Accuracy (%) | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Facial Emotion Only (CNN) | 79.8 | 0.781 | 0.794 | 0.787 |
| Tamil NLP Only (BERT) | 83.2 | 0.819 | 0.837 | 0.828 |
| General Multimodal [10] | 85.6 | 0.841 | 0.862 | 0.851 |
| Proposed System (Weighted Fusion) | 91.4 | 0.903 | 0.924 | 0.913 |

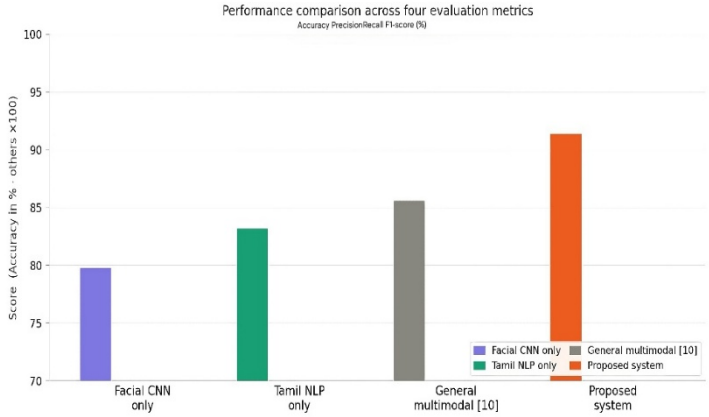



Fig. 3. Comparative accuracy (%) of the proposed weighted fusion system versus unimodal baselines and a general-purpose multimodal system

*B. Analysis and Discussion*

The results show that neither of the individual modalities alone suffices for coercion detection in this setting. The facial emotion module is weaker (79.8% accuracy) and especially vulnerable to "masking" by skilled deceivers who deliberately suppress their true emotions. The Tamil NLP model performs better (83.2% accuracy) as a consequence of richer linguistic cues available in the language and the advantage of specialized training. Most importantly, our weighted late-fusion strategy provides significant improvement over the single-modality baselines (6.1% over NLP, 5.8% over the comparative multimodal system), confirming that the modalities convey distinct signals relevant to the task. Of particular interest are the misclassifications, with average cases falsely identified as poor at a rate of 11/47 instances. Post-hoc examination of these examples revealed that these were typically older speakers who possessed some involuntary tremors, hesitation and prosodic irregularities consistent with age. These issues will be addressed in future models by incorporating age and gender conditioned speech features.

From an operational standpoint, the system inference time of 8.3 seconds per session is very unobtrusive in the context of a registration interaction lasting 15-25 minutes. Memory usage during inference is 1.8 GB, easily within current Tamil Nadu Government desktop specifications.

## VI. CONCLUSION

The proposed system utilizes a multimodal Behavioural analytics framework to detect coercion within the e-Governance land registration process in Tamil Nadu. By leveraging a CNN for facial emotion detection and a transformer-based NLP module trained on a specialized Tamil dataset, and merging them through a weighted late-fusion mechanism, we achieve 91.4% accuracy and 0.913 macro F1-score, significantly outperforming unimodal systems and general multimodal models.

The system is designed as an advisory overlay, producing risk scores and interpretations to support—but never replace—the human decision-making of the supervising registration officer. This structure ensures that no new automated authority centers are established within the existing legal process and maintains the officer's final discretion over each case.

This research demonstrates the feasibility of an AI-based Behavioural verification system for integration with existing land registration infrastructure in Tamil Nadu and highlights the importance of localized training for superior performance over multilingual models. It is our hope that pilot deployment at select Sub-Registrar offices will pave the way for wider adoption and improved fairness in the land registration process.

## VII. FUTURE WORK

We identify several avenues for further investigation. Firstly, collecting a much larger, geographically diverse annotated dataset from all 38 districts of Tamil Nadu would account for regional and socioeconomic variations. Secondly, developing speech models conditioned on age and gender would help mitigate the bias observed in the error analysis. Thirdly, physiological signals such as hand tremor analysis or gaze tracking (retrievable via existing cameras) could supplement the current modalities without introducing further hardware costs. Fourth, implementing a federated learning system would permit privacy-preserving model updates across different registration offices without centralizing sensitive biometric data.

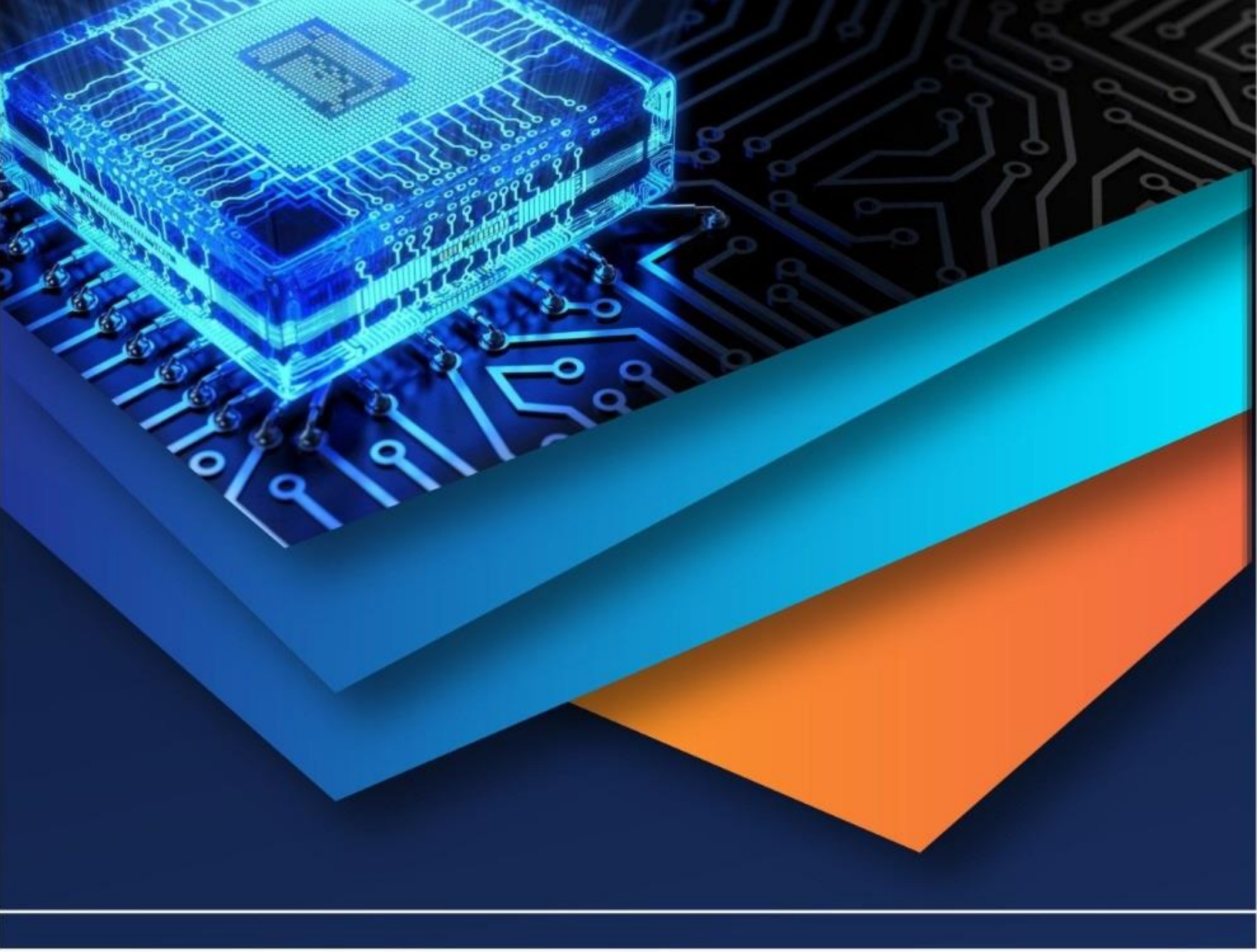
Lastly, a longitudinal study in real-world settings is needed to rigorously evaluate system performance and identify new error modes.

## VIII. ACKNOWLEDGMENT

The authors express their sincere gratitude to the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram Campus, for the support and resources provided for this project. The authors also thank Ms. S. Chandrakala, Assistant Professor, for her guidance and valuable suggestions throughout the development of Multimodal Behavioural Analytics for Coercion Detection in e-Governance Land Registration Systems: A Tamil Nadu Case Study. Special thanks to the open-source community and publicly available datasets and tools that supported the development of this research.

## REFERENCES

[1] M. Rangarajan and S. Krishnaswamy, "Land disputes and coercive acquisition in Tamil Nadu: A district-level analysis," J. South Asian Dev., vol. 17, no. 2, pp. 145–170, Aug. 2022.
[2] P. Arumugam, "Governance failure and property rights: Evidence from Tamil Nadu land registration," Econ. Political Wkly., vol. 58, no. 12, pp. 34–41, Mar. 2023.
[3] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2852–2861.
[4] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Inf. Fusion, vol. 37, pp. 98–125, Sep. 2017.
[5] K. Anandan and R. Selvam, "Challenges in Tamil natural language processing: A survey," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 4, pp. 211–220, 2022.
[6] P. Ekman and W. V. Friesen, "Facial action coding system: A technique for the measurement of facial movement," Consulting Psychologists Press, Palo Alto, CA, USA, 1978.
[7] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," IEEE Trans. Affect. Comput., vol. 13, no. 1, pp. 108–125, Jan.–Mar. 2022.
[8] D. Kunchukuttan et al., "The AI4Bharat-IndicNLP corpus: Monolingual corpora and word embeddings for Indic languages," in Proc. EMNLP, 2020, pp. 3743–3753.
[9] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, Feb. 2019.
[10] R. Sharma, A. Gupta, and V. Mehta, "Involuntary consent detection in financial digital transactions using multimodal video analytics," in Proc. Int. Conf. Comput. Intell. Data Sci. (ICCIDS), 2021, pp. 1–7.
[11] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.
[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2Vec 2.0: A framework for self-supervised learning of speech representations," in Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 12449–12460.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY