International Conference on Intelligent Computing and Explainable AI (ICICEA) – April 2026

RESEARCH ARTICLE OPEN ACCESS

Smart Recruitment Interview System Using Emotion-Aware AI

Dr. S. Padmapriya*, K. Swathilakshmi**, N. Dhara**, B. Dharshni**, V. Jayakiruba**

*Professor -Computer Science and Engineering, A.V. C. College of Engineering, Mannampandal, Mayiladuthurai,

**UG Students-Computer Science and Engineering, A.V. C. College of Engineering, Mannampandal, Mayiladuthurai, India

ABSTRACT

Recruitment processes traditionally rely on human judgment, which may introduce subjectivity, bias, and inconsistency in candidate evaluation. While technical skills and qualifications are considered during interviews, critical behavioral attributes such as emotional intelligence, stress management ability, confidence level, and communication clarity are rarely measured objectively. To address these limitations, this paper proposes a Smart Recruitment Interview System using Emotion-Aware Artificial Intelligence that automatically analyzes candidate emotions during the interview process. The proposed system captures real-time video, audio, and textual responses from candidates through a web-based interview platform. Facial expressions are analyzed using a deep learning model based on MobileNetV2, while speech signals are processed using MFCC feature extraction and a Deep Neural Network (DNN – Fully Connected Neural Network) to identify voice-based emotional patterns. Candidate responses are also evaluated using natural language processing techniques to assess response clarity and relevance. A multimodal fusion mechanism integrates the outputs from facial, voice, and textual analysis modules to generate an overall emotional intelligence score and behavioral evaluation. The results are presented through an HR dashboard that provides structured insights and hiring recommendations. The proposed system aims to improve fairness, reduce interviewer bias, and enable data-driven recruitment decisions.

Keywords —Emotion Recognition, Artificial Intelligence, Smart Recruitment System, Deep Learning, Multimodal Fusion, MobileNetV2, Voice Emotion Analysis.

I. INTRODUCTION

Recruitment is a critical process in organizational development, as selecting suitable candidates directly influences productivity, innovation, and long-term success. Traditional recruitment systems rely primarily on manual interviews conducted by human evaluators. While such methods allow recruiters to assess candidate skills and personality traits, they are often affected by subjective judgments, interviewer bias, and inconsistent evaluation criteria.

In many cases, interviewers may unintentionally evaluate candidates based on first impressions, body language interpretations, or personal preferences rather than objective analysis. Additionally, conventional recruitment methods typically focus on technical qualifications and communication skills while overlooking essential behavioral attributes such as emotional intelligence, stress handling capability, and confidence level during interviews. Recent advances in artificial intelligence and machine learning have enabled the development of intelligent systems capable of analyzing human behavior through facial expressions, speech signals, and textual content. Emotion recognition technologies can provide valuable insights into candidate behavior by identifying emotional patterns during interactions. Integrating these technologies into recruitment systems can improve the accuracy and fairness of candidate evaluation.

This paper proposes a Smart Recruitment Interview System using Emotion-Aware AI, which performs automated emotion analysis during candidate interviews. The system captures video, audio, and textual responses from candidates and processes them using deep learning models. Facial expressions are analyzed using a MobileNetV2-based convolutional neural network, while voice signals are analyzed using MFCC-based Deep Neural Network (DNN) model. Textual responses are evaluated using natural language processing techniques. The outputs of these modules are integrated using a multimodal fusion mechanism to generate a comprehensive emotional intelligence score.

II. LITERATUREREVIEW

This section reviews key research on emotion recognition, affective computing, and artificial intelligence–based recruitment systems, with emphasis on facial emotion recognition, speech emotion recognition, and multimodal behavioural analysis. Existing studies demonstrate the growing potential of deep learning techniques for analysing human emotional responses in digital environments. However, several challenges remain related to multimodal integration, computational complexity, and real-time deployment in intelligent recruitment platforms. Table I summarizes representative research works in this domain, highlighting their methodologies, application contexts, and identified research limitations.

Table I presents recent studies focusing on facial emotion recognition, speech emotion analysis, text-based emotion detection, and AI-assisted recruitment frameworks. These approaches demonstrate advancements in emotion recognition technologies using deep learning models, domain adaptation techniques, and multimodal behavioural analysis. Despite these developments, many existing systems rely on singlemodal emotion detection and lack comprehensive integration of multiple emotional cues for automated interview evaluation.

TABLE I

COMPARATIVEANALYSISOFEMOTIONRECOGNITIONANDAIBASED RECRUITMENTSYSTEMS

Author & Year	**Technique / Model**	**Application Context**	**Technical Contribution & Identified Gap**
A. N. Salman et al. (2025)	Mixture of Emotion Dependent Experts	Video-based Facial Emotion Recognition	Improves temporal facial emotion recognition in videos; limited to facial modality with high complexity.
A. Amjad et al. (2025)	Multi-Domain Adaptation	Speech Emotion Recognition	Enhances crossdataset speech emotion recognition; focuses only on speech signals.
A. Sharma et al. (2025)	Hybrid Deep Learning Facial Model	HR Interview Emotion Analysis	Enables real-time facial emotion prediction for HR systems; lacks multimodal emotion analysis.
S. J. Colaco & D. S. Han (2025)	Facial Landmark Attention Network	Context-Aware Facial Emotion Recognition	Improves facial emotion classification using landmark attention; requires heavy preprocessing.
O. Araque & C. A. Iglesias	Emotion Signal Extraction with Semantic Similarity	Text-Based Emotion Detection	Detects emotional signals from text using semantic similarity; limited to textual modality.
E. Dikbiyik et al. (BiMER)	Bimodal Emotion Recognition	Speech and Text Emotion Analysis	Combines speech and text emotion recognition; increased model complexity.
Zhao Huijuan et al.	Deep Local Domain Adaptation	Cross-Corpus Speech Emotion Recognition	Improves crosscorpus speech emotion recognition; limited to speech modality.
Nishad Mathur et al. (SAGA)	AI Recruitment Analysis System	AI-Based Interview Evaluation	Provides AI-based audio-video interview evaluation; limited focus on multimodal emotion fusion.

Recent advancements in emotion recognition and AIbased recruitment systems have shown promising results in analysing candidate behaviour using facial expressions, speech signals, and textual responses. Various deep learning and machine learning models have been proposed to improve emotion detection accuracy in individual modalities such as facial, speech, and text data. However, most existing approaches focus on single-modality analysis, which limits their ability to capture comprehensive emotional cues during interviews. Recent studies attempt to integrate multiple modalities such as audio, video, and text to improve emotion recognition performance. Techniques including deep neural networks, attention mechanisms, and domain adaptation methods enhance feature extraction and cross-dataset generalization. Despite these improvements, many systems still face challenges related to computational complexity, real-time processing, and effective multimodal fusion.

Overall, the literature reveals several research gaps, including limited integration of multimodal emotion signals, insufficient real-time interview analysis, and lack of scalable AI-driven recruitment frameworks. These limitations motivate the development of an intelligent multimodal interview analysis system that combines facial, speech, and textual emotion recognition to improve candidate evaluation accuracy and decision-making in recruitment processes.

III.RESEARCHPROBLEMANDOBJECTIVES

Modern recruitment processes increasingly rely on intelligent systems to evaluate candidate performance during interviews. AI-based interview analysis systems aim to automatically assess candidate emotions, behaviour, and responses using data from multiple sources such as facial expressions, speech signals, and textual responses. Although various machine learning and deep learning models have been developed for emotion recognition and candidate evaluation, their practical adoption in real-world recruitment environments remains limited due to challenges related to multimodal data integration, computational efficiency, and real-time analysis.

A. Research Problem Statement

The primary research problem addressed in this work is the lack of an efficient and reliable multimodal emotion recognition framework for automated interview analysis. Most existing systems focus on a single modality—such as facial expression analysis, speech emotion recognition, or text sentiment analysis—leading to incomplete understanding of candidate emotions and behaviour. Furthermore, current approaches face several limitations, including: (i) difficulty in effectively integrating facial, speech, and textual emotion signals, (ii) high computational complexity affecting real-time interview processing, (iii) reduced accuracy due to variations in lighting conditions, speech quality, and candidate behaviour, (iv) limited scalability for large-scale recruitment platforms, and (v) lack of intelligent evaluation mechanisms for objective candidate assessment. These challenges highlight the need for a robust multimodal AI-based interview evaluation system that can provide accurate and automated candidate analysis.

B. Research Objectives

To address these challenges, this paper aims to: develop a multimodal emotion recognition framework integrating facial, speech, and textual data for interview analysis, implement deep learning techniques to improve emotion detection accuracy across multiple modalities, design an automated candidate evaluation system capable of analysing emotional and behavioural cues during interviews, improve system efficiency for real-time interview processing, and enhance decision-making in recruitment by providing objective AI-based candidate insights.

C. Contributions of the Paper

The main contributions of this work are summarized as follows:

- Development of an AI-based multimodal interview analysis system integrating facial, speech, and text emotion recognition.
- Implementation of deep learning models for accurate candidate emotion detection and behavioral analysis.
- A unified framework for automated candidate evaluation to support intelligent recruitment decisions.
- An evaluation approach demonstrating the effectiveness of the proposed system for AI-driven interview assessment.

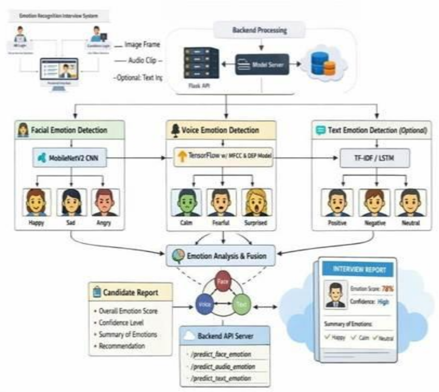

Fig. 1. Multimodal emotion recognition architecture for AI-based interview analysis integrating facial, speech, and textual emotion detection modules with backend processing and emotion fusion.

As shown in Fig. 1, the proposed AI-based interview emotion recognition system integrates facial, speech, and textual emotion analysis to evaluate candidate behaviour during interviews. The system combines multiple machine learning models with backend processing to automatically detect emotional states and generate a comprehensive interview assessment report. By analysing multiple modalities simultaneously, the proposed framework improves the reliability and accuracy of emotion detection compared to single-modality systems. The architecture consists of a frontend interview interface, backend processing server, and multimodal emotion detection modules. The frontend allows HR personnel and candidates to participate in the interview through a web-based platform. During the interview session, the system captures video frames, audio signals, and optional textual responses, which are transmitted to the backend server for processing and analysis.

A. Design Principles

The proposed system is designed based on the following principles:

Multimodal emotion analysis, where facial, speech, and textual data are analyzed together to improve emotion recognition accuracy. Automated interview evaluation, enabling objective candidate assessment using AI techniques.

Real-time processing, ensuring that emotion detection and analysis occur during the interview session.
Scalability, allowing the system to support multiple interview sessions simultaneously.

Efficient decision support, providing HR professionals with summarized candidate insights and confidence scores.

B. System Architecture

The architecture consists of three major components: Frontend Interface, Backend Processing, and Emotion Detection Modules, each responsible for different functions in the interview analysis system.

1. Frontend Interview Interface

The frontend interface enables HR personnel and candidates to interact with the interview system. HR users can configure interview questions and initiate interview sessions, while candidates join the video interview through the platform. During the session, the system captures image frames from video streams, audio clips from speech responses, and optional textual input from candidate answers. These inputs are transmitted to the backend server for emotion analysis.

2. Backend Processing Server

The backend processing layer manages the data processing, model execution, and system communication. A Flask-based API server receives the captured multimedia inputs and forwards them to the corresponding emotion recognition models. The backend includes a model server responsible for executing machine learning models for facial, voice, and text emotion detection. It also coordinates data processing and returns emotion predictions to the main analysis module.

3. Multimodal Emotion Detection Modules

The system includes three emotion detection modules responsible for analysing different types of candidate data.

Facial Emotion Detection:

Facial expressions are analysed using a MobileNetV2 convolutional neural network (CNN) model. Video frames captured during the interview are processed to detect facial features and classify emotions such as happy, sad, and angry. This module enables real-time visual emotion recognition.

Voice Emotion Detection:

Speech signals are analysed using Mel-Frequency Cepstral Coefficient (MFCC) feature extraction technique combined with DNN Model. The extracted features are used to classify emotional states such as calm, fearful, and surprised, providing insights into the candidate's vocal tone and emotional state during responses.

Text Emotion Detection (Optional):

Candidate responses in textual form are analysed using TFIDF feature extraction combined with LSTM-based classification models. This module identifies sentiments such as positive, negative, and neutral, providing an additional layer of emotional analysis.

C. Emotion Fusion and Interview Analysis

The outputs from the facial, speech, and text emotion detection modules are combined in an emotion fusion layer, which integrates predictions from multiple modalities to generate a unified emotion assessment. This fusion approach improves prediction reliability by considering complementary emotional cues from different sources.

The system then generates a candidate evaluation report, which includes the overall emotion score, confidence level, and a summary of detected emotions throughout the interview session. This report assists HR professionals in making more informed recruitment decisions by providing objective emotional and behavioural insights.

IV.ROLEOFARTIFICIAL INTELLIGENCEININTERVIEWEMOTIONRECOGNITION

Artificial Intelligence (AI) plays a crucial role in modern recruitment systems by enabling automated analysis of candidate emotions, behavior, and communication patterns during interviews. AI techniques improve the accuracy, efficiency, and objectivity of interview evaluations by processing multimodal data sources such as facial expressions, speech signals, and textual responses. By integrating machine learning and deep learning models, the proposed system can identify emotional cues and behavioral indicators that assist recruiters in making more informed hiring decisions.

A.Deep Learning for Facial Emotion Recognition

Deep learning models are widely used for detecting emotions from facial expressions in real-time video streams. Convolutional Neural Networks (CNNs) such as MobileNetV2 are capable of extracting facial features and identifying emotional states including happiness, sadness, and anger. These models automatically learn important facial patterns from image frames captured during interviews. By applying lightweight CNN architectures, the system ensures efficient emotion recognition while maintaining high accuracy and computational efficiency.

B.Speech Processing for Voice Emotion Detection

Speech analysis techniques help detect emotional states based on variations in tone, pitch, and speech patterns. In the proposed system, Mel-Frequency Cepstral Coefficient (MFCC) features are extracted from candidate audio responses to capture important speech characteristics. A Deep Neural Network (DNN – Fully Connected Neural Network) is used to classify emotional states such as calm, fearful, and stressed. This approach provides efficient and accurate voice emotion detection suitable for real-time interview analysis.

C.Natural Language Processing for Text Emotion Analysis

Natural Language Processing (NLP) techniques enable emotion detection from candidate responses expressed in text form. In the proposed framework, TF-IDF feature extraction combined with LSTM-based models is used to analyze textual responses and determine sentiments such as positive, negative, or neutral. Text-based emotion analysis complements facial and speech recognition by identifying emotional cues present in candidate language usage and response patterns.

V. PRIVACY, SECURITY, AND ETHICAL CONSIDERATIONS

Privacy, security, and ethical compliance are critical in AIbased interview analysis systems, as they process sensitive candidate data including facial images, voice recordings, and textual responses. The proposed emotion recognition interview system ensures privacy protection by securely handling candidate data during collection, processing, and storage. Interview data are transmitted through secure communication channels, and access to the system is restricted through authentication mechanisms to prevent unauthorized use. Security mechanisms are implemented to protect the system from potential data breaches and misuse. Candidate audio, video, and textual information are processed within controlled server environments, and only the necessary features required for emotion recognition are stored. Encryption and secure APIs are used to ensure safe data transmission between the frontend interview interface and the backend processing server. These measures reduce the risk of data interception and unauthorized access. Ethical considerations are also addressed to ensure responsible use of AI in recruitment. Emotion recognition results are used only as supporting insights for HR professionals rather than as the sole basis for hiring decisions. The system aims to reduce subjective bias in candidate evaluation by providing objective emotional analysis while maintaining transparency through interpretable interview reports. Additionally, candidates should be informed about the use of AI-based emotion analysis during interviews to ensure transparency and consent. Overall, the proposed framework promotes data privacy, system security, fairness, and ethical responsibility, enabling trustworthy deployment of AI-based emotion recognition technologies in modern recruitment processes.

VI. EXPERIMENTALFRAMEWORK

To evaluate the performance of the proposed multimodal interview emotion recognition system, a comprehensive experimental framework is designed. The evaluation process focuses on analyzing the effectiveness of facial, voice, and text emotion recognition modules as well as the overall system performance in automated interview analysis. This section describes the datasets, baseline models, evaluation metrics, and implementation protocols used for system validation. Fig. 2 illustrates the experimental workflow adopted for this evaluation.

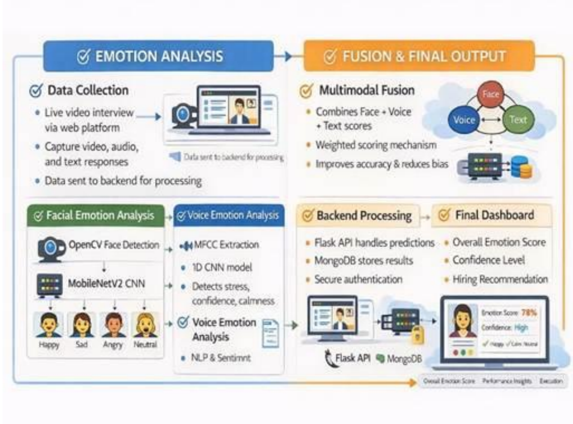

Fig. 2. Experimental framework illustrating the evaluation pipeline of the proposed multimodal interview emotion recognition system including data collection, emotion analysis modules, multimodal fusion, backend processing, and final candidate evaluation dashboard.

A. Data Collection

The system collects multimodal interview data through a web-based interview platform. During the interview session, the system captures candidate information in three formats:

- **Video Data:** Live video frames obtained from the candidate webcam during the interview.
- **Audio Data:** Speech recordings from candidate responses to interview questions.
- **Text Data:** Optional textual responses provided by candidates during interview interactions.

These data streams are transmitted to the backend server for emotion analysis and processing.

B.Datasets

To The proposed system utilizes publicly available datasets commonly used for emotion recognition research:

- **FER2013 Dataset:** Used for training and evaluating facial emotion recognition models. The dataset contains labeled facial images representing emotions such as happiness, sadness, anger, and neutral expressions.
- **RAVDESS Dataset:** Used for speech emotion recognition, containing audio recordings labelled with emotional states including calm, fear, and surprise.
- **Sentiment Text Dataset:** Used for training text emotion detection models to classify textual responses as positive, negative, or neutral.

These datasets enable reliable training and evaluation of the multimodal emotion recognition modules.

C.Baseline Models

To validate the effectiveness of the proposed system, the following models are used for comparison:

- Traditional Machine Learning Models: Support Vector Machine (SVM) and Random Forest classifiers used for emotion detection.
- Single-Modality Emotion Recognition Models: Individual facial, speech, and text emotion recognition models without multimodal fusion.
- Proposed Multimodal System: Integration of facial, speech, and text emotion analysis with fusion-based prediction.

These baseline comparisons help evaluate the advantages of the multimodal approach.

D.Evaluation Metrics

The performance of the proposed system is evaluated using several quantitative metrics:

- **Accuracy:** Measures the percentage of correctly classified emotional states.
- **Precision and Recall:** Evaluate the correctness of emotion predictions and detection capability.
- **F1 Score:** Provides a balanced performance measure combining precision and recall.
- **Processing Time:** Measures system efficiency during real-time interview analysis.
- **Emotion Confidence Score:** Indicates the reliability of emotion predictions.

E.Experimental Protocol

The experiments follow a standardized evaluation process:

1. Train facial, speech, and text emotion recognition models using respective datasets.
2. Extract relevant features from video frames, audio signals, and textual responses.
3. Perform emotion prediction using trained models.
4. Apply multimodal fusion to combine emotion scores from different modalities.
5. Generate final candidate evaluation reports containing emotion scores and confidence levels.
6. This protocol ensures consistent evaluation of the proposed interview analysis system.

F.Implementation Details

The system is implemented using Python-based machine learning and deep learning frameworks. Facial emotion recognition is performed using OpenCV and MobileNetV2 CNN, while voice emotion detection uses MFCC feature extraction combined with a Deep Neural Network (DNN) model. Text emotion analysis is performed using Natural Language Processing techniques with TF-IDF and LSTM models. The backend processing server is developed using Flask API, and the system stores results using MongoDB databases.

G.Reproducibility Considerations

To ensure reproducibility of results, all datasets, model parameters, and training configurations are carefully documented. Experiments are conducted multiple times with fixed random seeds to reduce variance in results. Performance metrics are averaged across multiple runs to ensure reliable evaluation of the proposed system.

VII.FUTURERESEARCHDIRECTIONS

Although the proposed multimodal interview emotion recognition system demonstrates promising capabilities for automated candidate evaluation, several research opportunities remain for further improvement and real-world deployment. Future work can focus on improving the accuracy and robustness of emotion recognition models by incorporating advanced deep learning architectures such as transformerbased models and attention mechanisms. These models can capture complex emotional patterns from facial expressions, speech signals, and textual responses more effectively than traditional neural networks. Another potential direction involves enhancing multimodal fusion strategies. Current fusion approaches combine emotion predictions from facial, voice, and text modalities using weighted scoring methods. Future research may explore more sophisticated fusion techniques such as deep multimodal learning frameworks that dynamically learn relationships between different emotional signals. Improving the real-time performance and scalability of the system is also an important area for future development. Optimizing model architectures and implementing efficient edge-based processing can reduce computational overhead and latency, enabling deployment in large-scale recruitment platforms and online interview systems. Future enhancements may also include advanced behavioural analytics, such as gesture recognition, eye movement tracking, and speech pattern analysis. Integrating these additional behavioural indicators can provide deeper insights into candidate confidence, engagement, and emotional responses during interviews. Finally, future research should focus on improving fairness, transparency, and ethical AI practices in automated recruitment systems. Developing explainable AI techniques can help HR professionals better understand how emotion predictions are generated, while bias detection mechanisms can ensure that the system provides fair and unbiased candidate evaluations across diverse user groups.

VIII. CONCLUSION

This paper presented a multimodal emotion recognition framework for automated interview analysis by integrating facial expression recognition, speech emotion detection, and text sentiment analysis. The proposed system leverages deep learning and natural language processing techniques to analyze candidate emotions from multiple data sources, enabling a more comprehensive and objective evaluation of candidate behavior during interview sessions. The system utilizes convolutional neural networks for facial emotion recognition, MFCC-based speech processing using a Deep Neural Network (DNN) for voice emotion detection, and NLP techniques for text sentiment analysis. By combining these modalities through a multimodal fusion mechanism, the proposed framework improves emotion detection accuracy and reduces the limitations associated with single-modality approaches. The backend architecture, implemented using a Flask API and database storage, enables efficient data processing and generation of candidate evaluation reports. Overall, the proposed system demonstrates the potential of artificial intelligence in enhancing modern recruitment processes by providing data-driven insights into candidate emotions and behavioural responses. The integration of multimodal emotion analysis can support HR professionals in making more informed and unbiased hiring decisions while improving the efficiency and effectiveness of interview evaluation systems.

IX. References

[1] A. N. Salman, M. Alhussein, and A. Alzahrani, "Facial emotion recognition using a mixture of emotion-dependent experts," *IEEE Access*, vol. 13, pp. 21543–21555, 2025.

[2] A. Amjad, M. Hussain, and S. Lee, "Multi-domain adaptation for robust speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 16, no. 1, pp. 125–136, 2025.

[3] A. Sharma, R. Gupta, and P. Singh, "Hybrid deep learning model for facial emotion recognition in interview analysis systems," in *Proc. IEEE Int. Conf. Artificial Intelligence and Machine Learning*, 2025, pp. 312–317.

[4] S. J. Colaco and D. S. Han, "Facial landmark attention network for context-aware facial emotion recognition," *IEEE Access*, vol. 11, pp. 84562–84573, 2023.

[5] O. Araque and C. A. Iglesias, "An approach for emotion detection from text using semantic similarity," *KnowledgeBased Systems*, vol. 145, pp. 139–150, Apr. 2018.

[6] E. Dikbiyik, H. Erdogan, and A. K. Coskun, "BiMER: A bimodal emotion recognition framework using speech and text modalities," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6474–6478.

[7] H. Zhao, Y. Zhang, and L. Wang, "Deep local domain adaptation for cross-corpus speech emotion recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 1567–1579, 2022.

[8] N. Mathur, S. Sharma, and R. Jain, "SAGA: An AIbased recruitment analysis system for automated interview evaluation," in *Proc. IEEE Int. Conf. Smart Computing*, 2023, pp. 215–220.