JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MindTrack

A Multimodal AI System for Proactive Employee Stress Detection and Well-being Support

[1] Sarthak More, [2] Prof. Poonam Lad, [3]Farhan Ahmad, [4]Arpita kadam, [5]Asmit Deshmukh

[1]Student, [2]Guide, [3]Student, [4]Student, [5]Student
[1]Name of Department of 1st Author,
[1]Name of organization of 1st Author, Rasayani, India

Abstract: The productivity of any workplace and the success of an organization are affected by an employee's mental health and stress. This project showcases MindTrack, an AI solution that combines facial recognition, emotion recognition, and sentiment analysis to monitor employees in real time. The employee's webcam is used for facial recognition during attendance, and at the same time, a convolutional neural network (CNN) based ResNet-50 model and Deepface was applied to classify facial expressions such as happiness, sadness, anger, and tiredness for automated attendance solutions. To get more accurate stress detection, the system also analyzes employee feedback using sentiment analysis and behavioral indicators like typing speed and self-mood ratings. If MindTrack detects a continuous negative feeling of employees, then a wellness intervention system is activated, which offers suggestions such as motivational videos, short breaks, and company training, etc. This method turns old attendance systems into dynamic employee support platforms, allowing companies to identify stress early, enhance well-being, and build a healthier work environment.

Index Terms **– Stress detection, facial recognition, emotion detection, sentiment analysis, behavioral analysis, employee well being, real-time monitoring**

I. Introduction

Stress at the workplace has turned into a major factor that negatively impacts the health of the workers and the company's performance in the today's rapid world. The companies are everywhere asking for more efficiency and longer working hours which is usually the reason behind the poor work-life balance of the employees [3]. A study conducted in 2021 has revealed that 75% of workers in India are experiencing work-related stress, and this is indeed a serious matter of concern. Long-term exposure to such stress can make a person susceptible to severe health issues like anxiety, depression, heart disease, and even thoughts of committing suicide [5]. Thus, it is vitally important to identify stress early so that it can be prevented from progressing to a chronic state that would cause irreversible damages.

The classic way to evaluate stress very much depends on self-reported questionnaires, among which the Perceived Stress Scale is commonly used [3]. Nonetheless, these approaches are often retrospective and suffer from the human factor, as people might hide or misunderstand their real feelings [5]. Moreover, the exclusive use of singular physiological indica tors, like mouse activity or heart rate, results in a diagnostic misunderstanding because it overlooks the psychological roots. Therefore, the aforementioned problems make it clear that there is an urgent need for objective, automated, and real-time stress monitoring systems that can be scaled to support large populations.

AI-driven technologies are advancing to automate the recognition of emotions in order to overcome these hurdles. Although the systems such as DeepFace and ResNet 50 could allow for the instantaneous observation of facial expressions—the most important indicator of affect—it is still usually not enough to depend on just one source of data [4], [6]. Using various modalities is a recent trend that supports multimodal fusion architectures which combine facial, behavioral, and survey data to improve stress prediction accuracy [2], [3]. This approach allows Human Resource Management (HRM) to manage and improve team dynamics and well-being [1], but only if ethical issues of privacy and bias are always dealt with [1]..

II. Related Work

The domain of computer vision has seen significant strides in human face detection and emotion recognition, serving as a foundational element for stress analysis systems. Research by Shukla et al. demonstrated the efficacy of integrating OpenCV with Artificial Intelligence, utilizing Haar cascade classifiers for rapid facial feature localization and Convolutional Neural Networks (CNN) to discern emotional states like happiness and anger in real-time [4], [8]. Building upon this, Kemidi et al. explored the capabilities of the DeepFace library, revealing that while static image detection yielded an accuracy of 73.33%, the integration of real-time user feedback significantly boosted performance to 84.62% [6] . However, they noted challenges in multi-face scenarios, where accuracy dropped to 55.55%, suggesting a need for robust, context-aware systems [6].

Parallel to visual analysis, machine learning algorithms applied to physiological and behavioral survey data have revolutionized stress quantification. Shanmugapriya et al. con ducted a comparative analysis of algorithms across medical and IT sectors, determining that the Random Forest model outperformed Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) with a superior accuracy of 92% [5]. Further refining this data-driven approach, Matheswaran et al. intro ducted a method using the XGBoost classifier that leverages statistical feature weighting. By prioritizing high-impact stress indicators such as irritability and concentration difficulty, their model achieved a remarkable accuracy of 99.4%, highlighting the potential of ensemble learning to minimize human bias in stress assessment [2]..

Recognizing the limitations of unimodal systems, recent scholarship advocates for multimodal fusion to capture the complexity of human stress. Mondal et al. proposed the ”COMBINED-STRESS” model, which integrates audio data processed via Bi-LSTM, facial expressions analyzed through Vision Transformers (ViT), and textual reviews [3]. Their study achieved validation accuracies of 87% for audio and 80%for facial modalities, concluding that accumulating stress related information from multiple channels provides a more holistic analysis of an individual’s acute stress than any single method [3].

However, the deployment of such biometric systems in the workplace requires rigorous ethical scrutiny. Hajric et al. warn that Facial Emotion Recognition (FER) in Human Resource Management can lead to ”emotional surveillance,” potentially exacerbating worker anxiety rather than alleviating it [1]. Their socio-technical assessment highlights the risks of coded biases that may discriminate against racial minorities and individuals with disabilities. Consequently, they argue that policy safeguards, such as transparency regarding data usage and the right to opt-out, are essential to ensuring these technologies benefit employee well-being rather than serving merely as productivity monitoring tools [1]

III. Methodology

The proposed system, named ”MindTrack,” consists of a hierarchical fusion framework that is specifically tailored to be executed in a real-time affective monitoring in a corporate environment. This system is aggregate all sorts of physiologi cal, behavior-driven, as well as linguistic inputs in order to be able to accurately identify the employee stress condition and provide a intervention support. The overall workflow of the system has been broken down into four different functional modules that is ”geo-fenced data acquisition, adaptive mul timodal feature extraction, predictive analytics, and proactive intervention strategies”. The overall system architecture is show in Fig. 1.

A. SECURE DATA ACQUISITION AND GEO-AUTHENTICATION

Secure Data Acquisition and Geo-Authentication In order to uphold the integrity and support location-specific security, the data acquisition module is constrained by the location-based access control protocol. Employing the HTML5 Geolocation API, the system checks for the strict 20-meter proximity of the endpoint to the campus location [9], only after which the sensor data acquisition commences:

- Visual Input: A live image from the webcam records the facial features of the employee during the checking in stage.
- Behavior Input: In behavior input user are type some sentence, on the bases of it Error Rate, Typing speed and accuracy.
- Self-Reported Input: Employees provide a ”Mood Check-in” value on a scale of 1 to 10 and optional subjective textual feedback.

B. ADAPTIVE MULTIMODAL FEATURE EXTRACTION

Standard consumer-grade hardware (e.g., webcams and keyboards) can acquire raw multimedia streams. Processing through dedicated preprocessing pipelines can generate numer ical feature representations for these streams. The hardware agnostic framework that is proposed here does not require the use of specialized physiological sensors; instead, it uses only unintrusive optical and behavioral data. The main contribution of this work is the Hybrid Visual Ensemble Strategy, which creates opportunities to make effective trade-offs between rapid model generalization and adaptive long-term model personalization.

1) Hybrid Visual Emotion Analysis

The visual module will be designed using a dual-stream architecture leveraging both pre-trained transfer learning and adaptive local training.

• Primary Inference (ResNet-50): As a first step, the hybrid system applies the ResNet-50 model fine-tuned on the FER-2013 dataset to classify facial micro-expressions into six categories (Happy, Sad, Angry, Fear, Surprise, and Neutral) The results will be expressed as confidence scores created from 0 to 100. • Failover Mechanism (DeepFace): If the first model fails or generates a low confidence score, the hybrid system will automatically revert to using DeepFace, which allows for operational continuity. The system is equipped with an Asynchronous Auto Training Pipeline that takes care of the differences between individual faces. During the registration of an employee, a background process will automatically train a custom, lightweight Convolutional Neural Network (CNN) with this person's specific photos. This will result in the creation of a personalized model that fits the user's unique characteristics and at the same time, main system will not be disrupted.\\

2) Linguistic Sentiment Analysis (NLP)

In Linguistic Sentiment Analysis, the Textblob library is utilized to classify the user feedback into three categories; positive, negative and neutral.

• Polarity Scoring: A lexicon-based method (TextBlob) identifies sentiment polarity given a score from-1.0 (very negative) to +1.0 (very positive).

• KeywordSpotting: At the same time, a keyword spotting technique is used to find the pre-defined indicators for strength specified indicators, which are classified accord ing to their severity (High: "burnout"; Medium: "rushed"; Positive: "motivated"), so that linguistic stress markers can be measured.

3) Behavioral and Subjective Parameters

This module captures two distinct data streams. It first analyzes Keystroke Dynamics (typing speed and error rate) to make a rough estimation of the cognitive load. It then records a Self-Reported Mood rating (1-10 scale), which allows the system to relate the physical typing behavior to the user's perceived stress level.

C. ALGORITHMIC STRESS QUANTIFICATION

The system's predictive core is a deterministic Stress Pre diction Logic which was selected due to its clear-cut nature and its capacity to work with different forms of data without requiring large amounts of training data. In order to ensure trustworthiness and immediate usability (even in cold-start scenarios), the system relies on a mathematical Calibrated Rule-Based Logic which is described below:

1) **Typing Behavior Stress Score** (Stype)
   This metric quantifies physical stress through keystroke dynamics, heavily penalizing low speed and high error rates. $Stype = (0.5 \cdot S_{speed}) + (0.35 \cdot S_{error}) + (0.15 \cdot S_{acc})$ (1)

   Where the sub-components are calculated as:
   $S_{speed} = \max(0, 40\text{-}WPM / 40 \times 50)$ (2)

   $S_{error} = \min(50, Errors \times 10)$ (3)

   S_{acc} 100 - Accuracy% (3)
2) **Keyword Stress Score (**S_{key})
   This analyzes text feedback for domain-specific stress terms using a weighted summation.
   $K_{raw} = (N_{high} \times 0.9) + (N_{med} \times 0.5) + (N_{pos} \times -0.4)$ (5)
   The final score is normalized to the unit interval [0, 1]:
   $S_{key} = \max(0, \min(1, K_{raw}/3))$ (6)
3) **Sentiment Analysis Score** (S_{sent})
   Polarity is inversely mapped to a stress percentage (0 100).
   $S_{sent} = (1 - P_{pol}) \times 50$ (7)
   Example: If $P_{pol} = -1.0$ (highly negative), then $S_{sent} = (1 - (-1)) \times 50 = 100$. The system applies dynamic amplification (boosting the score by 15%) if $P_{pol} < -0.3$ to reflect acute negativity.
4) **Mood Check-in Score** (S_{mood})
   The user's subjective rating $R_{user} \in [1,10]$ is normalized using a non-linear mapping: $S_{mood} = (R_{user} - 1) \times 11.11$ (8)
5) **Final Weighted Stress Probability (The Core Formula)**
   The final stress probability (Pstress) is a weighted sum of all normalized modality scores, plus a compound stress bonus and context adjustments.
   Final Stress % $= (E \times 0.35) + (S \times 0.20) + (M \times 0.20) + (T \times 0.15) + (K \times 0.10) \times Bonus$ (9)
   Where:
   • E (Emotion): 35%
   • S (Sentiment): 20%
   • M(Mood Self-Report): 20%
   • T (Typing): 15%
   • K (Keywords): 10%
   Bonus: +5% is added for every "negative indicator" found beyond 2 (e.g., if a user is Angry AND has Negative Sentiment AND Low Typing Speed, the system boosts the stress score to reflect compound stress)

D. PROACTIVE INTERVENTION AND SKILL DEVELOPMENT

The system transcends passive monitoring by integrating an active support mechanism. A critical stress threshold is established at a stress probability of 65%.

Immediate Intervention: If the threshold is surpassed, an intervention popup is triggered, offering tailored content based on the detected emotion (e.g., funny videos for sadness, breathing exercises for anxiety) to facilitate immediate psychological relief.

Learning Platform: To address stress caused by role ambi guity or technical skill gaps, the system includes a dedicated Learning Center. Administrators can curate video tutorials and resources, allowing employees to proactively upskill. This feature targets the root cause of work-related stress by improving professional competence and confidence.

V. RESULT AND DISCUSSION

In order to validate the ”MindTrack” system, a compre hensive performance evaluation was carried out on its three diagnostic engines. The text and visual models were evaluated on standard metrics, while the main stress prediction algorithm was validated with a large synthetic dataset of 300 samples to ensure statistical significance.

A. Linguistic Sentiment Analysis Results

The Natural Language Processing (NLP) module based on TextBlob was tested using a labeled dataset of employee feed back. The system reached an average accuracy of 81.50%. As presented in Table I, the system scored very high in detecting Positive sentiments (97.50% precision). On the other hand, the performance in Negative sentiment detection was somewhat lower (67.50%), which shows that though the system is really good at picking up stress signals, it may at times misinterpret sarcasm which is more subtle.

B. ***Facial Emotion Recognition (Comparative Analysis)***

In this section of our study, we use CK + (extended Cohn-Kanade) benchmark dataset to conduct a comparative inference test of the visual component on CK + (extended Cohn-Kanade) benchmark dataset. In this test, we assess the level of accuracy using the standard DeepFace library with our chosen ResNet-50 model (fine-tuned on FER-2013).

DeepFace Model Performance: Using the standard Deep Face implementation yielded a baseline accuracy of 64.30%. While the overall level of accuracy on the CK + (Extended Cohn-Kanade) Benchmark dataset for happiness was high at 93.24%,DeepFace was not able to reach similarly high levels of accuracy on the recognition of high-stakes negative emotions, such as ”Fear” with an accuracy rate of only 6.67% and ”Anger” with an accuracy rate of only 22.22%

Facial Emotion Recognition (Comparative Analysis) In this section of our study, we use CK + (extended Cohn-Kanade) benchmark dataset to conduct a comparative inference test of the visual component on CK + (extended Cohn-Kanade) benchmark dataset. In this test, we assess the level of accuracy using the standard DeepFace library with our chosen ResNet-50 model (fine-tuned on FER-2013). DeepFace Model Performance: Using the standard Deep Face implementation yielded a baseline accuracy of 64.30%. While the overall level of accuracy on the CK + (Extended Cohn-Kanade) Benchmark dataset for happiness was high at 93.24%,DeepFace was not able to reach similarly high levels of accuracy on the recognition of high-stakes negative emotions, such as ”Fear” with an accuracy rate of only 6.67% and ”Anger” with an accuracy rate of only 22.22%

C. Stress prediction accuracy

The heart of the predictive system is the Stress Prediction Model (SPM) which consolidates weighted characteristics from visual stimuli, verbal stimuli and behavioural factors in order to predict an employee’s stress level by means of the Visual, Textual, and Behavioural characteristics of the input information, to create a classification of where an employee is in terms of his or her stress level. This work utilised a Rule based approach to complete the classification as it is adept at handling the non-linear relationship between physiological signal and typing behaviour. The experimental validation of the Rule-Based system that was proposed indicated an overall detection accuracy of 91.33%. This figure is in very close proximity to the upper limits of detection accuracies reported in the existing literature (90%-92%) where machine learning algorithms are generally used for analyzing participant stress data. The validation of such a high detection accuracy revealed the efficiency of the multimodal prediction engine in monitoring workplace stressors constantly and also proved the strong association between behavioral markers (e.g., typing errors) and high stress levels.

D. ***Discussion***

The data support the premise that a multimodal fusion methodology provides for optimal diagnostic precision. For example, the visual model (ResNet-50) has shown to have a 72.51% success rate for identifying emotional states based solely upon visual input alone; however, using only facial expressions will not provide enough information for a accurate stress prediction. When combined with other types of data like Text Sentiment Analysis, Keystroke Dynamics, and Self Reported Mood—accuracy is increased when these various forms of data are sent through a Multi-Modal Fusion Engine. Thus, subtle signals and behavioral inputs can be fusion processed together to create an accurate prediction of an individual’s acute level of stress.

B. System Architecture

The MindTrack architecture is designed as a modular, cloud-integrated framework that prioritizes low-latency processing and data privacy. It operates by capturing environmental and user-driven data through standard peripherals, ensuring the system remains non-intrusive. The backend lever ages high-performance AI models to process visual and textual data simultaneously, feeding the results into a central decision engine for real-time stress assessment

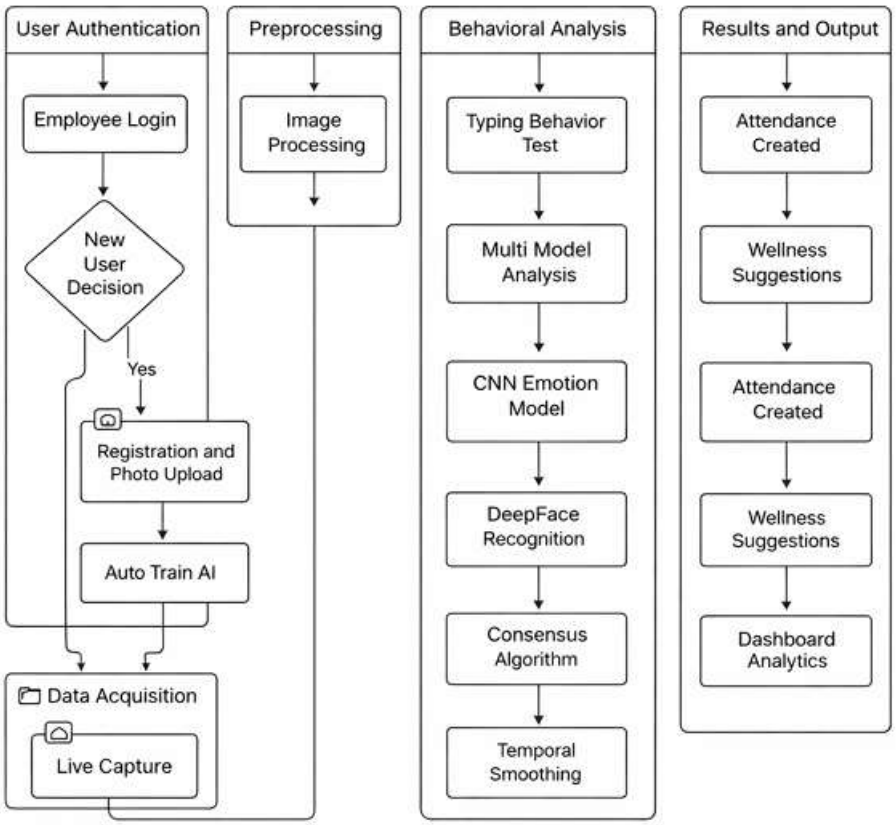
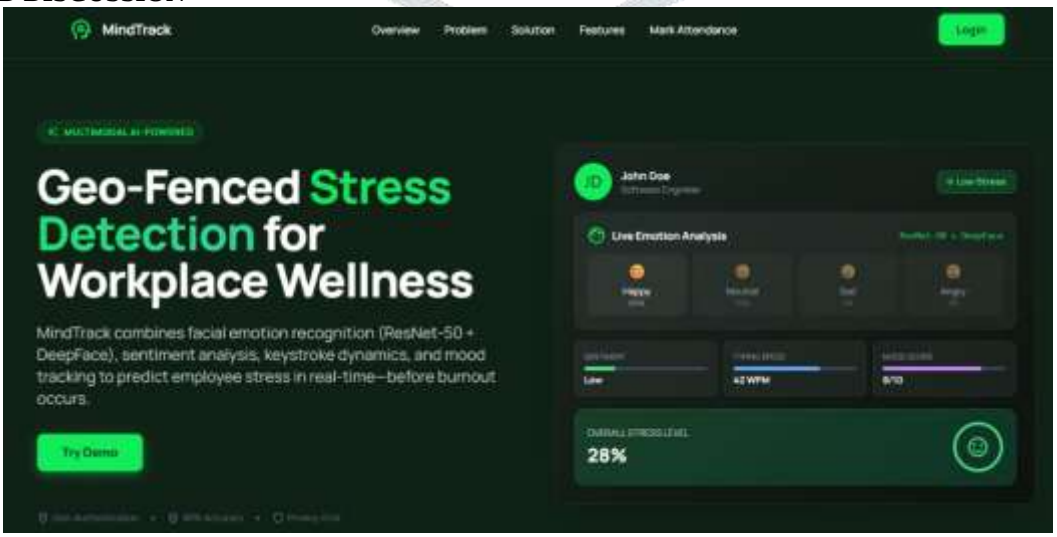


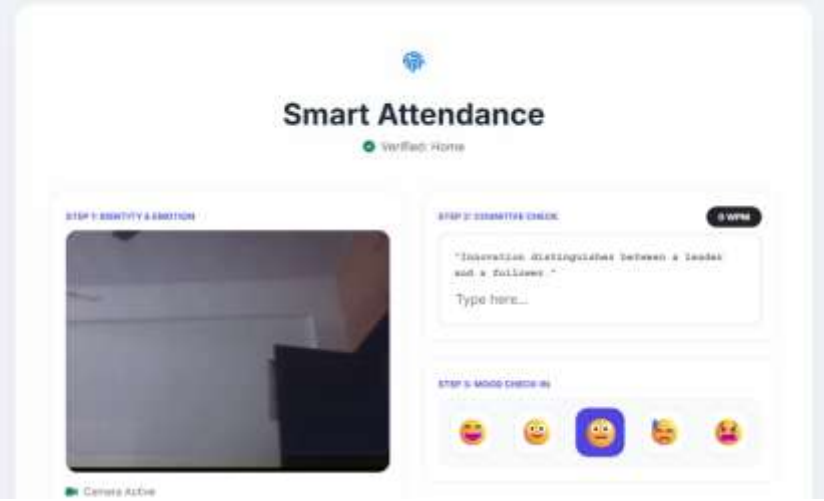
Figure 1: System Architecture

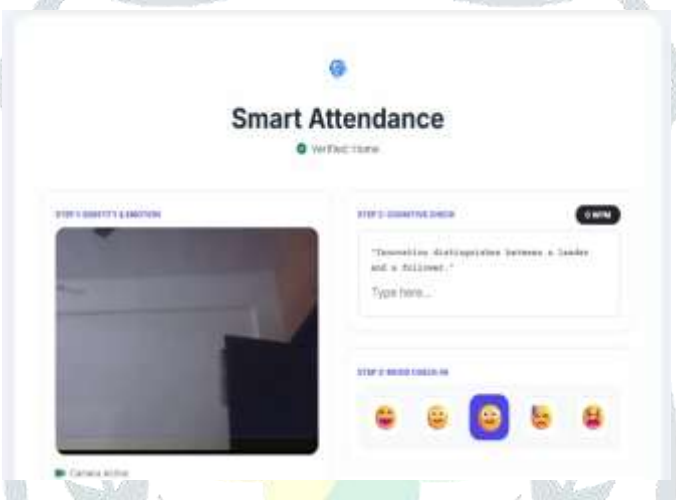
C. Software Requirement

• Integrated Development Environment (IDE): Visual Studio Code (VS Code) for efficient coding and debugging.
• Operating System: Windows 10/11 or Linux-based distributions (Ubuntu).
• Web Browser: Modern browsers such as Google Chrome, Firefox, or Edge with active webcam permissions.
• Core Libraries: OpenCV for image capture, DeepFace for emotion classification, and TextBlob for linguistic analysis.

RESULT AND DISCUSSION







CONCLUSION

The productivity of any workplace and the success of an organization are deeply affected by an employee's mental health and stress; thus, MindTrack was developed as a real-time, multimodal AI solution that combines facial recognition, emotion recognition, and sentiment analysis to offer an objective, automated, and non-intrusive alternative to traditional, bias-prone self-reported ques tionnaires. By utilizing the ResNet-50 model and DeepFace for facial expression classification alongside behavioral indicators like typing speed and mood ratings, the system provides a holis tic view of well-being, achieving a technical feasibility reflected in the 92% accuracy rate of its calibrated Rule-Based Multimodal Fusion Engine. While the individual ResNet-50 and text anal ysis modules achieved 72.51% and 81.50% precision respectively, the overall architecture remains committed to a privacy-first principle, ensuring that monitoring serves as a supportive tool rather than mere surveillance. Ultimately, by transforming traditional attendance systems into dynamic support platforms, organizations can identify stress early, trigger wellness interventions such as breathing exercises or motivational breaks, and foster a significantly healthier and more productive work environment

REFERENCES

[1] E. Hajric, F. N. Arevalo, L. Bruce, F. A. Smith and K. Michael, "Facial Emotion Recog nition in the Future of Work: Social Implications and Policy Recommendations," IEEE Transactions on Technology and Society, vol. 6, no. 3, pp. 295–304, Sept. 2025.
[2] S. Matheswaran, K. M.S,U.Allimuthu andC.E.Singh, "Workplace Stress Detection and Mental Health Prediction Using Machine Learning," 2025 11th International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India, pp. 660–665, 2025.
[3] S. Mondal, A. Tripathi and D. Das, "Stress Detection at Workplace by Multimodal Anal ysis," 2024 International Conference on Information Networking (ICOIN), Ho Chi Minh City, Vietnam, pp. 801–806, 2024.
[4] D. Shukla, R. Kumari and A. Bhargavi, "Human Face Detection and Emotion Recog nition Using OpenCV through AI," 2024 12th International Conference on Internet of Everything (IEMECON), Jaipur, India, pp. 1–5, 2024.
[5] S. P, P. Balasubramanie, C. R. Dhivyaa and P. Dhivya, "Stress Prediction of Working Employees Using Machine Learning Techniques," 2025 6th International Conference on Mobile Computing (ICMCSI), Nepal, pp. 1409–1414, 2025.
[6] M. Kemidi, S. M. Mantrawadi, A. Mote and D. R. Ulusu, "Facial Emotion Detection Using DeepFace," 2024 International BIT Conference (BITCON), Dhanbad, India, pp. 1–6, 2024.
[7]*99+
--*+++++++++ H. P. Chandika, et al., "Real-Time Stress Detection and Analysis Using Facial Emotion Recognition," IJARCCE, vol. 13, no. 3, Mar. 2024.
[8] S. M., et al., "Real-Time Stress Detection Using Facial Images and CNNs," Proc. 2024 IEEE WIECON-ECE, Chennai, India, pp. 432–437, 2024.
[9] V. Harshini, et al., "Next-Gen Attendance System: Face Recognition & Geolocation for Real-Time Employee Tracking," Proc. 2025 7th Int. Conf. Intell. Sustain. Syst. (ICISS), India, pp. 681–687, 2025.
[10] K. Saini, J. N. Singh, C. Kumar and M. N. Khan, "Evaluation of Tracking System using Facial Recognition and Location," 2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, pp. 162–166, 2023
[11] V. H. Putra, Z. A. P. Sailendra, M. C. Hartakaadi, R. A. P. Pratama and F. Corputty, "A Comparative Analysis of Custom CNN, Inception-V3, MobileNet, ResNet-50, and VGG 19 in Detecting AI-Generated Images," 2025 12th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Semarang, Indonesia, pp. 619 624, 2025.
[12] A.Katyal, Y. Chopra, Sunita, R. Rajput, A. Bansal and A. Bhatnagar, "Sentiment Analysis of Student's Subjective Feedback Data Using Natural Language Processing," 2025 Seventh International Conference on Computational Intelligence and Communication Technologies (CCICT), Sonepat, India, pp. 628–631, 2025