Mind Matters - Mental Health in Tech Industry

"Mental health is not a destination but a process; it's about how you drive, not where you're going." - Noam Shpancer, Ph.D.

Zainab Hasnain
A20516879
zhasnain1@hawk.iit.edu

Aditya Sharma
A20516668
asharma55@hawk.iit.edu

Muhammad Ans Qadri
A20521573
mqadri@hawk.iit.edu

Abhishek Jaiswal
A20380004
ajaiswal1@hawk.iit.edu

Abstract

This study focuses on the pressing issue of mental wellness in the tech industry, where the symptoms of mental illness are becoming increasingly pronounced. Using the OSMI survey responses, the research examines the distribution of mental health prevalence and attitudes among employees in the tech industry based on age, gender, and nationality. The study aims to identify the most powerful predictors of mental illness in the workplace and provide insights into effective predictive and inferential analysis techniques to detect and address mental health conditions, promote employee well-being, and encourage timely treatment-seeking. The research analyzes the pre- and post-COVID data to understand the trends in data and identify any changes in perception. The study finds that mental health's interference with work, family history, employer-provided benefits, and gender are strong predictor variables. The Naive Bayes model most accurately classifies whether an employee will seek treatment based on survey responses. Although the data is heavily unbalanced, the research provides valuable insights into the tech industry and suggestions for improving the treatment of employees in the workplace.

***Index Terms*—Mental Health, OSMI Survey, Random Forest, Naive Bayes, Decision Trees, Support Vector Machine (SVM)**

I. Overview

Being techies by heart and profession, we were taken aback by the high number of psychotropic medications, and the use of prescribed drugs, which included medications for mental health disorders such as antidepressants, anti-anxiety agents, and antipsychotics. Mental health disorders, including anxiety, depression, bipolar disorder, schizophrenia, and other mental disorders, are some of the most burdensome health concerns in the United States and across the globe.

According to the 2021 OSMI Mental Health in Tech Survey, over 90% of workers in the tech sector reported having been diagnosed with a mental health disorder. Additionally, nearly 64.7% of respondents noted that their mental health issue had an impact on their productivity. On top of that World Health Organization (WHO) study suggests that - worldwide, approximately 264 million people suffer from depression, which can reduce cognitive performance by about 35%. The WHO conducted a study on mental health in the workplace and found that depression and anxiety disorders alone cost the global economy one trillion U.S. dollars annually in lost productivity.

For a considerable amount of time, the tech culture has received criticism for taking a toll on the mental health of its employees. Burnout is prevalent in the industry, and issues such as isolation and a lack of work-life balance have been exacerbated by the COVID-19 pandemic.

Therefore, is it possible for the culture to evolve and provide improved support for the individuals who shape and constitute it?

We aim to utilize the responses of the OSMI survey to shed light on the factors contributing to mental health disorders among employees in the tech industry. Specifically, we seek to uncover the variations in mental health prevalence and attitudes across different geographical locations, ages, and gender and identify the most powerful predictors of mental illness in the workplace. Ultimately, our research aims to provide insights into the most effective predictive and inferential analysis techniques with the use of various statistical methods for detecting and addressing mental health conditions, promoting employee well-being, and encouraging timely treatment-seeking.

II. Project Motivation & Problem Statement

Through our research and project, we have attempted to address the interrelations between mental health issues prevalent in the tech industry based on several associated features. Our goal was to find effective solutions to alleviate the burden of mental health disorders among tech workers and improve their overall well-being. A few featured problem statements are as follows:

- To what extent do work conditions impact mental health?
- Could additional factors, such as age and family history, provide insight into the connection between workplace stress and mental health?
- How does the prevalence of mental health disorders and attitudes towards mental health vary based on factors such as geographic location, company size, family history, and other related characteristics, according to our data?
- To what extent can a predictive model determine whether an employee will seek mental health treatment?

III. Proposed Methodology/Approach

A. Data Gathering

For this project, we utilize the OSMI survey responses from 2014 as our baseline dataset. In addition to that, we also had access to the survey responses conducted in the following years- 2015, 2016, 2019, 2020, and 2021 (*Note- there was no data available for 2017)*. Our initial goal was to carry out an analysis of the data for each year and compare our findings.

However, to narrow down our problem statement we decided to combine these yearly datasets into two groups. The two groups represent the pre-covid (2014, 2015, and 2016) and post-covid (2019, 2020, and 2021) periods. This approach allowed us to gain insights into the trends and patterns of mental health implications among workers in the tech industry.

B. Data Preprocessing - Pipeline Details

In order to prepare the dataset for analysis, several steps were taken, such as ensuring data consistency by assigning a proper data type to each data point, handling missing values by imputing them with either a fixed value or a statistical measure such as mean, median, or mode, and dealing with invalid entries. Moreover, we employed extensive effort to remove redundant or unnecessary columns.

We also used techniques such as outlier detection and hot encoding to refine the data further and prepare it for analysis.

C. Data Cleaning: Issues and Adjustments

The type of survey questions for each year varied greatly, hence we spent a great deal of time cleaning the data. Here are some of the challenges we came across in this process:

- Each of the years had a different number of columns, hence combining different years' data by merging the columns was a time-consuming task.
- We dropped the unnecessary columns to reduce the dimensionality of the data. For example, we removed the columns regarding the previous workplace conditions, to keep our dataset focused on our problem statement.
- We also had to manually specify the column names as the original file contained long questions as the column names. This process was trivial yet cumbersome.
- Some of the survey questions did not have the correct format for the answer. For example, it seems that the *Gender* column, which should have been a drop-down menu, was a text box.
- Therefore, there were a varying number of responses indicating the same value. It was interesting to see the variety of responses, but since R makes factors of character-type columns, we had to manually specify conditions to filter and assign appropriate values in each of the 27 columns.

D. Summary Statistics and Data Visualization

Since most of our data was categorical, it did not make sense to find statistical summaries like mean values, or variance. Therefore, we resorted to segmenting the data into different rows (grouping into different years) or different columns to analyze the effect of the predictor variables on the response variable.

The methods employed in this study include using bar plots to analyze the relationships between various factors and the response variable, such as *gender*, *family history*, and *remote work*. To observe any trends in a specific group of people, colored bar plots are drawn by label (seek treatment) with categorical variables. We will take a look at some of the plots which display important relationships in the subsequent sections.

Plotting libraries such as ggplot, and treemap was used to visualize the results of exploratory data analysis (EDA) and interpret findings. Additionally, box plots, histograms, heatmaps, and bar charts were used appropriately to map the relationships between predictors such as *country*, *family history*, and *employer provided mental health benefits* with the response of *seek treatment* variable.

E. Distribution Among Different Groups

The responses came from a number of different groups, such as countries, gender, and age groups. To analyze this distribution, we plot the following graphs and gained some valuable insights. To make our visualizations easy to interpret, the color schemes vary for pre-covid and post-covid datasets.

*1) **Pre-Covid Responses**:* We can infer from these graphs that most respondents are between the ages of 20 to 40 and the majority of them are male.

Fig. 1. Gender distribution of the respondents (Pre-Covid)

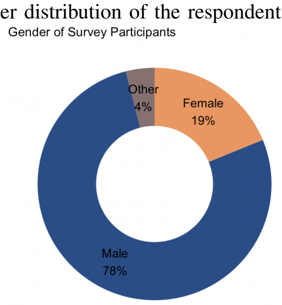


Fig. 2. Age distribution of the respondents (Pre-Covid)

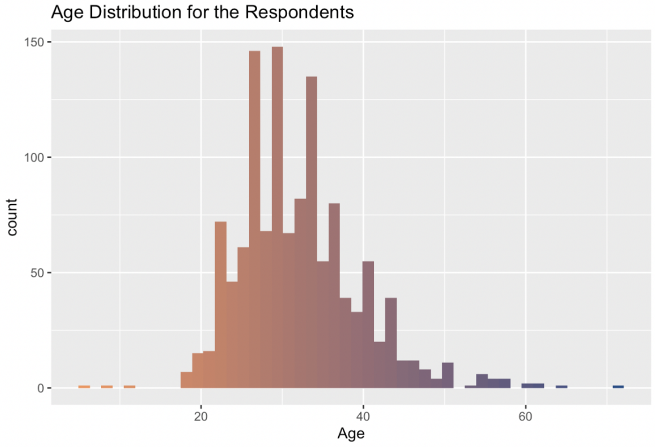
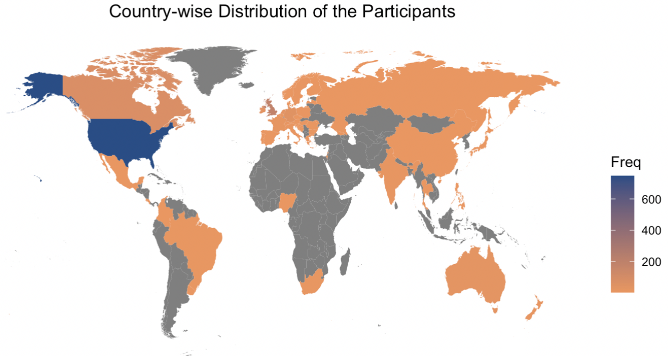


Fig. 3. Country-wide distribution of the respondents (Pre-Covid)



Moreover, the majority of the respondents are from North America as compared to the rest of the world, and, over half of the respondents are specifically from the United States.

*2) **Post-Covid Responses:*** Again, the majority of the respondents are between the age of 20-40 and are male. However, the ratio of female to male respondents has increased from 19% to 30% post-covid. This change in the ratio of gender in the response could be related to increased employment of females in the tech industry post-covid. However, since correlation does not imply causation, no conclusive remark can be added here about the increased number of responses by females as compared to males.

Fig. 4. Gender distribution of the respondents (Post-Covid)

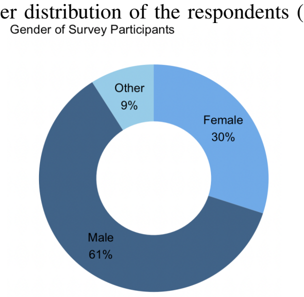


Fig. 5. Age distribution of the respondents (Post-Covid)

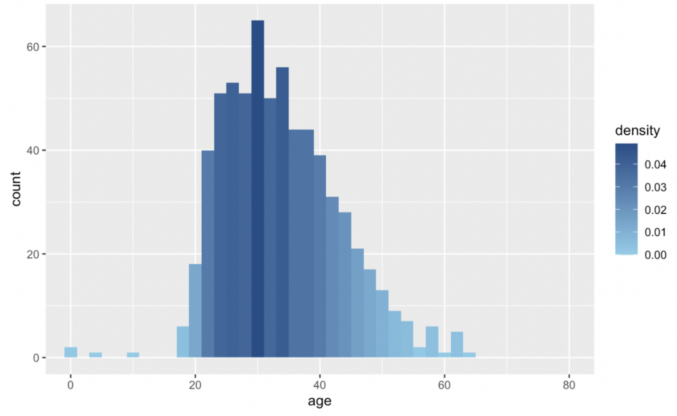
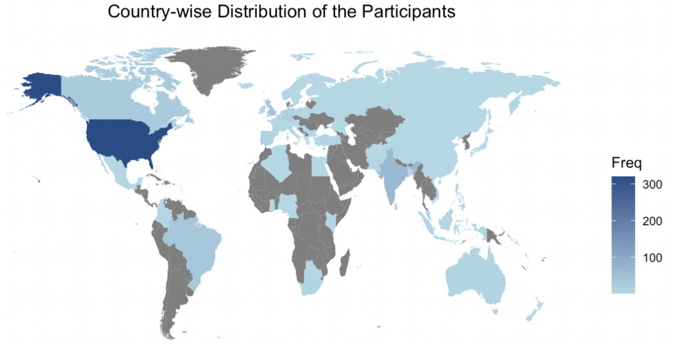


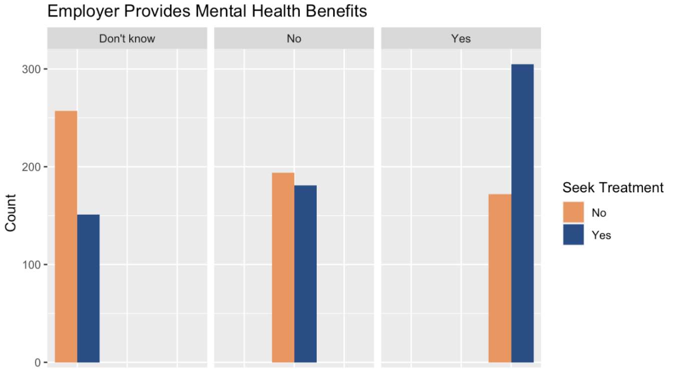
Fig. 6. Country-wide distribution of the respondents (Post-Covid)



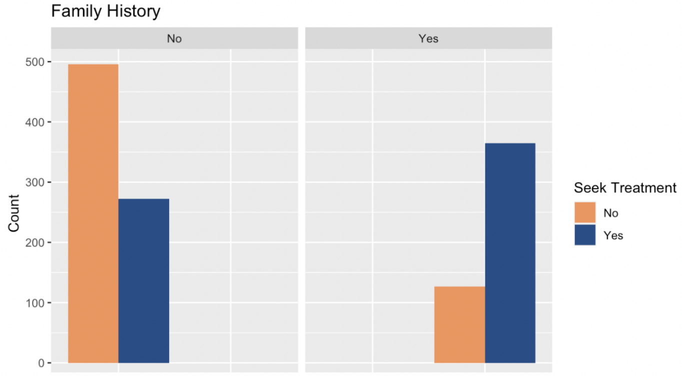
Lastly, the country-wise distribution post-covid does not change significantly, with most of the respondents being from North America, and the United States specifically.

F. Relationship Between Predictors and Response

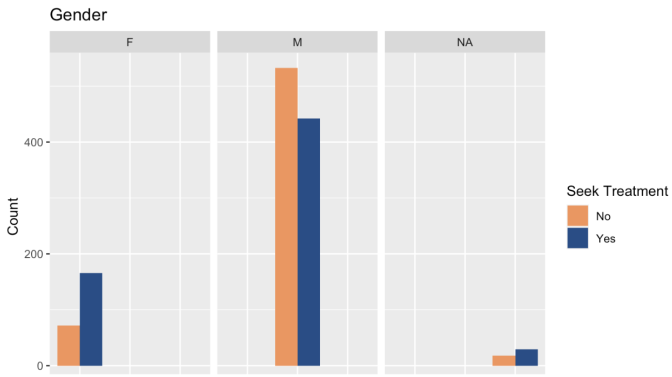
In this section, we will explore some of the relationships between the predictor variables and the response variable, *seek treatment*, by segmenting the data into groups of different values for a particular variable.



From the above bar plot, there is a clear distinction. Employees are more likely to seek for mental health help if their employer provides mental health benefits.



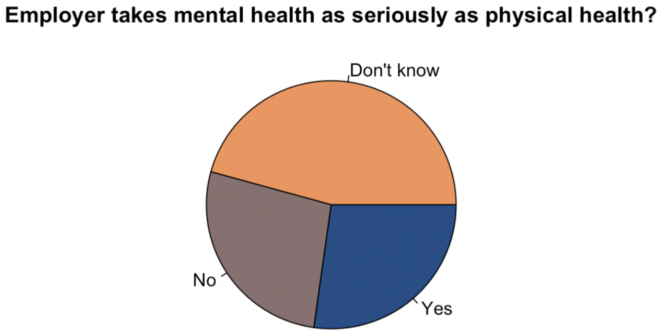
Employees who have a family history of mental health problems and were aware of it are the ones who sought treatment more often than those who did not.



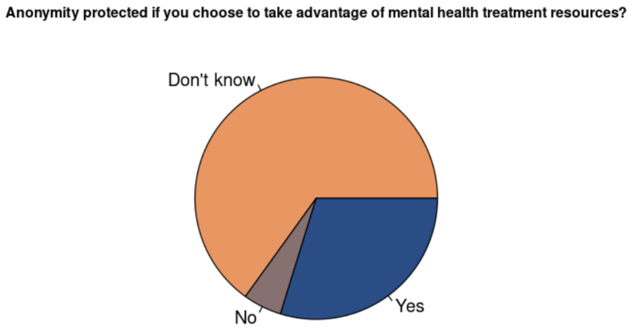
Even though there are fewer female employees in the tech industry, they are more likely to seek for mental health treatment than their male colleagues.

G. *Attitude Towards Mental Health*

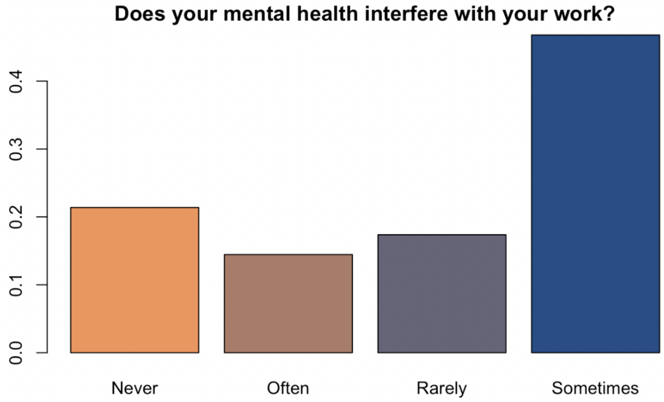
In this section, we will look at the overall behavior of employers and employees towards mental health at the workplace. We do this by summarizing the survey responses for some of the questions, and by visualizing them as a pie chart or a bar plot.



For simplicity, let's consider those who answered *Don't Know* as *No*. It is evident from the pie chart above that only one-fourth of the employees are of belief that their employer takes their mental health as seriously as physical health.



The stigma around mental health exists even in today's day and age and therefore it is important that an employee's anonymity is maintained if they choose to seek help. However, our data shows that the majority of the employees, around 70%, do not have the confidence that their identity will not be disclosed if they choose to take professional help.



We will see in the later sections whether interference with mental health problems affects workplace productivity. However, our data shows that over half the employees who suffer from some kind of mental condition responded to the question "If you have a mental health problem, does it interfere with your work?" as *Often* or *Sometimes*.

H. *Strongest Predictor Variables - Feature Extraction*

We used the aforementioned visualization techniques to analyze the significant predictor variables. However, we needed a more robust technique to identify the strongest predictors. For that purpose, we implemented logistic regression to classify the response variable, *seek treatment*, and picked the statistically significant variables to identify as the strongest predictors.

The most important survey questions, along with the response type, up to 95% significance level (in decreasing order of significance) are given below:

(i) If you have a mental health condition, do you feel that it interferes with your work? *(Never, Rarely, Sometimes, Often)*

(ii) Do you have a family history of mental health problems? *(Yes, No)*

(iii) Does your employer provide mental health benefits? *(Yes, No)*

(iv) Gender *(Male, Female, Other)*

(v) Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources? *(Yes, No, Don't Know)*

(vi) Have you experienced an unsupportive/supportive response to a mental health issue at your workplace? *(Yes, No, Maybe/Not Sure, Yes I observed, Yes I Experienced)*

The results of our p-test seem to be analogous to our intuition. Moreover, they also align with the bar plots shown in the previous section, highlighting how employees in certain groups were more strongly correlated with the seek treatment response variable.

Those employees who thought their mental health was affecting their productivity at work, were the most inclined to seek for mental health help.

Secondly, employees with a family history of mental health problems also actively sought treatment. Lastly, it is interesting to note that the employer's attitudes, such as the provision of mental health benefits, as well as their experience towards a mental health issue at work, and protecting the anonymity of their employee's medical decisions, also make it our list of strong predictors.

I. *Evaluation Metrics*

As our focus is on running classification models for predictive modeling, we have chosen evaluation metrics that are specific to classification tasks. These metrics include *Classification Accuracy*, which determines the ratio of correct predictions to the total number of samples, and the *Confusion Matrix*, which presents a matrix-like output that summarizes the model's performance, including true positive, false negative, and other features used to determine accuracy, precision, and recall.

It is crucial to use multiple evaluation metrics to assess your model's performance. While a model may perform well based on one evaluation metric, it may perform poorly based on another. Therefore, evaluating your model through different metrics is critical to ensure it operates optimally and correctly.

By employing appropriate evaluation metrics, we can effectively gauge the strengths and weaknesses of our classification models, enabling us to make informed decisions and continuously improve our modeling efforts.

J. *Model Selection*

Our focus is on a qualitative response variable (yes/no), and to analyze this data, we have employed classification-based algorithms such as Logistic Regression, Decision Trees, Naive Bayes, and Ensemble Methods, including Random Forest and Support Vector Machines (SVM). These models enable us to evaluate which one provides better performance and gain insights into the relationship between predictors and the response variable.

Our selection of models is based on their complexity and high dimensionality within the data, given the limited computational power available. To ensure comprehensive coverage, our model selection is divided into 'generative' and 'discriminative' models, providing us with a thorough understanding of which models are reliable.

Overall, our implementation of classification-based algorithms provides us with a reliable means of analyzing the data and identifying which model performs the best. This approach allows us to gain insights into the relationship between the predictors and the response variable, enabling us to optimize our models for better performance.

To benchmark and assess the performance of the different classifiers, we have chosen 'Logistic Regression' as the baseline. This choice is based on its simplicity, interpretability, and its ability to facilitate fine-tuning for other models. This baseline model serves as a valuable reference point for comparison with other models, enabling us to determine which model performs better and providing us with the necessary information to optimize our models.

By implementing various classification models and utilizing a reliable benchmark, we can efficiently evaluate our models' performance and make informed decisions to improve our predictive analysis.

IV. Results

A. *Testing Results*

TABLE I
MODEL'S PERFORMANCE ON TESTING SET

Model	**Testing Accuracy in %**
Logistic Regression	50.78
Naive Bayes	71.96
Decision Trees	71.97
Random Forest	77.28
Support Vector Machines (Linear)	51.22

To prepare and assess the model's performance, the dataset was divided into two sets: an 80% training set and a 20% test set. The training set was subsequently employed to train various models, and the model's accuracy was evaluated on the test set.

Our analysis revealed that the **Random Forest** algorithm outperformed the other models. Random Forest, which is an ensemble method that enhances the performance of weak classifiers by combining the results of multiple decision trees, performed better than the decision trees model. Random Forest is particularly useful for datasets with high dimensionality and missing data.

Similarly, Naive Bayes and Decision Trees demonstrated comparable performance, with both models achieving moderate testing accuracy. However, due to the small size of the dataset, both models tended to overfit, which could be why the testing accuracy was not significantly higher.

During our model analysis, we found that our baseline model, which was a 'Logistic Regression' model, did not perform satisfactorily and had low testing accuracy. This was likely due to the limited amount of available data, which may have impacted the model's ability to generalize well. Furthermore, Logistic Regression models are known to be sensitive to outliers and may not perform well on data with non-linear relationships. To address this issue, we could use regularization techniques or alternative metrics to improve the model's performance.

The 'Support Vector Machine(SVM) with a Linear Kernel', which is typically effective for large datasets with high dimensionality, showed signs of overfitting due to the limited data, resulting in lower accuracy. To mitigate this problem, we could use regularization or techniques such as cross-validation to enhance model performance.

B. Performance Criteria

Our primary focus is on the categorical target variable (treatment), and we have employed classification methods for our predictive modeling. The metrics we have chosen are specifically tailored for classification problems. The two primary metrics used are Accuracy (the ratio of correctly predicted values to the total number of observations) and Confusion Matrix, which provides a summary of actual and predicted values. These metrics can be further utilized to evaluate other metrics such as precision, recall, etc.

In our analysis and modeling, we observed that the **Random Forest** algorithm performed better in terms of accuracy compared to the other models we evaluated.

C. Risks: Unbalanced Dataset

When conducting machine learning modeling and analysis, it is essential to consider the potential biases and risks that can arise based on the data and models used. Our analysis has identified a few biases and risks that we need to address to ensure fair and accurate outcomes.

- Our analysis revealed that the majority of our respondents were based in the ”United States”, resulting in a geographical bias that could affect the performance of the models on observations or unseen data outside of the USA.
- During our preliminary analysis, we observed that the majority of the respondents were ’male’, and the models disproportionately affected certain genders. Since gender is a strong predictor and the data is not balanced and underrepresents a particular group, the models had unfair biases towards a particular gender.
- Some of the predictors like ’mental health benefits at work’ and ’diagnosed with a mental health disorder’ had missing values or no response. Other predictors like ’Response of a mental issue at work’ and ’Leave because of a mental health issue’ had many ’Maybe/Not Sure’ responses, which affect the model’s performance and induce bias.
- We observed limited data due to limited responses to the tech survey in some years, leading to biased and underperforming models. To improve the models’ performance and capture important trends and patterns within the data, we need more data to ensure fair and accurate outcomes.

V. Conclusion

Addressing mental health challenges in tech culture is essential for human empathy, the benefit of organizations, and tech culture. When IT professionals feel unsupported and experience mental health challenges like anxiety, depression, and burnout, organizations will inevitably face problems such as presenteeism, absenteeism, and attrition, among others. This in turn will lead to a decrease in productivity and business revenues.

Therefore, investing in mental health is a need of the hour. It will lead to better outcomes for organizations, including stronger retention rates, revenues, and growth. According to Deloitte’s 2022 Mental Health and Employers: The Case for Investment - Pandemic and Beyond report, training and awareness-raising programs that target mental health generate ROIs of 6:1 and 5.3:1, respectively.

The data clearly supports the argument for better support of employees and their mental health. Thus, during a period of high turnover when many talented individuals are searching for employers who prioritize their interests, it’s crucial to meet their expectations.

Furthermore, we need to look for newer dimensions/solutions to deal with mental health in the tech industry. But could these different dimensions be? We don’t need to look far as most of the answers lie around us. For example- to cope with stress, distress, burnout, etc. individuals may benefit from participating in sports, hobbies, or taking a relaxing vacation.

The use of a stress score can aid in identifying those who may be susceptible to stress-related physical ailments, with individuals scoring above 300 being at risk for such illnesses. Early intervention is important to relieve stress and prevent potential health issues. By promoting the health of employees, the overall performance of the community can also improve. Therefore, an annual stress assessment should be conducted, and individuals scoring above 300 should be provided with anti-stress management resources [11].

VI. Future Aspirations

Going forward we would be interested in merging and examining Mental Health in Tech Survey data sets from other recent years to compare against the baseline year data set 2014. It would be insightful to observe if any changes have occurred in mental health within the tech industry since the world came out of the pandemic and what impact the pandemic has left on the global tech industry.

In order to gain a comprehensive understanding of attitudes toward mental health disorders in the workforce, we plan to expand our research beyond the technology industry. We believe that examining attitudes towards mental health in other industries such as healthcare, hospitality, transportation, and others can help us identify differences or similarities in how mental health is perceived and addressed in different work environments. For example, in the healthcare industry, mental health issues may be more commonly recognized and addressed due to the nature of the work, whereas in hospitality or transportation, the focus may be more on meeting customer needs rather than employee well-being.

Through this broader investigation, we hope to gain insights that can inform policies and initiatives to promote mental health and well-being in different industries.

VII. Source Code & Data

- Data: Google Drive
- Source Code: Github
- Reference Data: OSMI Health Survey, Kaggle

ACKNOWLEDGMENT

- We would like to express our sincere gratitude to *Prof. Jawahar Panchal* for providing us with guidance and support throughout this project. His teachings, insights, and feedback have been invaluable in shaping the final project and report.
- We would also like to thank all the participants of the *OSMI survey*. Their willingness to share their experiences and insights has contributed significantly to the findings and recommendations presented in this report.

REFERENCES

[1] Syed, N. U., Syed, K. U. (2019). Analysis of the Impact of Mental Health on Employee Productivity and Work Attitudes in the Technology Industry.

[2] Mohata, A., Sutar, P. (2021). A Comparative Analysis of Machine Learning Algorithms for Predicting Mental Health Disorders in the Tech Industry.

[3] Saini, R., Singh, N. (2020). Mental Health in Tech: Exploring the Relationship Between Work Stress, Job Satisfaction, and Burnout.

[4] Daptardar, A. V., Raisinghani, M. S., Dasgupta, A. K. (2018). Mental Health and Job Performance in the IT Industry: A Comparative Study of India and the United States.

[5] Thompson, David. 2022. "Mental Health - an Important Conversation in the Tech Industry." Tech Times. February 4, 2022.

[6] LinkedIn - How agile can contribute to peoples' mental health in the workplace?

[7] Figueroa, C. A., Aguilera, A. (2019). The Need for a Mental Health Technology Revolution in the COVID-19 Pandemic. Frontiers in Psychiatry, 11. https://doi.org/10.3389/fpsyt.2020.00523

[8] Conference: 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)

[9] Johnson, A., Dey, S., Nguyen, H., Groth, M., Joyce, S., Tan, L., Glozier, N., Harvey, S. B. (2020). A review and agenda for examining how technology-driven changes at work will impact workplace mental health and employee well-being. Australian Journal of Management. https://doi.org/10.1177/0312896220922292

[10] Naslund, J. A., Aschbrenner, K. A. (2021). Technology use and interest in digital apps for mental health promotion and lifestyle intervention among young adults with serious mental illness. Journal of Affective Disorders Reports, 6, 100227. https://doi.org/10.1016/j.jadr.2021.100227

[11] https://news.stanford.edu/2022/12/05/explains-recent-tech-layoffs-worried/