Research on an AI Interview Evaluation System Integrating Multi-Agent Systems and Virtual Digital Humans

Jiayi Wu[1], Jiaqi Zhang[1], Liuyang Gao[1], Jialiang Feng[1], Bo Meng[1], Yifan Wu[1], Mingming Gong[2]
[1]SArtificial Intelligence and Software Engineering, Henan University of Technology, Zhengzhou, Henan, China
[2]iFLYTEK COLTD, Hefei, Anhui, China

Abstract: Current AI interview systems face challenges in achieving natural interaction, conducting multidimensional assessments, and ensuring interpretability. This paper proposes and validates an intelligent evaluation framework integrating multi-agent collaboration with 3D virtual digital humans (VDH). The system enables complex question-answering and multidimensional evaluation through the coordinated operation of four functional agents: resume analysis, interview skills, written test training, and job recommendation. Leveraging a 3D Virtual Digital Human interviewer, the system supports multimodal fusion interaction encompassing voice, text, and facial expressions. It automatically generates interview questions, processes multimodal data, and produces multidimensional scoring alongside interpretable feedback reports. Experiments demonstrate that this research significantly enhances the immersion and naturalness of human-machine interaction while strengthening the objectivity, professionalism, and explainability of evaluations. It provides an innovative solution and theoretical foundation for intelligent talent assessment.

Keywords: Multi-agent; Virtual Digital Human; AI Interview; Multimodal Interaction; Comprehensive Evaluation; Explainability

1. Introduction

Traditional recruitment interviews suffer from issues such as low efficiency, high subjectivity, and high costs[1]，In recent years, AI interview systems have garnered increasing attention. However, existing systems predominantly employ simple question-and-answer formats, suffering from issues such as insufficient interactive realism, limited evaluation dimensions, and poor decision explainability. To address these challenges, this paper designs and implements an AI interview evaluation system that integrates multi-agent collaboration with virtual digital human interaction. The system achieves multimodal data collection and fusion of voice, video, and text, constructs a multi-agent question-answering and evaluation mechanism, and generates explainable multidimensional feedback reports. Key innovations include: (1) VDH-driven immersive interaction mode; (2) Multi-agent collaborative evaluation framework; (3) Configurable multi-dimensional assessment model with adjustable weighting.

This study aims to provide a viable solution for intelligent recruitment that combines immersive experiences, objective analysis, and personalized feedback.

2. System Design and Workflow

2.1 System Architecture Design

The system employs a clearly layered architecture design. At the user level, it supports applicants accessing the platform via web browsers through cloud or local connections. At the data level, it centrally stores resume documents, multimodal data generated during interviews (including audio, video, and text), assessment results, and model parameters. The application layer serves as the system's core, integrating a high-fidelity 3D virtual digital human interface while managing real-time audio/video capture and rendering workflows. The system invokes a suite of external service APIs, such as speech recognition and facial expression analysis, with the multimodal assessment module conducting comprehensive analysis. The data management module ensures all information is stored in a structured manner and enables efficient access. The detailed system architecture is illustrated in Figure 1.

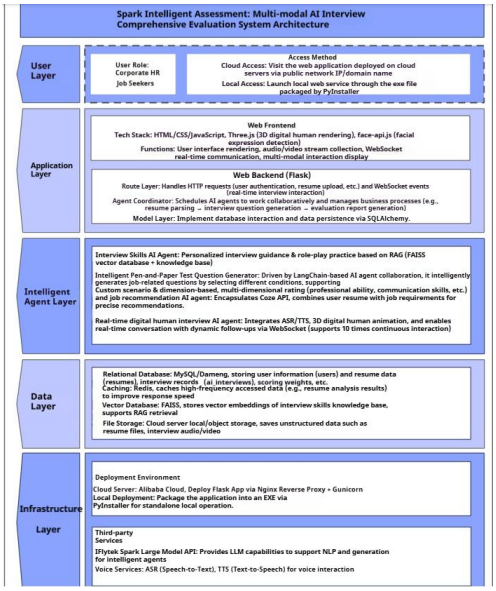

Figure 1. System Architecture Diagram

2.2 System Workflow

The system workflow is divided into three phases: preparation, interview, and reporting. The project flowchart is shown in Figure 2.

During the preparation phase, candidates upload their resumes and select target positions. The system initializes interview parameters based on this information and loads the corresponding evaluation model. Figure 3 shows the resume upload interface.

The interview is conducted by a virtual digital human serving as the primary examiner. The system collects candidates' voice, video, and interactive text data, while the underlying multi-agent system simultaneously performs real-time processing and preliminary evaluation. Figure 4 shows the system's interview interface.

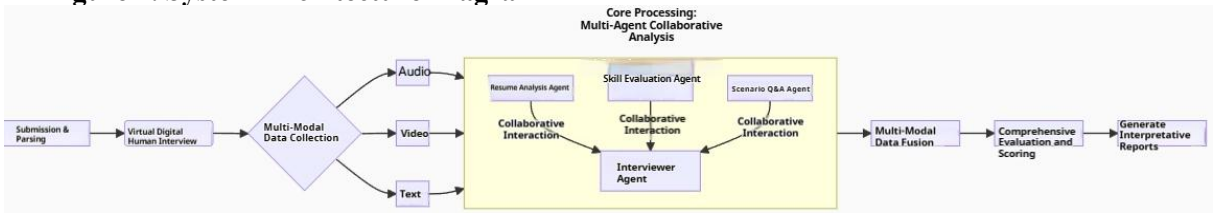

Figure 2. Project Flowchart

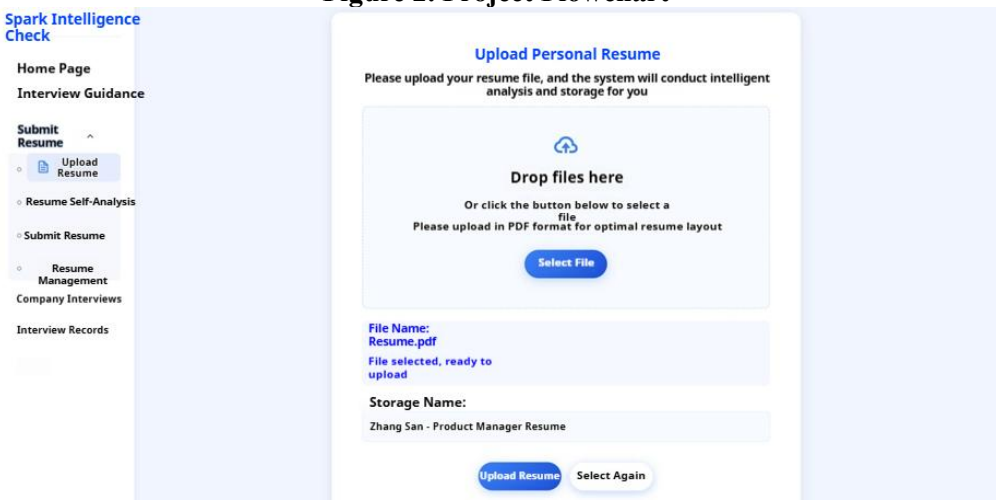

Figure 3. Preparation Phase

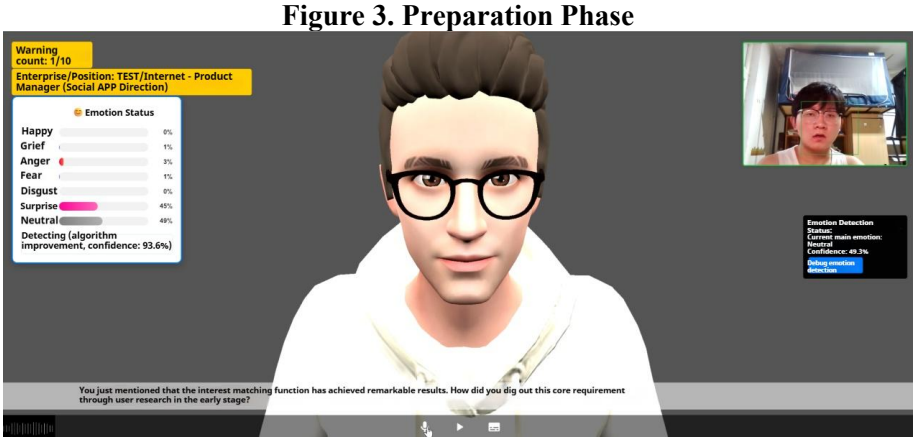

Figure 4. Digital Human Interview

Following the interview, the system will consolidate all data to generate a comprehensive score and a detailed visual feedback report. The scoring encompasses five dimensions: professional competence, learning ability, teamwork, problem-solving, and communication skills. These scores are integrated to produce a multidimensional competency radar chart, as illustrated in Figure 5.

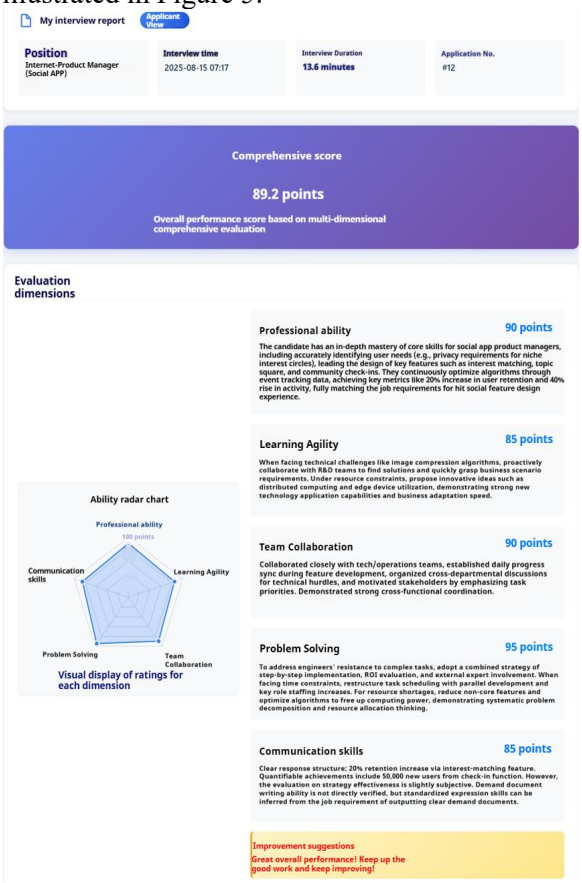

Figure 5. Overall Score

3. Multimodal Information Acquisition and Processing

3.1 Data Source

Data is sourced from multiple origins, including complete historical Q&A records, JSON data generated from multimodal evaluations, and structured scoring tables.[2-3]

3.2 Speech Data Processing

The system employs Intelligent Audio Transcription (IAT) technology to convert real-time audio streams into text. During this process, noise reduction algorithms are utilized to enhance the signal-to-noise ratio, while text-to-speech (TTS) technology converts the text into natural-sounding speech. Figure 6 illustrates a recorded example of speech-to-text conversion. Voice activity detection (VAD) technology is applied to accurately identify the start and end points of speech segments, thereby improving the accuracy and coherence of the transcribed text.[4-6]

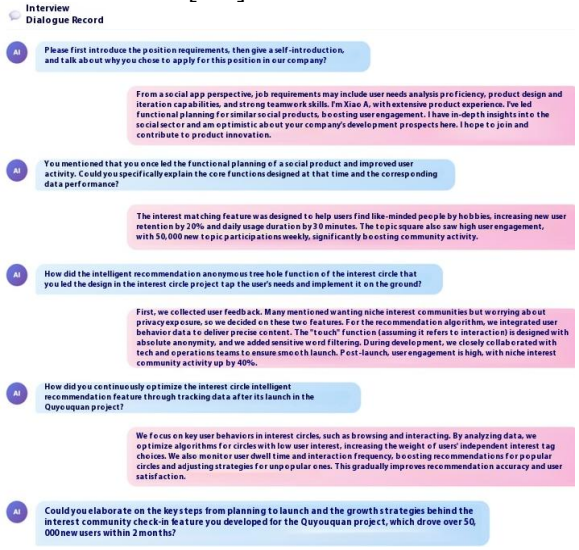

Figure 6. Interview Voice-to-Text

3.3 Video Data Processing

Using computer vision technology, facial images of interviewees are captured in real time via cameras. These images are fed into a pre-trained facial expression recognition model, which identifies subtle changes in expression (such as joy, confusion, or confidence). This nonverbal information is then converted into quantifiable data.[7] Facial expressions serve as critical nonverbal signals reflecting a candidate's emotional state. The system employs computer vision technology for real-time analysis, utilizing the MTCNN algorithm to detect facial regions. On the CPU side, it achieves a detection speed of 30 frames per second with a face bounding box localization accuracy rate as high as 98.6%. The pre-trained FERNet model identifies seven basic emotions: anger, disgust, fear, happiness, sadness, surprise, and neutrality. It outputs probability values for each emotion as evaluation metrics, achieving 82.3% classification accuracy on the CK+ dataset. Facial action unit (FAU) features are extracted—e.g., AU12 (upward mouth corners) signifies positive emotions. And AU4 (brow furrow) signifying confusion. A 5-second sliding window analyzes dynamic emotional changes, providing data support for the "emotion management" dimension assessment, as shown in Figure 7.

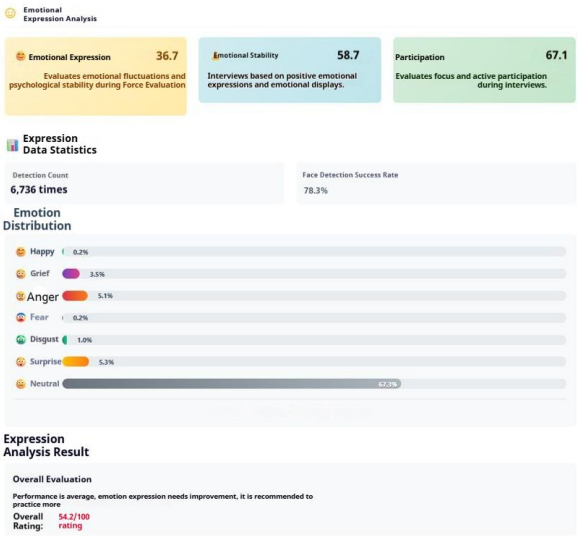

Figure 7. Facial Expression Analysis during Interviews

3.4 Synchronized Management of Text and Data

Utilizing the SocketIO real-time communication mechanism, audio and video streams along with recognized text data are tagged and compared to ensure precise alignment of accuracy across different modalities, preparing the groundwork for subsequent analysis. All raw and processed data is stored in the database. Key tables include ai_interviews for storing basic interview information and interview_scores for storing various scores. Raw audio and video files are stored as files in the database, while structured data—such as recognized text, emotion labels, and scores—is stored in relational format. The interview results synthesized from these two components are saved in the "Interview Records" module, where users can click to view their interview details. [8-9] As shown in Figure 8.

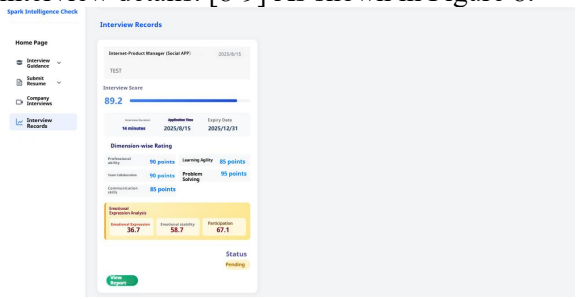

Figure 8. Keep Interview Records

4. Multi-Agent Interview Question-Answering Evaluation

4.1 Multi-Agent Collaborative Model

The system adopts a multi-agent collaborative framework comprising four agents: the interviewer agent, resume analysis agent, skill assessment agent, and scenario-based Q&A agent. These agents collaborate through message passing and shared working memory to ensure a smooth and adaptive interview process. Figure 9 illustrates the multi-agent collaboration diagram.

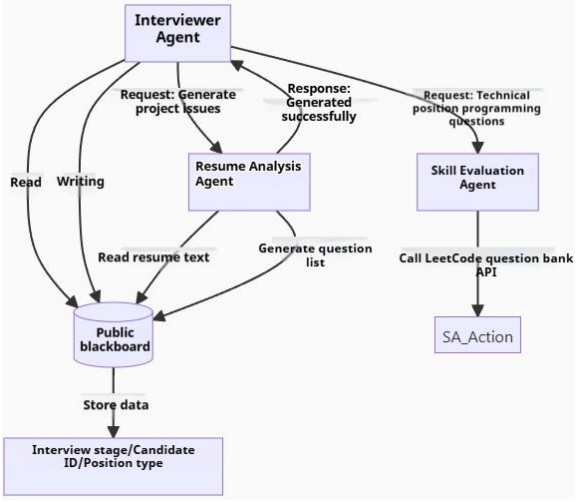

Figure 9. Multi-Agent Collaboration

4.2 Intelligent Agent Function Design

The primary functions of the aforementioned four types of agents are as follows: (1) Interviewer Agent: As the dominant entity in the interactive interface, it controls the interview pace and facilitates questioning and dialogue; (2) Resume Analysis Agent: Parses candidate resumes to generate personalized questions related to experience and skills; (3) Skill Assessment Agent: Focuses on designing technical Q&A to evaluate candidates' professional hard skills and knowledge reserves; (4) Scenario Q&A Agent: Simulates real-world work scenarios to pose hypothetical or behavioral questions, evaluating soft skills such as communication, collaboration, and adaptability. [10] The agents were developed using the Coze application and agent development platform. Figure 10 illustrates the agent construction process.

4.3 Decision-Making Logic Mechanism

The system possesses dynamic adjustment capabilities, intelligently switching the primary questioning agent based on the interviewee's responses and emotional feedback (such as expressions of confusion), thereby achieving a highly personalized and human-like interview process.

4.4 Evaluation Dimension Design

The evaluation system design is divided into two dimensions: one is the content dimension, covering technical competence, learning potential, and communication skills; the other is the nonverbal dimension, which includes facial expressions, speaking pace, intonation, and body language.

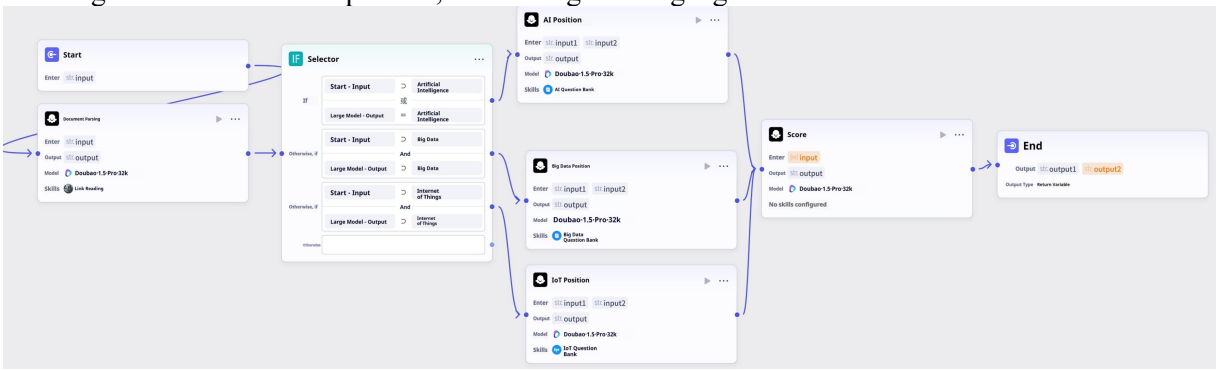

Figure 10. Coze Platform Workflow Agent Development

4.5 Weighted Scoring Model

For the comprehensive scoring, the system employs a weighted sum model, with the specific formula being:

$$S = w_1 \cdot S_{\text{Skills}} + w_2 \cdot S_{\text{Study}} + w_3 \cdot S_{Communication} + w_4 \cdot S_{Emotion} + w_5 \cdot S_{\text{Rate}} + \quad (1)$$

The respective weights (w_1, w_2, ...) can be dynamically configured and adjusted based on the specific requirements of different positions, highlighting recruitment priorities.

5. Experiments and Results Analysis

5.1 Experimental Environment and Dataset

High-performance servers equipped with NVIDIA GPUs accelerate deep learning inference. Furthermore, the dataset includes multimodal samples collected from real resumes, job descriptions, and simulated interviews.

5.2 Experimental Protocol and Evaluation Metrics

This experiment specifically designed a comparison group involving two distinct conditions: single-modality versus multi-modality evaluation, and with versus without VDH interaction. The experimental design is illustrated in Figure 11. The primary evaluation metrics comprise three aspects: first, objectivity, measured by the correlation between system scores and human expert ratings; second, comprehensiveness, assessing whether the system effectively utilizes multimodal information; and third, user experience, evaluated through questionnaires to gather candidate feedback on interaction naturalness and immersion [11].

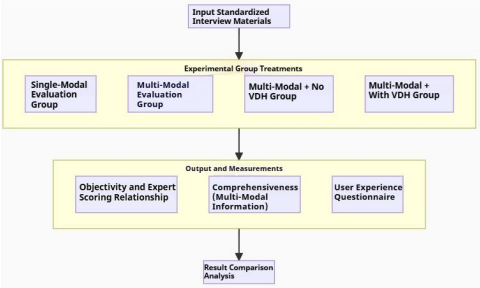

Figure 11. Experimental Design Approach

5.3 Experimental Result and Discussion

Experimental results demonstrate that the system integrating multi-agent systems with VDH outperforms other methods across all metrics. Multimodal evaluation effectively enhances scoring objectivity, achieving a correlation exceeding 0.85 with human ratings. The incorporation of VDH significantly improves interaction realism and user experience, with questionnaire scores increasing by 30%. The system-generated interpretable reports have received high recognition from the company. Figure 12 presents the experimental results.

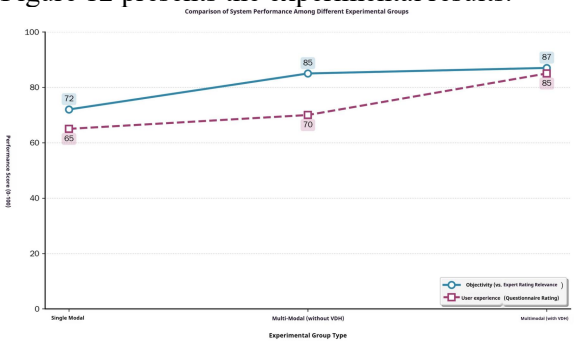

Figure 12. Performance Rating

As shown in figure12, the system demonstrates superior performance under multimodal conditions and with VDH interaction. However, there remains room for improvement in facial

expression recognition accuracy, which will be a key focus for future optimization.

6. Summary and Outlook

6.1 Research Summary

This study successfully developed an AI interview evaluation system integrating multi-agent systems with virtual digital humans. By leveraging collaboration among multiple agents and multimodal information fusion, the system automates the entire interview process—including question-answering, multi-perspective assessment, and report generation. Experimental results demonstrate significant improvements in enhancing interaction immersion, assessment objectivity, and result interpretability.

6.2 Research Outlook

Future work will focus on the following key areas: Continuously enhancing the realism of digital human expressions and movements to strengthen interactive experiences; Expanding system application scenarios, such as group interview training and customer service simulation exercises; Expanding corporate and interview resource databases to enable users to access real-time company operations and recruitment status online, thereby deepening their understanding of various organizations; Optimizing the AI agent's RAG (Retrieval-Augmented Generation) question-answering solutions and process control to drive higher interview success rates.

References

[1]Altınbaş G G,Yorulmaz O .Enhancing Professional Skills: Assessing the Impact of a Computer-Based Training Program on Interview Skills, Self-Efficacy, and Anxiety Levels in Psychology Students. Journal of Rational-Emotive & Cognitive-Behavior Therapy, 2025, 43(4): 51-51. DOI:10.1007/S10942-025-00615-Z.

[2]Fosci P, Psaila G .Evolving J-CO-QL+ with fuzzy evaluators for flexible queryisng of JSON data sets. Neurocomputing, 2025, 633129621-129621. DOI:10.1016/J.NEUCOM.2025.129621.

[3]Gonçalo A, Filipe M ,Rogério C , et al. JSON Schemas with Semantic Annotations Supporting Data Translation. Applied Sciences, 2021, 11(24): 11978-11978. DOI:10.3390/APP112411978.

[4]Kinouchi T, Ogawa A, Wakabayashi Y, et al.Domain adaptation using non-parallel target domain corpus for self-supervised learning-based automatic speech recognition. Speech Communication, 2025, 174103303-103303. DOI:10.1016/J.SPECOM.2025.103303.

[5]Tao Y, Liu J, Lu C, et al. CMDF-TTS: Text-to-speech method with limited target speaker corpus. Neural networks: the official journal of the International Neural Network Society, 2025, 188107432. DOI:10.1016/J.NEUNET.2025.107432.

[6]Chen C, Liu K, Jing J, et al. Wide-Frequency vibration positioning of asymmetric interferometers based on dual endpoint detection and iterative VMD. Optics and Laser Technology, 2025, 187112882-112882. DOI:10.1016/J.OPTLASTEC.2025.112882.

[7]Astrid V, Laura M, Michele B, et al. Automatic detection of charcoal kilns on Very High Resolution images with a computer vision approach in Somalia. International Journal of Applied Earth Observation and Geoinformation, 2023, 125. DOI:10.1016/J.JAG.2023.103524.

[8]Technology-Communication Technology; Researchers' Work from Henan Agricultural University Focuses on Communication Technology (Analysis of Socket Communication Technology Based On Machine Learning Algorithms Under Tcp/ip Protocol In Network Virtual Laboratory System). Internet Weekly News, 2019. 1023.

[9]Dong-Jo K, Hyun-Ju P. A design and implementation of transmit/receive model to speed up the transmission of large string-data sets in TCP/IP socket communication. Journal of the Korea Institute of Information and Communication Engineering, 2013, 17(4): 885-892.

[10]Castillo O, Dinçer H, Yüksel S, et al. Publisher Correction: Assessing carbon–neutral supercapacitors in renewable energy systems with self-improving agent-based molecular fuzzy intelligent algorithms. Scientific Reports, 2025, 15(1): 33712-33712. DOI:10.1038/S41598-025-22245-2.

[11]Christoph T ,Ann-Christin B ,Felix M , et al.Multi-modal framework to model wet

milling through numerical simulations and artificial intelligence (part 1).Chemical Engineering Journal, 2022, 449. DOI:10.1016/J.CEJ.2022.137794.

