Leveraging Audio and Physiological Data to Predict Job Performance in Healthcare Settings

Abdul Mannan Kanji
akanji@usc.edu

Maksim Siniukov
siniukov@usc.edu

Mirmahdi Seyedrezaei
seyedrez@usc.edu

Jiachen Zhang
jzhang70@usc.edu

1 Problem Definition

The evolution of today's work environments, particularly in healthcare requires refined methods to assess and predict job performance. As healthcare continues to expand its service range from rehabilitation nursing to intensive care, the consistency of service quality across all domains becomes crucial [1]. Thus, in these high-stress environments, accurate evaluation of job performance is essential not only for maintaining the quality of patient care but also for ensuring the safety of both staff and patients [2].

However, measuring job performance in healthcare settings poses unique challenges. Traditional metrics such as task precision and error rates may not fully capture the subtleties of performance deterioration due to the internal and external complexities involved. Moreover, self-assessments are often biased, as medical workers may not always recognize their own errors or lapses in performance. This complexity necessitates a robust data-driven approach to uncover these relationships more accurately [2] .

Capturing the intricate dynamics between an employee's physiological conditions, behavioral patterns, environmental stressors, and their combined impact on workplace outcomes, such as performance. The TILES dataset, with its extensive collection of physiological, behavioral, and audio data from intensive care units, provides a unique foundation for developing predictive models of job performance in stressful environments and, more particularly, intensive care units [3].

This project aims to leverage the TILES-2019 dataset to create a predictive model that uses physiological and audio data to assess job performance among medical workers. By synthesizing data across multiple sensors, instantaneously processed audio data, physiological metrics such as heartbeat and sleep quality, and self-reported evaluations, we aim to establish a predictive framework that can serve as a reliable indicator of job performance in healthcare settings. This endeavor not only addresses the immediate needs of healthcare organizations to maintain high standards of patient care but also contributes to the broader understanding of how individual and environmental factors interact to affect performance in high-stress work environments.

2 Literature Review

In this section, we explored the trends of making use of physiological data and audio data to predict job performance. At the same time, we tried to prove the theoretical foundations of the availability of conducting predictive analysis between predictor variables (physiological and audio data), and target variable (job performance).

The exploration of audio and physiological data has become a significant area of research, providing deep insights into human emotions and behaviors in various settings. A foundational study published in 2004 explored the potential of audio features within a composite emotion classification system in multimodal contexts [4]. The findings suggested that both global features such as intensity and pitch, alongside time variation-based features like spectrograms and MFCCs, are effective in predicting human emotions, which in turn may influence job performance. Despite the limitations of the pre-deep learning era, the application of classical machine learning techniques, including K-nearest neighbors and kernel regression on static audio features, as well as Hidden Markov Models (HMM) on dynamic features, demonstrated considerable accuracy. This research laid the groundwork for utilizing processed rather than raw audio data in analyzing job performance.

Advancements in audio data collection techniques have also been notable. A recent study introduced innovative methods featuring Voice Activity

Detection (VAD) and instant feature generation, marking a significant improvement over earlier systems such as the Environmental Audio Recorder (EAR), which collected speech at constant intervals [5]. The new VAD system includes a passive human voice detection module capable of operating effectively in diverse acoustic environments. Following a minimal delay, this system extracts global audio features from raw signals and then discards the raw data, retaining only the valuable processed information. This approach aligns seamlessly with frameworks that utilize global audio features for machine learning models.

The TILES dataset has been instrumental in advancing research on multimodal data to understand workplace dynamics. A notable study within this dataset examined the potential of non-linear complexity measures derived from respiratory activity to monitor stress and anxiety in workplace settings [6]. This research highlighted the predictive power of physiological signals regarding mental states that directly impact job performance and satisfaction. By demonstrating that breathing patterns can effectively indicate stress levels among hospital workers, this study validated the critical role of physiological data in occupational research and facilitated the development of real-time monitoring tools aimed at enhancing employee well-being and productivity.

Moreover, the openSMILE toolkit has emerged as a crucial development for extracting audio features from the TILES dataset, enabling detailed analysis of both verbal and non-verbal cues in workplace communication [7]. This toolkit's ability to handle complex audio data underscores the significance of acoustic signals in revealing the emotional and cognitive states of employees. Research utilizing openSMILE has shown that speech patterns and vocal expressions are potent markers of emotional states, enhancing our understanding of how interpersonal communication within the workplace influences team performance and individual well-being and job performance.

Physiological data, encompassing various body signals, serves as a critical indicator of stress and consequently, job performance. A study from 2008 highlighted that short-term heart rate variability is a significant marker of health status, suggesting potential indirect links between stressful work conditions and heart rate variations [8]. Building on this foundation, subsequent research has integrated multimodal features to enhance affective prediction tasks. For instance, a social-interaction system study not only leveraged audio data but also incorporated physical behaviors like movement and gestures, employing embedding techniques to transform sparse features into denser, more meaningful predictors [9].

Further advancing multimodal affective analysis, another pivotal study applied Gaussian Mixture Models (GMM) to improve prediction accuracy by utilizing covariance matrices between features. This study methodically partitioned multimodal and emotional data into regular time frames to monitor human behaviors continuously, illustrating a novel approach to fusing multimodal information at the model level [10]. As the deep learning paradigm gained traction, numerous studies began exploring the relationships between emotions and multimodal physiological data. For example, Tao et al. utilized a cross-modal LSTM network to analyze EEG, eye movement, and differential entropy features to predict emotional states under normal and sleep-deprivation scenarios, revealing that sleep deprivation markedly reduces positive emotions which can affect job performance [11].

Moreover, a Harvard neuroscience study demonstrated that heart rate time series, processed through spectral and higher-order statistics (HOS) feature extraction, could effectively identify personalized emotional patterns when analyzed with support vector machine (SVM) models [12]. Recently, deep learning techniques have been increasingly applied to analyze physiological signals directly; for instance, Harper's work employed BiLSTM and CNN to process heartbeat time series data, enhancing predictive accuracy through a Bayesian framework integrated into neural networks, thus marking a significant step forward in end-to-end physiological data analysis for emotion recognition [13].

Furthermore, the use of wearable sensor data in predicting job performance has proven the viability of employing real-time physiological and environmental data to enhance workplace outcomes. Research involving the TILES dataset has demonstrated how variables such as heart rate variability, skin temperature, and environmental noise levels can be indicative of stress, cognitive load, and overall job satisfaction [6].

The cumulative body of research underscores the transformative impact of multimodal data in creating healthier, more productive workplace environments, thereby illustrating the pivotal role of the TILES dataset in advancing occupational health and organizational psychology. These studies have clearly delineated the utility of physiological data in job performance analysis. Key takeaways include the close relationship between audio and physiological features with human emotions, the effectiveness of multimodal fusion as a method to understand interactions between different data modalities, and the superiority of complex modeling techniques over simpler ones. The use of deep neural networks, sequential modeling, and model-level fusions are highlighted as particularly effective in handling the intricate, often non-linear relationships inherent in multimodal data. Our research methodologies and experiments have been guided by these insights, aiming to develop robust models for emotion recognition and the prediction of job performance-related factors.

3 Data

For our research, we utilized the TILES-2019 dataset, a detailed compilation of physiological and behavioral information collected from 57 medical residents across a three-week period [3]. This dataset, of a medium size, includes two primary types of data: subjective assessments from personal surveys and objective measurements from a variety of sensors. The physiological aspects were mainly recorded using Fitbit devices, shedding light on various health metrics, while the collection of audio data through a dedicated audio-feature recorder provided further insights into behavioral patterns. Additionally, data from apps like RescueTime, Owl-in-one, and TIS added more depth to our analysis, offering details on phone usage, signal strength, and social interactions.

The TILES-2019 dataset divides its survey data into two main types: baseline and periodic surveys. The baseline survey captures each participant's initial mental health state, assessing stress levels, burnout, depression, anxiety, and other relevant factors. Subsequent periodic surveys, conducted mid-day, end-of-day, and end-of-week, track changes over time. The mid-day survey focuses on stress levels, while the end-of-day survey collects comprehensive data including stress, the work environment, self-assessed job performance, job satisfaction, and sleep quality. The end-of-week survey extends to cover broader aspects such as stressors, tobacco and alcohol use, and social interactions. When diving into the survey data, particularly looking at the job performance indicator used as the target variable for our predictive models, we noted some instances of missing data, though these were relatively minimal. It's also important to point out that the data collection wasn't continuous. There were moments when participants might miss some workdays, introducing gaps in the dataset. Over the three-week survey period, this typically resulted in obtaining 10 to 15 data points for each participant, leading to varying sample sizes across the group of participants.

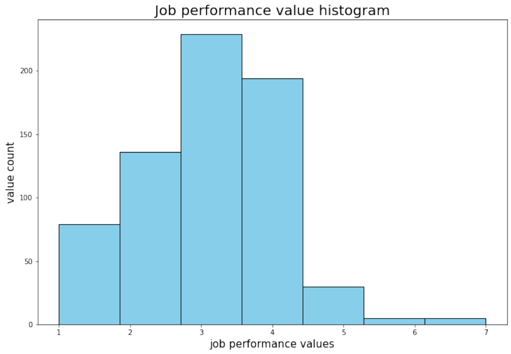

Figure 1: Histogram of job performance value

Count	Mean	Std	Min	Median	Max
678	2.99	1.15	1.00	3.00	7.00

Table 1: Summary statistics of job performance value

The Fitbit data within the TILES-2019 dataset provides detailed insights through daily summaries, heart rate, step counts, and sleep metrics for each participant over a three-week period. These daily summaries include comprehensive physiological parameters such as step counts, calories burned in various activities (cardio and fat burn), sleep quality, sleep duration across stages (deep and light), and resting heart rate, compiled into a single table covering roughly 20 days. However, this subset of data poses some challenges: it includes records for only 50 participants, not the entire volunteer group, and some of the sleep data is missing. To address missing sleep data, interpolation techniques are suggested, though the issue of incomplete participant records remains complex.

In addition, the dataset organizes physiological data into three distinct tables—heart rate, step count, and sleep data. Heart rate data, recorded at

Feature group	Feature names
Waveform	Zero Crossing
Signal Energy	Root Mean Square, log
Loudness	Intensity, approx loudness
Fourier Spectrum	Phase, magnitude(dB)
ACF, Cepstrum	Autocorrelation, Cepstrum
Pitch	F0, voice prob
Voice quality	HNR, Jitter, Shimmer

Table 2: Feature groups and their core features

intervals of 5, 10, or 15 seconds during the workday, does not uniformly align with other data types, prompting the use of average pooling to streamline these measurements into daily summaries. Similarly, sleep and step count data were detailed but have been simplified to enhance consistency. While sleep data was categorized by phases such as deep and light, the analysis focused on broader daily summaries rather than phase-specific details. Step count data was recorded every minute, but since aggregated step counts were included in daily summaries, minute-by-minute logs were bypassed to prevent data redundancy and support streamlined analysis.

The audio data within our project utilizes the TAR framework as outlined in [5], where each CSV file captures auto-detected human speech for an individual tester over a single day. The recordings were made in a highly granular 10ms timeframe, resulting in substantial data volumes. Feature extraction was performed using two packages: TarsosDSP [14] and OpenSMILE [7]. TarsosDSP processes the audio to provide features such as pitch estimation, time-stretch, resampling, filters, pitch shifting, audio effects, and Q-transform. OpenSMILE, on the other hand, offers a broader range of advanced features, including cepstral, waveform, spectrum, loudness, voice quality, signal energy, and various pitch-related attributes. The dataset includes a total of 29 audio feature columns, with a detailed breakdown provided in the subsequent table. Based on our literature review, it was likely that these features hold significant statistical value for predicting job performance.

Audio features in our project were initially sampled in a highly granular 10ms timeframe, necessitating a pooling process along the time dimension to align with the daily collection of job performance data. For this, we calculated the mean, standard deviation, and maximum of audio features within a 3-hour window prior to each self-assessment timestamp, ultimately transforming each raw audio feature into three new features, leading to a total of 87 audio features. We hypothesize that vocal features proximal to the self-evaluation are more indicative of stress and, consequently, job performance.

In terms of preprocessing, we undertook several steps to prepare the data for analysis. Initially, we merged separate data tables and addressed missing values. After pooling, all data fields were aligned with time intervals of one day, and we concatenated the pooled data within the same field (physiological, EMA, audio) using (id, date) as the primary key. The subsequent joining of these tables creates an early fused dataset, encompassing features across the three modalities, resulting in a table dimension of 1730 by 171.

To address data shortcomings, we implemented three main strategies: First, we removed records with incomplete data, such as those missing physiological or audio data, and eliminated records lacking the target variable. Second, we handled missing values by dropping columns predominantly null (e.g., 'Sleep3Efficiency') and imputing others using median values or categorizing missing values as 'missing'. Lastly, we transformed categorical features according to model requirements; models like Catboost or LightGBM treat categorical attributes differently than models like Random Forest or Linear Regression, where one-hot encoding is applied to integrate these features effectively.

4 Method

We experimented on multiple machine leaning models to capture the relationships between features from different modalities. The 3 modalities are **Survey, Audio and Physiology**. We conducted identical experiments on all of the following modality settings: (1) Physiology only. (2) Audio only. (3) Survey only. (4) Physiology and Survey. (5) All modalities.

The primary objective of these model experiments was to assess the efficacy of multimodal fusion in enhancing predictive performance. We primarily utilized early fusion, where features from different modalities are concatenated prior to being fed into machine learning models, allowing the models to discern complex inter-feature relationships.

Our model selection ranges from the simplest

Linear regression models, to deep neural networks. All the models come with proper feature engineering and hyperparameter tuning. The details of application of each model are listed as follows:

- **Random Forest:** This model uses a bagging strategy with each decision tree trained on randomly sampled features and data. Predictions were aggregated via majority voting. For feature engineering, we imputed 'missing' for categorical and median values for continuous features and applied one-hot encoding to categorical data. The model was configured with 100 decision trees and default pruning settings, making Random Forest a strong baseline for non-linear methods due to its generalizability.
- **XGBoost:** This model trains decision-tree-regressors sequentially on second-order approximation residuals, using MSE for parameter learning. XGBoost automatically handles categorical features and NA values, performing one-hot encoding and treating NAs as a separate category, respectively. To prevent overfitting, especially in smaller samples, L1 regularization was applied to limit node numbers and control weight variance. Parameters were set to utilize 60% of features per tree, with a 10-unit L1 regularization coefficient and a 0.1 learning rate.
- **Catboost:** Catboost extends the GBDT framework, adeptly processing categorical features by automatically applying one-hot encoding to low-cardinality categories and target-statistic transformations for high-cardinality ones. It uses a greedy bundling algorithm for variable splitting at each node and leverages first-order residuals with unbiased gradient estimation and ordered boosting to improve precision. No feature engineering was performed, and default parameters were used. The model focuses on key categorical variables: survey type (mid-day or end of day) and timestamp, treating other continuous variables similarly to GBDT. Catboost is particularly effective in small sample-size settings, optimizing accuracy through its advanced handling of non-linear relationships and gradient estimation.
- **LightGBM:** LightGBM improves upon XGBoost by using quantization to efficiently calculate data distribution, thereby reducing complexity and enhancing split accuracy. It adopts leaf-wise learning instead of level-wise, enabling it to capture more complex relationships. One-side sampling focuses training on samples with higher gradients, improving model effectiveness. LightGBM also employs a greedy bundling algorithm to expand feature space but introduces a risk of overfitting in small sample scenarios. To mitigate this, the maximum number of leaves per tree was capped at 31. LightGBM handles categorical features automatically with one-hot encoding. The model settings included 100 estimators and a learning rate of 0.01.
- **Linear and Lasso regression:** These models serve as our baseline and test whether complex non-linear models might overfit the predictor-target relationships. Linear regression also functions as a late fusion method, helping evaluate the significance of interactions between features from different modalities. For these experiments, we applied one-hot encoding to categorical features and set the L1-regularization penalty parameter to 0.1. Based on our literature review, we did not anticipate high performance from these models, as the relationships between predictors and the target were likely complex.
- **SVM Regressor:** SVM utilizes hinge loss to identify the optimal classification hyperplane and can incorporate non-linear kernel functions from the dual side. We applied one-hot encoding to categorical columns and set the default L2-penalty parameter to 1. However, SVM's performance is highly sensitive to the C parameter settings. Without prior feature selection or sample weighting, SVM is susceptible to both overfitting and underfitting, particularly in scenarios with small sample sizes.

To conclude, our model selection explored several directions: (1) extending the existing feature space, by bundling, smoothing and non-linear projection methods. (2) Switching to model representations and optimization techniques that are freer and more robust. (3) Improving the estimation of loss of true distribution by weighting samples and adopting unbiased mechanisms. (4) Focusing on the generalizability of models, using bagging, random sampling of features and data points, or regularization methods to get to an improvement.

For evaluation, we utilize MSE to assess the accuracy of regression predictions, R-squared to gauge the proportion of variance our model captures, and the Concordance Correlation Coefficient, which ranges from -1 to 1, to measure agreement similar to inter-rater reliability. The calculation function is as follows:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

5 Results

In this section, we will firstly display the comparison of different models in 2-modalities setting. We aim to find out which model class is more appropriate to be applied in this healthcare prediction settings. Also, we will try to simply discuss the phenomena we faced during training and our observations.

The validation result of all models using different modalities is shown in table 3.

The table shows that CatBoost has best performance over all models in unimodal settings. And in general, ensemble methods perform better than other models. During our observation of the learning process, deep neural networks undoubtedly overfits the data. Because the training error keeps reducing, but validation error oscillated in a very high value range, we concluded that deep learning models can hardly capture the relationships in this small sample settings. So does SVM, which got a really high error. Therefore, we focus our study of modality fusion effects on ensemble methods. As we had expected before, the Random Forest is a decent baseline, as it generalized better and learned the non-linear relationships. XGBoost, LightGBM, CatBoost tried to improve GBDT framework in different ways. In unimodal settings, XGBoost is not a good one, we believe it is because we regularized the model too much, as it has a much better CCC than randomized model that has 0 CCC value, but has very little R2 score. LightGBM and Catboost focused more on improving model complexities, aligning well with the need of Audio features mining. Therefore, although XGBoost is not good in this scenario, we believe it has potentials to perform better in multimodal settings.

Then we use Line Chart to depict the performance(MSE) variation in different modality settings mentioned in Section 4.

Through this figure, we can conclude that Survey-only prediction is the best unimodal settings. And Random Forest model in this setting is even better performed than all of the models in multimodal settings. Even if using all modalities, there is only slight improvement in the MSE error.

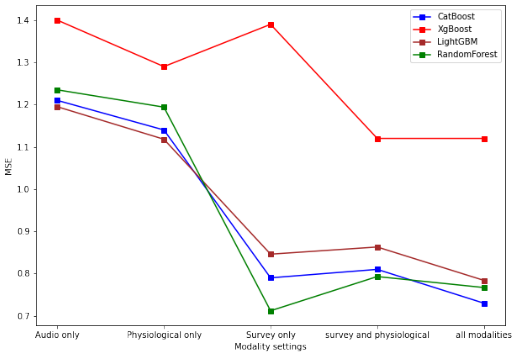

Figure 2: Multimodal fusion effect

6 Team members' contributions

In our project, each team member contributed to the project. Abdul Mannan Kanji focused on audio feature pre-processing, extraction, model experimentation, multimodal analysis while contributing to our midterm, final presentation and reporting results. Jiachen Zhang managed the extraction and integration of physiological data from Fitbit devices and led the significant portion of the final report. Mirmahdi Seyedrezaei conducted the literature review, statistical feature exploration, and was responsible for data preprocessing and analysis for the preliminary dataset, alongside contributing to the midterm and final presentation and final report. Maksim Siniukov developed and trained our predictive models, contributed to the literature review, and prepared the midterm report. Each member's specialized contributions were crucial to the project's success, enabling a thorough exploration of multimodal data to predict job performance.

References

[1] M. L'Hommedieu et al. Lessons learned: Recommendations for implementing a longitudinal study using wearable and environmental sensors in a health care organization. *JMIR Mhealth Uhealth*, 7(12):e13305, Dec 2019.

[2] Joanna C. Yau et al. Tiles-2019: A longitudinal physiologic and behavioral data set of medical residents in an intensive care unit. *Scientific Data*, 9(1):536, Sep 2022.

Model	All modalities			Survey			Audio			Physiology			Survey & Physio		
	MSE	R2	CCC	MSE	R2	CCC	MSE	R2	CCC	MSE	R2	CCC	MSE	R2	CCC
Random Forest	0.767	0.46	0.585	**0.712**	**0.50**	**0.685**	1.235	0.13	0.202	1.194	0.16	0.274	**0.793**	**0.44**	**0.596**
SVR	1.381	0.03	0.038	0.772	0.46	0.617	1.381	0.03	0.038	1.426	0.00	0.007	1.426	0.00	0.007
Lasso	1.341	0.06	0.190	1.441	-0.01	0.000	1.307	0.08	0.169	1.458	-0.02	0.015	1.458	-0.02	0.015
LightGBM	0.784	0.45	0.633	0.846	0.41	0.580	1.195	**0.16**	**0.256**	**1.118**	**0.22**	**0.337**	0.863	0.39	0.600
XGBoost	1.12	0.22	0.33	1.39	0.03	0.25	1.40	0.02	0.04	1.29	0.09	0.23	1.12	0.22	0.34
Catboost	**0.73**	**0.48**	**0.67**	0.79	0.45	0.64	**1.21**	0.15	0.25	1.14	0.20	0.33	0.81	0.43	0.43

Table 3: Comparison of MSE, R2 and CCC evaluation metrics across modality combinations

[3] Tiles: Tracking individual performance with sensors data set. https://tiles-data.isi.edu/dataset2019_details, 2024. Accessed: May 06, 2024.

[4] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211, 2004.

[5] Tiantian Feng, Amrutha Nadarajan, Colin Vaz, Brandon Booth, and Shrikanth Narayanan. Tiles audio recorder: an unobtrusive wearable solution to track audio activity. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, pages 33–38, 2018.

[6] A. Tiwari, S. Narayanan, and T. H. Falk. Breathing rate complexity features for 'in-the-wild' stress and anxiety measurement. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, Sep 2019.

[7] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, New York, NY, USA, Oct 2010. Association for Computing Machinery.

[8] Mirka Hintsanen, Marko Elovainio, Sampsa Puttonen, Mika Kivimäki, Tuomas Koskinen, Olli T Raitakari, and Liisa Keltikangas-Järvinen. Effort—reward imbalance, heart rate, and heart rate variability: the cardiovascular risk in young finns study. *International journal of behavioral medicine*, 14:202–212, 2007.

[9] Giovanna Varni, Gualtiero Volpe, and Antonio Camurri. A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. *IEEE Transactions on Multimedia*, 12(6):576–590, 2010.

[10] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2):137–152, 2013.

[11] Le-Yan Tao and Bao-Liang Lu. Emotion recognition under sleep deprivation using a multimodal residual lstm network. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[12] Gaetano Valenza, Luca Citi, Antonio Lanatá, Enzo Pasquale Scilingo, and Riccardo Barbieri. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. *Scientific reports*, 4(1):4998, 2014.

[13] Ross Harper and Joshua Southern. A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. *IEEE transactions on affective computing*, 13(2):985–991, 2020.

[14] Joren Six, Olmo Cornelis, and Marc Leman. Tarsosdsp, a real-time audio processing framework in java. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

[15] T. a. i. of stress, "workplace stress," 2023.

[16] Janete de Souza Urbanetto, Priscila Costa da Silva, Eveline Hoffmeister, Bianca Souza de Negri, Bartira Ercília Pinheiro da Costa, and Carlos Eduardo Poli de Figueiredo. Workplace stress in nursing workers from an emergency hospital: Job stress scale analysis. *Revista latino-americana de enfermagem*, 19:1122–1131, 2011.

[17] Elkana Timotius and Gilbert Sterling Octavius. Stress at the workplace and its impacts on productivity: A systematic review from industrial engineering, management, and medical perspective. *Industrial Engineering & Management Systems*, 21(2):192–205, 2022.