Depression Severity and Factor Analysis Using Reported Health Data with Machine Learning Models

1 INTRODUCTION

1.1 ABSTRACT

Depression is a significant global health concern, as it impacts individuals' daily lives and productivity. Challenges remain unexplored and unresolved, with a concerning factor that certain types of depression have little to no treatment [5], and it sometimes even remains undetected [8]. However, recent studies have demonstrated the potential of leveraging wearable device data, patient-generated health metrics and socio-demographic information into this complex health disorder [2] [3] [4].

Specifically, current research highlighted various factors, for example, a study done in Sweden found that socio-economic conditions, such as unemployment, economic hardships, as well as lifestyle factors (physical activity, alcohol consumption) contribute to mental health issues, and that young people more commonly have them. [2]. Digital engagement has also become a crucial aspect of mental health, with individuals suffering from mental health issues often turning to social media for support and information, particularly younger, lower-income women with chronic health challenges [3].

This study aims to address the gap in knowledge and investigate hypotheses by employing machine learning models to predict depression severity while exploring key contributing factors. Particularly, Support Vector Machines have shown promise in handling depression related predictions [4], and thus have been used in this project, with various kernel methods and optimization techniques, utilizing feature selection by PCA.

The question of this research is, what factors influence depression, and how accurately can the severity of the mood disorder be predicted using machine learning models?

Preliminary findings reveal correlations between depression severity and factors such as physical activity, sleep hygiene and comorbid health conditions. The optimized model used for the prediction achieved a high R^2 score of 0.94, and a low Mean Squared Error of 2.1. These results are consistent with existing literature, demonstrating the potential of integrating wearable technology and machine learning research.

2 PROBLEM FORMULATION

Depression is an often-overlooked, widespread mental health disorder, yet it significantly affects the lives of patients; Despite advancements in diagnoses within the field, there always exists room for improvement in identifying and evaluating the key factors that contribute to its severity. Therefore, the emphasis on this challenge is critical, given that globally, 5% of all adults worldwide suffer from depression [5]. By examining data metrics, such as wearable device metrics, patient-generated health-data and socio-demographic information, it is possible to shed light on the factors contributing to this mental health disorder. However, the characteristic of these datasets is high-dimensional, multi-faceted and complex—which makes it difficult to draw reliable conclusions. This study offers an opportunity to make exceptionally accurate predictions by leveraging machine learning.

This study aims to investigate the factors and relationships within the aforementioned generated data to predict the severity of depression accurately by comparing the effectiveness of various kernel methods and various machine learning models, such as SVM's linear, polynomial, RBF and sigmoid kernels, with cross-validating and parameter tuning. Ultimately, these insights can hold potential for enhancing identification of depression, and to support earlier intervention for healthcare professionals.

Additionally, this research proposes to test a specific hypothesis, as presented by the related papers: that individuals suffering from depression are less physically active, have poorer sleep quality and have chronic health problems. This is to further investigate our understanding of the condition and its factors.

3 DATASET DESCRIPTION

3.1 DATASET DESCRIPTION, SOURCE, FEATURES

The dataset consists of 35,694 rows and 154 columns, each row representing an observation for a specific participant during a specific month, over the time period of 12 months. In this case, the index is labelled in the following format: *[participant]_[month]* (e.g., 34_12 indicates data from participant 34 during month 12). All identifiers are anonymized. Each datapoint corresponds to an individual's monthly aggregated data. In total, there were 10,036 individuals surveyed who wore wearable devices and completed various surveys. It is important to note that there are numerous missing values.

The dataset includes a mix of categorical, binary, numerical and continuous features related to demographic, behavioural, and health status indicators. Since there are plenty of features to describe, they have been intuitively classified by type, as shown below:

- Wearable Device Data (PGHD)
  - Includes step count and sleep metrics gathered from Fitbit.
  - Daily metrics, which are aggregated monthly.
  - Data was recorded 8–14 days prior to each quarterly PHQ-9 assessment to closely align behavioural metrics with mental health data.
- Mental Health Surveys (PHQ-9)
  - Patient Health Questionnaire, assessing severity of depression.
  - Assessed every 3 months.
- Lifestyle and Medication Changes (LMC)
  - Monthly self-reported data on lifestyle adjustments and medication usage.
  - Monthly assessed.
- Demographic and Socioeconomic Data
  - Single time screener survey at the study's start, capturing demographic and socioeconomic data.

Each participant's data is structured into a non-overlapping, 3-month sample, totalling to 10,866 samples across 4,036 participants. The data frequency is aggregated to 3-month intervals.

The dataset used can be accessed from Zenodo [1] with a more detailed description, derived from the DiSCover Project.

3.2 Dataset Preprocessing – Imputation, cleanup

The handling of the dataset assumes that the timely intervals which the dataset was recorded do not bear any significance on the report, as the objectives stated before do not require them.

To have one singular decisive label for the models 'phq9_score_end' will suffice for predictions. Since there is an immense amount of missing label scores, the dataset was cleaned by dropping rows where a participant has not completed the PHQ-9 survey. This will directly ensure that all labels are covered, unchanged and true, without the need to use techniques such as imputation which can introduce noise and unrelated correlation with other features. Afterwards, rows with duplicates were dropped to reduce redundant entries, even though all data was unique.

For the remaining missing numerical data, the usage of Z-Scores was employed to detect outliers and remove them. Any datapoints with a Z-score greater than 3 thresholds in absolute value were removed, as 3 is typically used in the data preprocessing segment in the field [10], and it will only detect the most extreme outliers. The explanation behind this is that Z-scores identify data points that deviate significantly from the mean. This step ensures the dataset is normalized for the analysis.

3.3 Feature Selection and Engineering

After cleaning, this leaves 10866 rows to work with, which will be sufficient for the exploration of the study. However, not all missing values had been eliminated. Therefore,

a method of imputation is done to the dataset, which means that a model is used to estimate and fill missing values in the dataset by taking into account other features, its relations, correlations with each other. KNN was chosen to address this, as it leverages the relationship between similar data points, according to its nearest neighbour [6]. This method is effective particularly when the dataset has a high number of features and a consistent correlation across data. Although, it is important to note that KNN performs best with a low number of features, as it is sensitive to noise, which could create an overfitting problem [6]. However, the usage of Z-scores alleviates this issue.

For the fine-tuning of the KNN method, the training and test data were split with a 70-30 ratio. Different values of the number of neighbours (k) were tested. Evaluating it through multiple iterations, an optimal value of k=3 was chosen by determining whichever minimized error the most. With this, it is sufficient to use a KNN with k=3 to impute all data.

PCA, as a dimensionality reduction technique, was used for feature engineering and determining which features play a critical role. In general, PCA reduces dimensionality to mitigate overfitting while retaining key information and eliminating redundant ones. A 50% threshold was used to filter the amount of features, and the final count of features, as proven to be sufficient, is 53.

3.4 Complete Literature Review

3.4.1 Mental health symptoms in relation to socio-economic conditions and lifestyle factors – a population-based study in Sweden [2]

This article examines multiple factors which could contribute to mental health symptoms, with particular focus on socio-economic conditions, lifestyle habits and personal experiences that could correspond with anxiety and depression symptoms.

Researchers found that mental health issues were reported by 40% of women and 30% of men, with **younger people experiencing poorer mental** health condition than those in their senior years.

Key factors included **poor social supports, experiences of belittlement, unemployment, economic hardships, critical life events and functional disabilities**. Additionally, **lifestyle factors such as physical activity, body fitness and risky alcohol consumption** also played a role.

3.4.2 Social Networking Service, Patient-Generated Health Data, and Population Health Informatics: National Cross-sectional Study of Patterns and Implications of Leveraging Digital Technologies to Support Mental Health and Well-being [3]

The study examines the differences in health, digital engagement and social factors between individuals with and without mental health issues. The survey included people with conditions such as depression and anxiety, **revealing a higher prevalence of poor physical health, inadequate sleep and higher alcohol consumption** among those with mental health issues. These participants were **in general younger, female**

and of lower income, tending to face more **chronic health challenges**. They also showed higher engagement with social media platforms for health information, often using these platforms to share their experiences and participate in support groups.

3.4.3 Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions [4]

This review paper goes over how certain machine learning algorithms can be applied in order to **diagnose and detect depression**, along with other mental health conditions. The algorithms are categorized into three types: classification, deep learning, and ensemble methods.

It concludes that **Support Vector Machines are favoured for depression diagnosis**, particularly because of their high accuracy (75%~) and avoidance of overfitting.

4 METHODS

4.1 BASIC METHODS USED

For the study, numerous libraries were needed for a sufficient report:

- **NumPy** was used for linear algebraic operations to handle the manipulation of arrays.
- **Pandas** provided tools for data manipulation, for storing, organizing, cleaning, exploring, and analysing the data.
- **Matplotlib** helped by creating various visualizations, to show correlations, distributions and trends over time.
- **Seaborn** was used, dependent on Matplotlib, to create more easily interpretable, aesthetically pleasing visualization.
- **Scikit-Learn** was utilized to provide functionalities for preprocessing (e.g. Z-Scores), scaling, feature-selecting (PCA), imputing (KNN), evaluating (R2 Scores, MSE), cross validating, splitting and predicting data (SVR).

Descriptive statistics were performed to understand the data distribution and identify potential issues like skewness or data imbalance. Visualizations included histograms, line plots, correlation matrices to identify pattern and relationships.

4.2 SUMMARY OF PREPROCESSING

This pipeline was designed to prepare the dataset for the modelling phase. Multiple techniques were applied to ensure the data was ready for analysis:

- **Handling missing data:** Initially, rows with missing PHQ-9 scores were removed to ensure the dataset is complete. This exclusion ensured that no inaccurately imputed data would affect the results.

- **Removing outliers:** Using Z-Scores, the detection and removal of outliers was made possible. Data-points with Z-Scores greater than 3 in absolute value were identified as extreme, and therefore, discarded. This is to ensure that the model would not be influenced by extreme values which could distort predictions. [10]
- **Imputation:** To fill remaining missing values in numerical data, K-Nearest-Neighbours (KNN) imputation was employed. This method was chosen for the main reason that it takes into account the relationship between similar datapoints to estimate the missing values. An optimal value of k=3 was selected based on cross-validation of performance.
- **Dimensionality reduction:** Principal Component Analysis (PCA) was used to reduce the dimensionality of the feature space, to retain only the most important features for model training, to mitigate overfitting and ensure an accurate prediction. A 50% threshold for cumulative explained variance was chosen, as that yielded an adequate amount of features.

4.3 Advanced Methods for Prediction

To predict the severity of depression, machine learning models were employed, with a focus on Support Vector Machines (SVM). The reasoning behind this was motivated by a related research report **[3]**, along with SVM having the ability to handle complex datasets with high dimensional spaces, such as the dataset, while also handling non-linear relationships.

To ensure complete reliability and accuracy with this model, several kernel functions were tested, such as linear, polynomial, RBF, sigmoid. The training data was scaled per industry standards and reliability. Alongside this, HalvingGridSearch, a hyperparameter tuning method, was employed, to help optimize and determine the best parameters for the model **[9]**. HalvingGridSearch is a special type of GridSearch. It converges faster than the standard one, as it progressively increases its subsets of data in the early stages to find the best parameter combinations, iteratively increasing it to narrow it on the best combinations. This makes it better than standard GridSearch, which goes over every possible iteration, and is slower, more computationally costly.

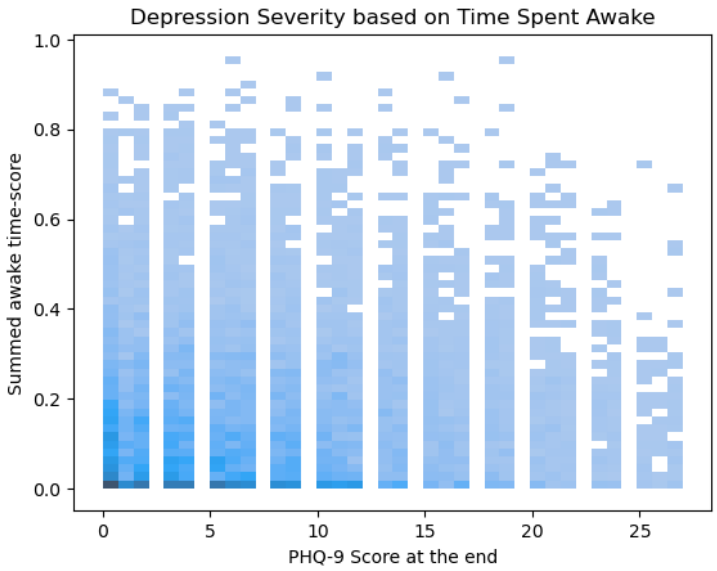
In order to evaluate the model's performance, cross-validation was used to ensure the best results, across different subsets of the data. Concisely, a 5-fold cross validation was performed, and the model was evaluated by R^2 scores and Mean Squared Errors. The results of the cross validation were averaged.

5 RESULTS

Firstly, a basic examination of features and its relations to the label was executed.

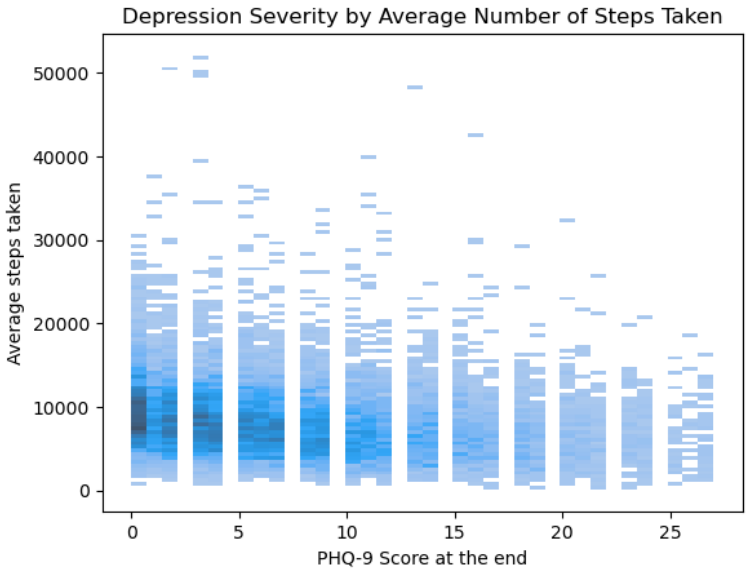
1. Figure: Correlation between Top 10 Features in the dataset. The Top 10 Features are, for phq9_end_score's column, in descending order are: phq9_cat_end, phq9_score_start, phq9_cat_start, comorbid_migraines, comorbid_neuropathic, num_migraine_days, steps__sedentary_day_count_, steps_awake_mean, steps__rolling_6_sum__max__intercept_.

Figure 1 implicates that comorbid conditions (migraines, neuropathic), the amount of sedentary days, steps taken on average in a 3-month period, and how much the individual walks in any 6 minute window/interval at maximum over the past 4 days are related to the severity of depression.



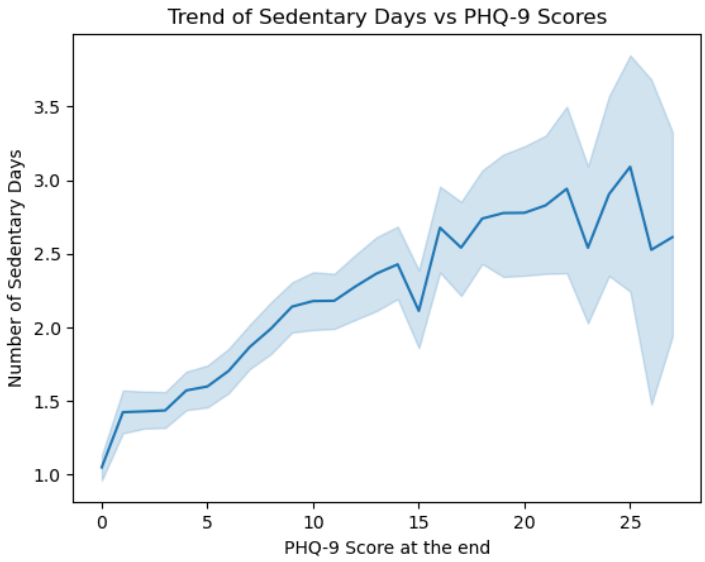
2. Figure: Correlation between PHQ9 Scores and Time Spent Awake. The data gets scarcer, the higher the scores are. The distribution of time spent awake is also scarcer.

Figure 2 implies that the more severe the depression score is, the less time the individual stayed awake for.



3. Figure: Depression Severity distribution by Average Number of Steps Taken. The higher the PHQ9 Score is, the less steps on average the individual takes.

Figure 3 implies that the more depressed the individual is, the less steps they take on average.



4. Figure The Trend of Sedentary Days, its mean and variance by PHQ-9 scores.

Figure 4 shows that more sedentary days were accumulated among people who have higher rates of depression.

The best parameters for the model were:

- 'C': 1
- 'degree': 4
- 'gamma': 'auto'
- 'kernel': 'linear'

For the best model, the Mean Squared Error was 2.1, and the R^2 score was 0.94.

6 Conclusion & Discussion

6.1 Result Interpretation

From the diagrams and the model's results, it is possible to infer that individuals suffering from comorbid conditions, such as migraines, prolonged migraines and neuropathic pain are associated with a higher depression score. Furthermore, higher physical activity is also heavily associated with lower depression scores; low scorers of PHQ-9 are associated with more steps taken, and more active days spent. Time spent awake is inversely correlated with the severity of depression, as it potentially indicates disrupted sleep patterns or fatigue.

Among all the various kernel functions tested, the linear kernel performed the best with an R^2 score of 0.94 and a Mean Squared Error of 2.1. This suggests that a linear relationship is present between the selected features and the label, and that it is a sufficiently strong predictor. Furthermore, comorbid conditions, steps taken and time spent awake emerged as the most influential predictors, supporting the initial hypothesis that these factors play a critical role in increasing PHQ-9 scores.

6.2 Implications

While these findings provide valuable insights, they must also be interpreted with caution, as correlation does not imply causation. The conclusions reached in this report are merely symptoms of a much more complex problem and can only be used as pointers for diagnosis. For example, reduced physical activity cannot imply that it is the cause of depression, as it could simply be an effect of an underlying problem or condition.

However, the results show that wearable device data, and socio-demographic information is sufficient to predict whether the individual is depressed or not, such as by simply measuring sleep hygiene and physical activity levels. These can be used for early detection, and even personalized interventions. In clinical assessments, such as in PHQ-9, a method of including this data could prove to be beneficial for a more accurate evaluation.

Overall, these implications, by majority, are consistent with the conclusions found in related literature on the topic of depression [1] [2].

6.3 Limitations and Shortcomings

In this study, there must be several limitations to be considered. A lot of the missing data (exactly 24828, ~70% of the dataset) was cut from the original dataset, and additionally, there were still missing values left, which meant that there presented a need to impute data. This method of imputation regarding the data can introduce biases and inaccuracies.

Analysing the individuals who were used for data gathering, it only consists of participants whom have agreed to wear devices for measurement and have completed surveys. It only consists of participants who agree to wear devices for measurement; therefore this population is more likely to be health conscious or have the means to access these devices to monitor their health', resulting in the fact that it might not generalize well to the general public.

And while SVM did perform well, certain features could have been overlooked during the elimination of them by the PCA method, as a means for dimensionality reduction.

6.4 Future Steps

To alleviate the limitations of the report, a richer dataset would greatly enhance the accuracy of predictions. More features to the dataset, such as data on social interactions, mood and dietary patterns would not only be beneficial, but necessary, as depression is an inherently complex mental health disorder.

Moreover, more models could be tested on the dataset. As the results showed there was a linear correlation, more linear models could be employed and compared to find an optimal prediction model for health care professionals.

A more complex, longitudinal investigation of this data could also shed light on how depression develops over time.

Finally, a more diverse, complete and larger dataset could mitigate potential selection bias, as highlighted in the above limitations.

6.5 CONCLUSION

The potential presented by the usage of wearable technology and machine learning models to deepen our understanding of depression is nontrivial—this has been demonstrated by the significant correlation between physical activity, amount of quality sleep gotten and comorbid condition between depression score, and the high predictions of the model deployed. This concludes that the processing of data gathered by wearable technology serve as a great foundation to improve our current services in healthcare, by showing the importance of leveraging modern, everyday technology for diagnoses and personalized interventions.

7 REFERENCES

[1] Makhmutova et Al. PSYCHE-D: predicting change in depression severity using person-generated health data (DATASET) https://zenodo.org/records/5085146

[2] Molarius, A., Berglund, K., Eriksson, C. et al. (2009) Mental health symptoms in relation to socio-economic conditions and lifestyle factors – a population-based study in Sweden. BMC Public Health 9, 302. https://doi.org/10.1186/1471-2458-9-302

[3] Ye, J., Wang, Z., & Hai, J. (2022). Social Networking Service, Patient-Generated Health Data, and Population Health Informatics: National Cross-sectional Study of Patterns and Implications of Leveraging Digital Technologies to Support Mental Health and Well-being. *Journal of medical Internet research*, *24*(4), e30898. https://doi.org/10.2196/30898

[4] Aleem S, Huda Nu, Amin R, Khalid S, Alshamrani SS, Alshehri A. (2022) Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions. Electronics. 11(7):1111. https://www.mdpi.com/2079-9292/11/7/1111

[5] WHO – World Health Organization Depressive disorder (depression). https://www.who.int/news-room/fact-sheets/detail/depression

[6] Aditya Kumar (2020) KNN Algorithm: When? Why? How?. Medium. https://towardsdatascience.com/knn-algorithm-what-when-why-how-41405c16c36f

[7] Aymane Hachcham (2023) The KNN Algorithm – Explanation, Opportunities, Limitations. Neptune Blog. https://neptune.ai/blog/knn-algorithm-explanation-opportunities-limitations#:~:text=KNN%20is%20most%20useful%20when,of%20desired%20precision%20and%20accuracy

[8] Handy, A., Mangal, R., Stead, T. S., Coffee, R. L., Jr, & Ganti, L. (2022). Prevalence and Impact of Diagnosed and Undiagnosed Depression in the United States. Cureus, 14(8), e28011. https://doi.org/10.7759/cureus.28011

[9] Brian Roepke (2024) 5-10x Faster Hyperparameter Tuning with HalvingGridSearch. Data Knows All. https://dataknowsall.com/blog/hyperparameter.html

[10] Sarah Thomas (2021) Z-Score: Formula, Examples & How to Interpret It. Outlier. https://articles.outlier.org/z-score-formula-examples-and-how-to-interpret#:~:text=Z-scores%20are%20measured%20in%20standard%20deviation%20units.&text=The%20further%20away%20your%20Z,standard%20deviations%20of%20the%20mean.