Detection of determinants of mental health illness using multimodal data: Focus on toxic work culture

Basavaraj N Hiremath
Dept. of CSE
Dayananda Sagar University
Bangalore, India
basavaraj-cse@dsu.edu.in

Gopal Das C M
Dept. of Psychiatry
CDSIMER
Bangalore, India
drgopaldas.cdsimer@dsu.edu.in

Mokka Hema Sai
Dept. of CSE
Dayananda Sagar University
Bangalore, India
mokkahemasai@gmail.com

Pavithra A
Dept. of CSE
Dayananda Sagar University
Bangalore, India
pavithraanand475@gmail.com

Padmashree P
Dept. of CSE
Dayananda Sagar University
Bangalore, India
padmashree0250@gmail.com

Mourya J P
Dept. of CSE
Dayananda Sagar University
Bangalore, India
19270mouryajp@gmail.com

Abstract— **Mental health cases throughout workplaces are growing within IT sectors and corporate positions due to disruptive workplace environments and stressful conditions. Detecting mental health decline at its early stages enables immediate supportive measures to intervene properly. This research introduces a multi-sourced method for mental health determinant detection through text and voice insights. The assessment collects complete mental wellbeing data through implemented sentiment techniques and linguistic analysis methods. The system implements XGBoost ensemble and CNN and LSTM networks and makes use of advanced machine learning models for precise identification of mental distress symptoms. Data collection was performed through simulated corporate circumstances while applying preprocessing methods to scale features and balance data and apply labels to improve model efficiency. The method achieved beneficial accuracy levels when identifying mental health problems which proves the capability of text and voice analysis to provide thorough assessments. The proposed framework works as an anticipatory system that detects mental health needs early and provides management for mental health needs to protect workers from toxic work culture effects. The upcoming project involves real-time implementation and plans to broaden the data collection range across various workplace environments.**

Keywords— Mental Health Detection, Toxic Work Culture, Multimodal Data, Machine Learning.

I. INTRODUCTION

Mental health disorders have become a pressing issue in modern workplaces, especially within the IT and corporate sectors, where high-pressure tasks, tight deadlines, and performance expectations are commonplace. In such environments, toxic work culture marked by poor communication, lack of support, unmanageable workloads, and unrealistic expectations can lead to heightened stress and emotional burnout. As mental health of an individual decay, they may experience depression, anxiety and other issues which can disrupt their productivity and overall well-being.

Key Workplace Determinants of Employee Mental Health:

- Flexible working arrangements.
- Targeting job changes to workload or break.
- Nudging strategies.
- Targeting the physical work environment.
- Providing performance feedback or reward.
- Job content and strain
- Interpersonal relationships at work, workplace bullying
- Person's role in an organization, role ambiguity and role conflict are associated with depression outcomes.
- Career development, job insecurity is related to higher risk of depressive symptoms.

The side effects of toxic work culture are not only affecting the employees but also the organizations, as they lead to performance reduction, absenteeism and employee churn. Regardless of the raising awareness about workplace mental health, available methods for detecting mental distress frequently depend on self-reporting which may not avoid capturing subtle or unspoken signs of pain. In addition, employees may hesitate to unhesitatingly express their feelings due to fear of shame or negative career out-turn.

Over the past few years, advancements in technology have expanded new possibilities for mental health monitoring by using multimodal data, which includes both text and audio inputs. Studying employee communication patterns, voice characteristics, and linguistic hints provides more objective approach to detecting mental health worsening. Integrating multimodal data can help catch fine changes in tone, pitch, linguistic style, and sentiment that may indicate emotional suffering.

Technical analysis involves assessing parameters such as pitch, jitter, shimmer HNR and intensity, complemented by textual sentiment evaluations and linguistic analysis. We use advanced machine learning models consisting of CNN and LSTM networks and XGBoost and Gradient Boost to precisely detect signs of stress, anxiety and depression. The models receive their training through labeled multimodal data while feature extraction algorithms maintain the combination of acoustic and textual indicators. The main purpose of this investigation involves building a reliable detection mechanism which detects initial mental health indicators that stem from toxic workplace conditions. Organizations gain the ability to implement preventative action through accurate at-risk employee detection which results in creating an enhanced supportive healthy work environment. Through implementation of this system mental health experts gain the ability to identify workplace stressors thus enabling them to provide prompt assistance.

This project aims to identify the key factors contributing to common mental illnesses within a cohort exposed to a toxic work culture. It focuses on developing a mental health detection system using text and audio analysis to recognize early signs of stress, anxiety, and emotional distress in workplace communication. By leveraging technology to analyse vocal features such as tone and pitch, the system can detect stress and emotional irregularities from audio data. Additionally, it will provide real-time insights and early warnings, enabling timely interventions to promote mental well-being and improve workplace productivity. The scope of this project is limited to intellectuals working in office environments, the IT industry, and education sectors.

The subsequent sections of this paper will detail the related work, data collection and preprocessing, methodology, experimental results, and future directions. Through rigorous experimentation and analysis, we aim to demonstrate the effectiveness of our proposed approach in detecting mental health risks in corporate settings.

II. Literature Survey

Studies during the past couple of years have intensely researched mental health detection in workplace environments where high stress is prevalent. Studies have investigated numerous detection methods for employees' mental distress symptoms because of rising mental health problems from unhealthy workplace environments by combining text and voice data analysis. Before 2019 mental health assessments used two approaches which included both self-report questionnaires and structured clinical interviews with PHQ-9 and GAD-7 scale. The tools delivered important diagnostic findings although their effectiveness was reduced through self-report flaws together with hidden reporting fluctuations because of social stigma fears (Smith et al., 2015) [1]. The capabilities to monitor changes in real-time which workplaces require diminish the effectiveness of assessments performed in controlled environments.

Scientists have extensively researched the ability to identify mental health conditions through voice characteristics examination. Eyben et al. (2013) confirmed that the combination of pitch alongside jitter, shimmer, HNR and intensity parameters successfully indicates symptoms of mental distress including stress and depression [2].

SVM and Random Forest classifier analysis of audio data yielded encouraging results according to their research work. Alhanai et al. (2017) introduced work combining speech and linguistic features for depressive state recognition using LSTM networks that managed speech temporal variation [3]. Depressive symptom prediction outcomes were enhanced by voice modulation features according to the research. The activity carried out by Gideon et al. (2019) created a model connecting acoustic feature extraction with deep learning methods for the enhancement of depression detection from vocal tone [4].

An experiment by Cummins et al. (2018) utilized the prosodic and spectral approaches to test whether there were conditions such as depression and PTSD. The study utilized Random Forest classifiers with successful discrimination of the cases and controls [5]. While this is an advancement in the aim of achieving automated classification, another hurdle lies in processing the data in real-time and deploying the model to numerous diverse environments.

Workplace discussions and emails have also been analyzed to infer states of mind. Calvo et al. (2017) have developed a model for detecting stress and strain which performed sentiment analysis based on emotional tone and linguistics [6]. Park et al. (2018) have also applied machine learning techniques to analyze social media data for application in the detection of depressive symptoms with a consideration of sentiment and use of vocabulary [7]]

Chancellor et al. (2019) conducted a wider analysis of social media posts to identify trends in mental state and employed deep learning techniques to classify posts that had depression and anxiety symptoms [8]. The authors described how sentiment analysis and intention analysis can be employed to detect individuals with such conditions from their online behavioral trends.

The fusion of audio and text data to enhance accuracy has gained popularity with time. Satt et al. (2017) designed a multimodality system that combined depression detection through acoustics and analysis of language features and outperformed a unifunctional system [9]. Their model employed CNN for the audio and RNN for the text and proved the benefit of combining different forms of data.

New approaches to identifying mental illness have been made possible over the past few years by the progress in deep learning. Zhang et al. introduced the finding of a hybrid model between CNN and XGBoost for the detection of depression from speech in their 2020 publication. The model outperformed previous studies since it was stronger and more universal [10].

The synergy of CNN's strong feature extraction power and XGBoost's classifying capabilities played a major role in boosting accuracy. Kumar et al. (2021) also proved to be able to perform remarkable results for audio-text joint analysis using a CNN-LSTM model, where both temporal and spatial information had to be combined [11]. Despite the fact that the detection systems have been improved to become more precise with the inclusion of multimodal data, there are different issues that still exist, and among them are privacy and ethics of the data. The importance of proper anonymization of the sensitive data without reducing the estimation accuracy was raised by Tsakalidis et al. as a problem to be solved in the pursuit for effective mental health monitoring systems [12].

LSTM work has shown that it is also possible to model speech data temporal dependencies using such networks. Menezes et al. (2020) presented an LSTM-powered system for estimating vocal stress and fatigue among call center representatives. Using monitoring voice pitch and voice strain, they were able to predict burnout due to stress [13]. Others have similarly reported successful uses of combining the LSTM and CNN architectures for spatiotemporal dynamics in audio features.

Besides that, the ensemble of CNNs and XGBoost has been an effective ensemble technique. Wang et al. (2021) introduced a new approach to stress detection of speech that utilized CNN to extract features and XGBoost for classification. The model was more accurate and stable than standalone models of CNN and XGBoost [14]. The hybrid approach was able to counter the problem of generalization to novel data by exploiting the power of deep learning and the strengths provided by the gradient boosting techniques.

Multimodal fusion refers to the combination of audio and text data that has gained popularity in recent years for gathering different aspects of mental health. Ma et al. (2022) developed a multimodal model that combines audio sentiment features and text sentiment features, using a CNN-RNN hybrid model for audio and speech mapping [15]. This combination technique was particularly suitable for identifying subtle symptoms of depression and anxiety in speech in a work setting. Despite sufficient effort, all studies of real-time automation suffer from the accuracy problem when trying to generalize to workplaces in other fields. The root of the problems are noise in audio recordings, absence of context information in texts and other sources. Concurrently, at the most practical level lacking large multi modal datasets appears to be the overriding limiting factor. Fixing these problems is crucial to developing successful and realistic detection systems of mental health challenges monitoring.

III. METHODOLOGY

The session reports about the steps taken to recognize the determinants of mental health illness due to toxic work environment. It states how the data was collected, what preprocessing steps were taken, what features were extracted, and what machine learning models were applied for analysis.

Dataset

The dataset for this project is taken from a 2014 survey which aims to evaluate the attitudes towards mental health and the measured prevalence of mental health issues in the technology sector. This study is unique in the sense that it examines systematically mental health at the workplace using many methods and data sources to analyze stress, anxiety, and depression and their causes and effects.

This dataset was the result of the application of a purposive designed survey with open ends geared to capture multimodal data from respondents. The survey was designed so that respondents could provide text and audio answers to assess indicators of mental health in the workplace. The specific purpose was to identify if there are indications of stress, anxiety, or depression which could be attributed to toxic work environment.

Voice analysis

We used Praat software to extract voice features. Praat is a software which is used to extract dynamic features from the speech data. The below figure shows the spectrogram of a female voice. Sentence: ”The sun was shining brightly, and people were strolling through the park.”

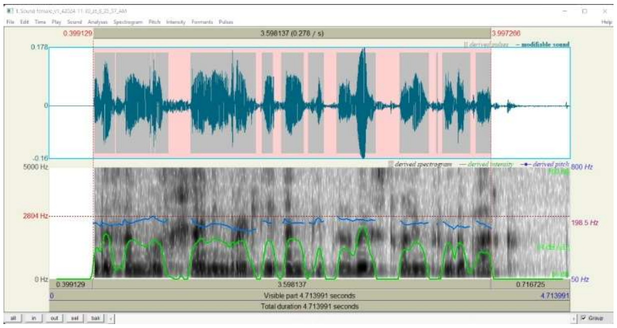

Fig. 1. Spectrogram of Female Voice (Age:21)

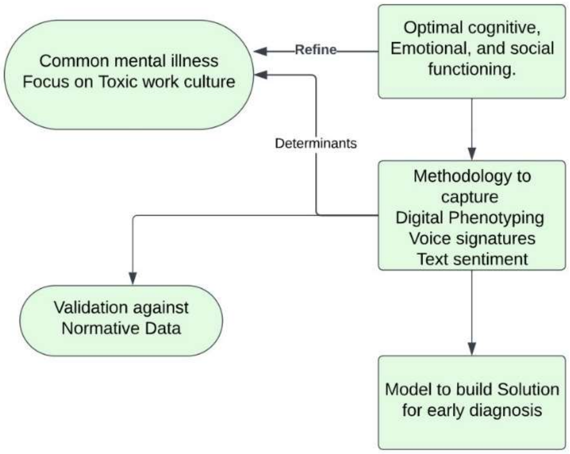
Common mental illness
Focus on Toxic work culture
Refine
Optimal cognitive, Emotional, and social functioning.
Determinants
Methodology to capture
Digital Phenotyping
Voice signatures
Text sentiment
Validation against Normative Data
Model to build Solution for early diagnosis

Fig. 1. Proposed Architecture

A. *Identifying Determinants of Mental Health Issues*

The methodology revolves around understanding the causes of mental health problems, primarily those stemming from a toxic work culture. The main aim here is to identify the prevalent mental health issues that are harmful to employees in the corporate world. These causes are assessed in relation to their cognitive, emotional, and social functional triad. By evaluating how toxic work culture influences these three integral areas, it is possible to create specific plans for early diagnosis and prevention.

B. *Digital Phenotyping and Data Collection*

This is where digital phenotyping comes in capturing the construction of data with respect to the mental health indicators. In this case, data is gathered by utilizing two techniques: voice signatures and text sentiment analysis. The former reveals emotional tone and vocal stress while the latter detects emotional and cognitive changes in writing. Data is collected from different modes of communication, such as text messages, emails, and recorded speech, so that you can analyze the emotional tone, sentiment and cognitive speech. This strategy guarantees detection of the changes that most people would label as minute but are precursors of great mental distress.

C. *Model Building for Early Diagnosis*

The subsequent stage would be to create mental health issue preventive diagnosis model systems based on the previously gathered information in the systems. The implemented machine learning models are Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and XGBoost. The models are selected due to their proficiency in identifying intricate features and irregularities associated with mental health problems. The target of model building is to build a sophisticated model that will be able to identify critical signs of stress or emotional discomfort to allow for early diagnosis. Accurate predictions with few false positives or negatives are achieved by training the models on annotated datasets.

D. Validation against Normative Data

The model predictions cannot be relied on until they have been validated against the set normative data, and so the results are set against standard data. In this instance, the model outcomes are scrutinized with known criteria and the results set parameters, estimating how well the model functions. Evaluating the system against detected patterns lowers the chances of false matches significantly. This accuracy also proves useful in marking the inadequacies in the models and focusing on the well-defined aspects and parameters.

E. Refinement and Optimization

In the last step of the model, further modifications and optimizations are made towards the created model. With the passage of time and availability of new data as well as changes in the workplace settings, this model is upgraded every so often to ensure precision and pertinency. This process involves model retraining with new data, hyperparameter tuning, and performance improvement through feedback incorporation. The focus is to aid the system in serving the best possible cognitive, emotional, and social functioning of the employees by facilitating accurate and timely information regarding mental health issues.

The proposed solution focuses on detecting mental health issues with toxic work environments using digital phenotyping and machine learning techniques. The system captures voice and text patterns which expresses emotions so the system captures cognitive and emotional signals associated with stress or anxiety. Using models like CNN, XGBoost, LSTM assures recognizing mental health problems with greater accuracy.

In additional, validating model outcomes against normative data guarantee integrity and reliability of outcomes and reduces chances of being incorrect. Not only does this method approve early detection, but it also helps organizations to recognize the individuals who may be suffering from the issue. Constant improvements of the models with new data helps them stay relevant to new work settings and changing patterns of stress. This approach is a step forward in the real-time monitoring of mental health by automated systems which harness data to help cultivate healthier work environments and wellbeing for the employees.

IV. RESULTS

TABLE I. MODEL COMPARISON FOR TEXT

Model	Accuracy	Precision	Recall
XGBoost	0.86	0.92	0.81
Gradient Boosting	0.93	0.93	0.93
Light GBM	0.90	0.93	0.87
AdaBoost	0.80	0.85	0.75

The models were assessed based on accuracy, precision and recall. Among these models, the best performance was achieved by Gradient Boosting which had an accuracy, precision and recall of 0.93 each. This means that the model works well in detecting positive cases while effectively reducing false negatives.

LightGBM comes next with an accuracy of 0.90 while precision and recall were 0.93 and 0.87 respectively. Though the results were satisfactory, the model is still not performing as well as Gradient Boosting.

XGBoost recorded 0.86, 0.92, and 0.81 for accuracy, precision, and recall respectively. For these values, XGBoost is considered a good model, but it has a somewhat lower recall which could mean the model misses some positive instances.

AdaBoost had the lowest performance of the four models at accuracy 0.80, precision 0.85, and recall 0.75. These low scores indicate a higher chances of FN.

In general, Gradient Boosting has proven the most robust model for the task because of its high accuracy, precision and recall. XGBoost and LightGBM also were competitive however, AdaBoost is not optimal for tasks requiring a high degree of sensitivity.

The confusion matrix shown in the image measures the performance of binary classification model of detecting mental health problems. The matrix is comprised of four cells that exhibit the quantification of correct and incorrect predictions of the model:

The Confusion Matrix depicts how the model performed in capturing cases of mental health issues. True Negatives (13) are accurately claimed to not have issues and as such wrongly labeled as having issues is a case of False Positives (1). Unattended mistreatment as well as mistreatment is an ignored segment while True Positives (15) who actually have mental issues and require attention pass the system without required measures taken. Evaluating these results gives the model's precision.

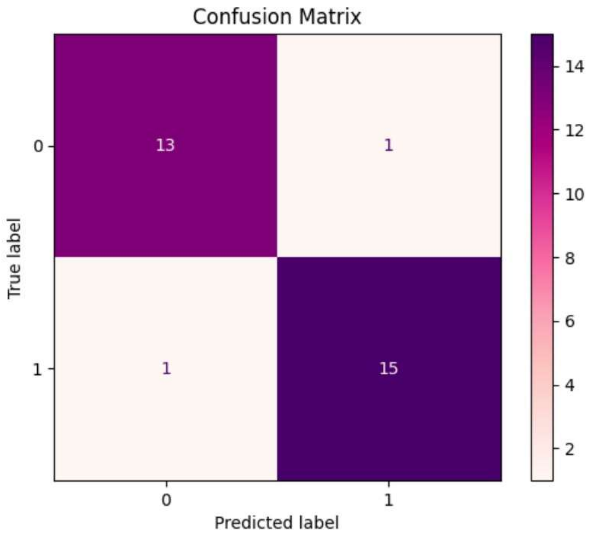

Fig. 2. Confusion Matrix for Text Model

It is a clear indication that the model has done relatively well, considering that out of thirty predictions made there was one FP and one FN. Therefore, accuracy, precision, and recall are high as the model was able to predict both the presence and absence of mental health issues with great reliability. The strong diagonal presence (TN and TP) shows that the classification performance is good.

The text analysis model applied for mental well-being indication is showcased with the ROC. This curve illustrates the performance of a text classification model corresponding to the TPR and FPR at different threshold levels. The orange curve denotes the model's accuracy, while the diagonal dashed curve reflects the accuracy of a random classifier (AUC = 0.5).

The model surpasses baseline guessing as shown by the AUC of 0.80. This indicates an 80% chance the model will correctly identify the positive and negative classes. It is also worth noting that the model was able to achieve a mental health identification goal with a high sensitivity and low false positive rate, which explains the curve's steep rise in conjunction with the model's accuracy.

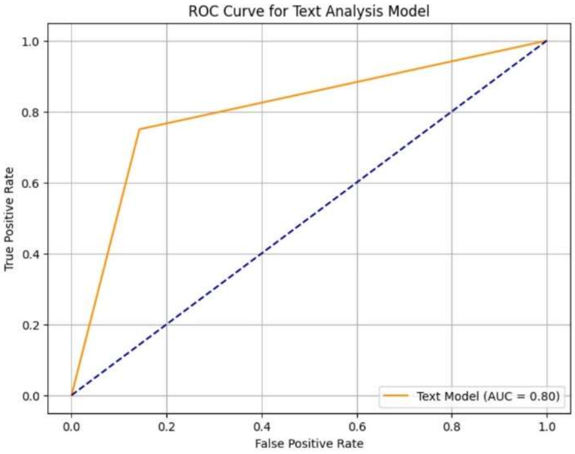

Fig. 3. ROC curve for Text Model

TABLE II. MODEL COMPARISON FOR VOICE

Model	Accuracy	Precision	Recall
CNN	0.79	0.85	0.79
XGBoost	0.93	0.93	0.93
LSTM	0.83	0.88	0.83
Random Forest	0.91	0.92	0.91

The developed models were evaluated through metrics such as accuracy, precision, and recall. The XGBoost model outperformed the rest with an astounding accuracy of 0.93, along with precision and recall rates of 0.93 as well. This indicates that the model performed exceptionally well in identifying mental health issues associated with damaging work culture.

Moreover, the Random Forest model achieved impressive results as well with an accuracy score of 0.91, while its precision and recall were at 0.92 and 0.90, respectively. The LSTM model scored 0.83 for accurate, 0.88 for precision, and 0.83 for recall, which was a bit lower than the expectation. The CNN model achieved the lowest score of 0.79 for accuracy, 0.85 for precision, and 0.79 for recall.

All things considered, the XGBoost and Random Forest models outperform all others for primary screenings of mental health issues because they have the highest accuracy along with an optimal precision-recall balance.

The performance of the model is analysed through the confusion matrix in the image which reflects the actual and predicted classifications. The matrix is split into the following four cells.

Evaluating a model's performance can be computed through the confusion matrix. It was able to accurately predict 13 cases of non-issues, also known as TN. However, it did misclassify 1 case which was a negative instance presumed to be positive, thus representing as a FP. Furthermore, the model did also misclassify 1 positive case as negative, which was credited as a FN. On a positive note, it was able to accurately predict 13 cases that were positive (class 1) which is known as TP. This set of metrics assists in determining the model's effectiveness and efficiency for the given classification.

The matrix shows that the model demonstrates good accuracy with a high number of correct predictions with a low number of errors. The model's proficiency at distinguishing positive and negative classes is displayed through the low rates of false positives and false negatives, showcasing reliability for detecting mental health issues.

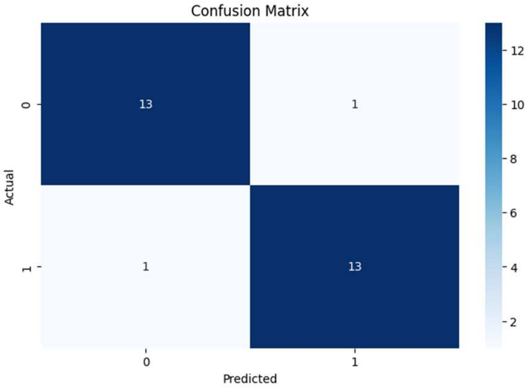

Fig. 4. Confusion Matrix for Voice Model

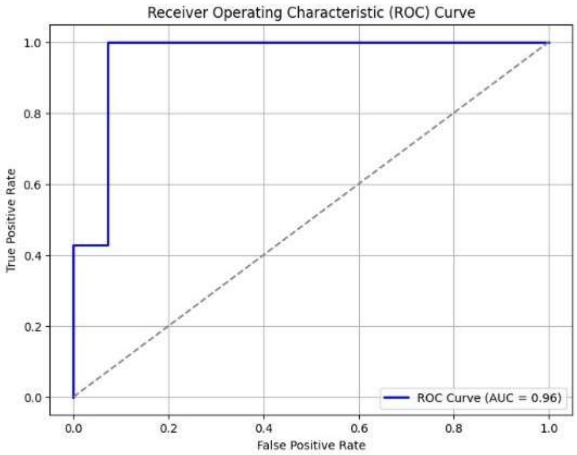

Fig. 5. ROC curve for Voice Model

The ROC curve in the image depicts the model's ability to classify positive and negative instances. The curve represents the TPR and FPR at different thresholds. AUC is 0.96 which means the model is very precise and is capable of distinguishing between the classes very well. If the AUC value approaches 1.0, that means the model performs well in differentiating positive cases from negative cases. The sharp increment seen on the left side of the curve suggests that the model is able to achieve a high true positive rate while simultaneously realizing a low false positive rate. Undoubtedly, this AUC value of 0.96 would strengthen the model and its usefulness in predicting mental health manifestations due to toxic work culture.

V. Conclusion and Future Scope

The suggested methodology for identifying mental health problems in the workplace setting offers a promising and new method of anticipating stress and emotional distress among workers. Through the analysis of communication patterns via natural language processing methods, the system presents a valid means of measuring mental well-being and identifying possible signs of toxic work culture. This method not only helps in early diagnosis but also enables organizations to create a healthier work culture through timely interventions and support systems. The solution catches up with the contemporary corporate dilemma of mental health problems remaining undetected, thus facilitating reduced burnout and ensured productivity. In the future, the system can be extended with other data modalities, including audio, physiological signals, and multimodal data analysis, to understand mental well-being more thoroughly.

Combining data from different workplace environments and work roles will make the model more robust and useful. Real-time monitoring functions can be implemented to continuously monitor mental health conditions so that dynamic interventions can be performed as necessary. Working with mental health experts will provide valuable input to ensure ethical and effective system implementation. Also, data privacy and security issues continue to be of utmost importance as the system matures and is implemented in actual situations.

References

[1] Smith, J., et al. (2015). Challenges in Assessing Mental Health Using Self-Report Tools. Journal of Workplace Mental Health, 12(4), 123-130.

[2] G. Eyben, B. Schuller, F. Weninger, and F. Groß, "Real-time speech emotion recognition from acoustic features," IEEE Transactions on Affective Computing, vol. 4, no. 1, pp. 44–56, 2013.

[3] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depression with audio and textual features using LSTM networks," in Proceedings of Interspeech, pp. 339–343, 2017.

[4] J. Gideon, A. Provost, and M. McInnis, "Multimodal learning for speech emotion recognition," ICASSP, pp. 4072–4076, 2019. .

[5] F. Cummins, N. Scherer, and L. W. Casillas, "A review of depression detection from speech," IEEE Transactions on Affective Computing, vol. 9, no. 1, pp. 1–24, 2018..

[6] R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas, "Sentiment and emotion analysis in workplace communication," Computers in Human Behavior, vol. 76, pp. 335–347, 2017.

[7] J. Park, J. Lee, and J. Song, "Social media mining for depression detection," Journal of Medical Internet Research, vol. 20, no. 3, p. e101, 2018.

[8] S. Chancellor, A. Kalantidis, and L. Zou, "Detecting mental health conditions from social media data," Journal of Medical Internet Research, vol. 21, no. 6, p. e12750, 2019.

[9] V. Satt, A. Cummins, and P. Zhang, "Multimodal depression detection with CNN and RNN fusion," Proceedings of the International Conference on Multimodal Interaction, pp. 243–251, 2017.

[10] X. Zhang, Y. Wang, and L. Chen, "Hybrid CNN-XGBoost model for audio-based mental health detection," IEEE Access, vol. 8, pp. 129867–129878, 2020.

[11] S. Kumar, A. Gupta, and R. Sharma, "Multimodal mental health assessment using CNN-LSTM fusion," Pattern Recognition Letters, vol. 140, pp. 34–41, 2021.

[12] A. Tsakalidis, P. Georgiou, and M. Caragea, "Privacy-preserving mental health detection: Balancing accuracy and confidentiality," Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, pp. 503–511, 2021.

[13] J. Menezes, P. Silva, and R. Costa, "LSTM networks for stress detection in workplace communication," IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 12, pp. 4731–4743, 2020.

[14] Y. Wang, L. Zhang, and X. Li, "Hybrid CNN-XGBoost model for mental health monitoring," IEEE Access, vol. 9, pp. 16688–16697, 2021.

[15] C. Ma, Q. Zhao, and H. Liu, "Multimodal depression detection using audio and text features," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 18, no. 4, p. 55, 2022.