Exploratory Data Analysis of ME/CFS vs. Depression

PREPARED BY: HEMAL MEWANTHA -S16231
SANJANA FERNANDO-S16145
MAHEESHA SEWMINI – S16349

Table of Contents

List of tables

List of figures

Abstract

This study explores the differentiation between Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) and Depression using a dataset of 1,000 patients with 16 variables, including demographic, symptom, and lifestyle factors. Descriptive analysis revealed key distinctions: ME/CFS patients exhibited higher fatigue severity and consistent fatigue levels regardless of work status, while Depression patients showed lower but more variable fatigue scores. Sleep quality was poor across all groups, though ME/CFS patients reported longer sleep durations, aligning with clinical observations of non-restorative sleep. Depression scores were highest in comorbid cases, with ME/CFS-only patients displaying milder symptoms. Gender and age distributions were similar across diagnoses, contradicting some real-world trends. Factor analysis indicated no clear clustering, suggesting complex interactions among variables. These findings highlight the need for advanced machine learning models and Explainable AI (XAI) tools to improve diagnostic accuracy and interpretability, addressing challenges like class imbalance and feature selection. The study underscores the potential of data-driven approaches to disentangle overlapping symptoms and support clinical decision-making.

Introduction

Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) is a complex multisystem disorder that is characterized by persistent fatigue, post-exertional malaise, sleep disturbances, and a wide array of additional symptoms including cognitive impairment and musculoskeletal pain. In clinical practice, ME/CFS is particularly challenging to diagnose, not only because of the absence of definitive laboratory tests but also due to its significant symptomatic overlap with other conditions, especially depression. Depression is itself marked by low mood, anhedonia, cognitive slowing, and fatigue, all of which may also occur in ME/CFS patients, further complicating diagnostic clarity. Recent studies have shown that integrating multi-modal data developing robust classification models can help disentangle these overlapping features. However, the high dimensionality of symptom profiles and the presence of confounding personal characteristics often limit the interpretability and clinical applicability of such models. Moving forward, continued research that prioritizes both model accuracy and interpretability will be essential for improving differential diagnosis and guiding more personalized, effective treatment strategies for ME/CFS and Depression related conditions.

Description of the Problem

Objective 1-Build a machine learning model to classify whether a person has ME/CFS or Depression using available data.

This objective focuses on the core task of developing a predictive model. It involves selecting appropriate machine learning algorithms, preprocessing the data (including handling missing values and encoding categorical variables), training the model on the available dataset, and evaluating its performance in accurately classifying individuals into the respective diagnostic categories (ME/CFS, Depression, or Both).

Objective 2-Find out which symptoms, survey answers, or personal details (like age or gender) are most helpful in telling the two conditions apart.

This objective delves into feature importance and selection. It aims to identify the specific variables within the dataset that contribute most significantly to the model's ability to differentiate between ME/CFS and Depression. Techniques such as feature importance scores from tree-based models or statistical tests can be used to highlight the key symptoms, survey responses, or demographic factors that are most discriminative.

Objective 3-Use Explainable AI tools (SHAP or LIME) to understand why the model makes certain predictions, so doctors and researchers can trust and use it.

This objective addresses the need for transparency and interpretability in the machine learning model. By applying XAI tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), we can gain insights into how individual features influence the model's predictions for specific instances. This enables a deeper understanding of the model's decision-making process, making it more trustworthy and useful for clinicians and researchers in identifying the factors that contribute to a diagnosis.

Description of the Dataset

The dataset utilized in this project, me_cfs_vs_depression_dataset.csv, comprises 1000 records and 16 variables.

Variable Name	Description	Type	Categories/Values
age	Age of the patient	Numerical	-
gender	Gender of the patient	Categorical	Male, Female
sleep_quality_index	Index representing the quality of sleep	Numerical	-
brain_fog_level	Level of brain fog experienced by the patient	Numerical	-
physical_pain_score	Score indicating the level of physical pain	Numerical	-
stress_level	Level of stress experienced by the patient	Numerical	-
depression_phq9_score	Score from the PHQ-9 depression questionnaire	Numerical	-
fatigue_severity_scale_score	Score from the Fatigue Severity Scale	Numerical	-
pem_duration_hours	Duration of post-exertional malaise in hours	Numerical	-
hours_of_sleep_per_night	Number of hours of sleep per night	Numerical	-
pem_present	Indicates the presence of post-exertional malaise	Categorical	Yes, No
work_status	Work status of the patient	Categorical	Working, Partially working, Not working

social_activity_level	Level of social activity	Categorical	Low, Very low, High, Very high, Medium
exercise_frequency	Frequency of exercise	Categorical	Daily, Never, Often, Rarely, Sometimes
meditation_or_mindfulness	Indicates if the patient practices meditation or mindfulness	Categorical	Yes, No
diagnosis	The target variable, indicating the diagnosis	Categorical	ME/CFS, Depression, Both

Table 1 Variable Descriptions

Data Preprocessing and Data Cleaning

Missing values were present across several variables in the dataset. To address this, a preprocessing pipeline was constructed using **ColumnTransformer** with **IterativeImputer**, which is based on the MICE (Multiple Imputation by Chained Equations) approach. Missing data in numerical variables were imputed using a **RandomForestRegressor**,while categorical variables were first encoded with **OrdinalEncoder**, then imputed using **IterativeImputer** with a **RandomForestClassifier.**

The dataset was split into a **training set (70%)** and a **testing set (30%)**, and the imputation pipeline was applied to each subset separately. After imputation, the resulting arrays were converted back into pandas DataFrames,and the categorical variables were mapped back to their original string labels for better interpretability.Additionally,univariate outliers in numerical features were assessed using the **Interquartile Range (IQR)** method. No duplicate records were found in the dataset.

Results of Descriptive Analysis

Target Variable-Diagnosis

The diagnosis variable consists of three categories: ME/CFS, Depression, and Both.

The distribution of diagnoses appears imbalanced, with ME/CFS and Depression being reported at similar and relatively higher frequencies, while the Both category show a notably lower count.

This imbalance suggests the need to consider resampling techniques such as SMOTE during advance analysis to address potential bias in model training.

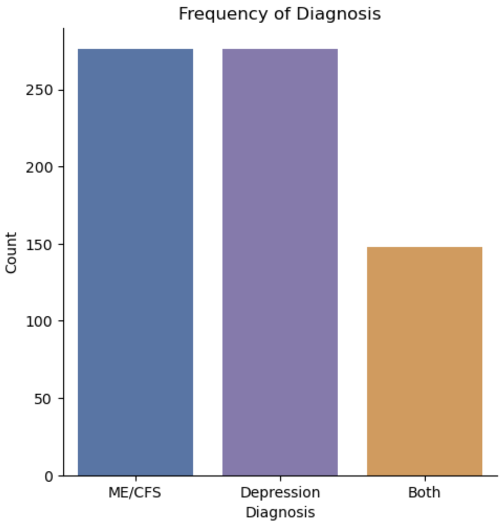

Figure 1 Bar chart of Diagnosis

Predictor Variable-Gender

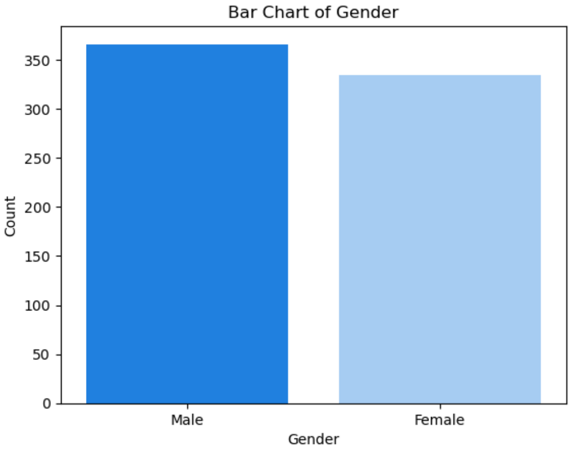

Figure 2 Bar chart of Gender

By Figure 2, males exhibit a higher count in the bar chart, indicating a greater likelihood of being affected by this diagnosis.

In contrast, females show a slightly lower count compared to males.

However, these historical clinical results present a slight contradiction with the real-world clinical records.

"Both disorders are more common in females, gender alone doesn't rule one out" [2].

Correlation among the variables

The correlation revealed a generally weak relationship among all the numerical variables. The highest correlation observed was a modest negative association between depression scores and fatigue severity, while most other variable pairs showed negligible correlations.

This lack of strong linear correlations is beneficial for advanced analysis, as it indicates minimal multicollinearity's result, future machine learning models can include these variables without concern for overlapping effect, enabling clearer interpretation of each factor's unique contribution.

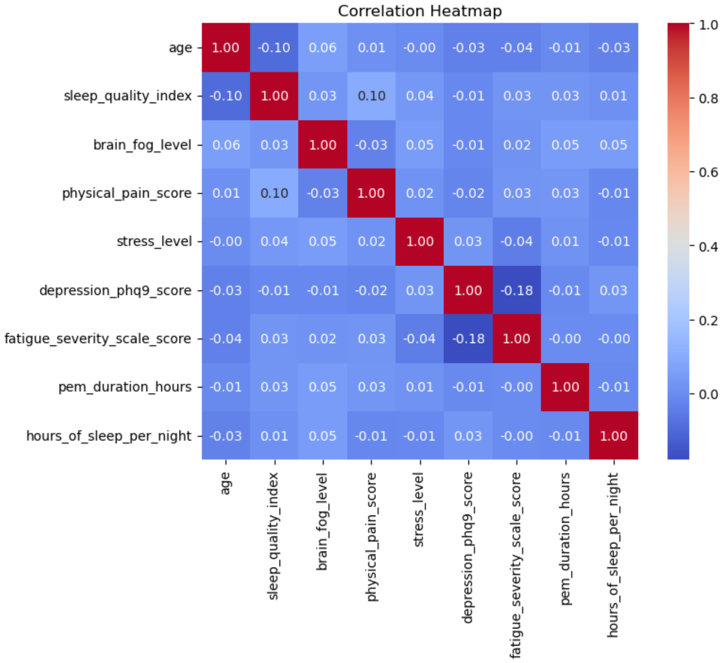

Figure 3 Correlation Heatmap

Exploring the Influence of Predictor Variables on the Response Variable

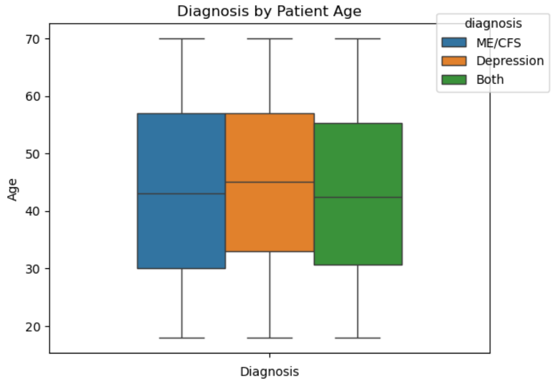

Figure 4 Box plot of Patient Age by Diagnosis

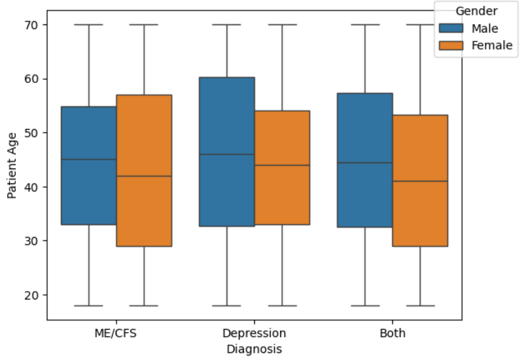

Figure 5 Box plot of Patient Age, Gender by Diagnosis

Figure 4 illustrates that most patients diagnosed with ME/CFS, Depression, or both conditions fall withing early to mid-adulthood, with relatively few cases observed in childhood or older age.

"For instance, NIH reports that the median age of ME/CFS onset is about 33 years" [3].

“Depression also most often begins in early adulthood – large studies report a median onset around age 31” [3].

Figure 5 further explores this pattern by incorporating gender. It shows that both male and female patients across all three diagnostic categories exhibit similar age distributions, predominantly clustered in early to mid-adulthood. There is no clear indication of substantial age differences between genders withing this diagnosis. This suggests that age at onset for ME/CFS, Depression, or their co-occurrence does not markedly differ between males and females.

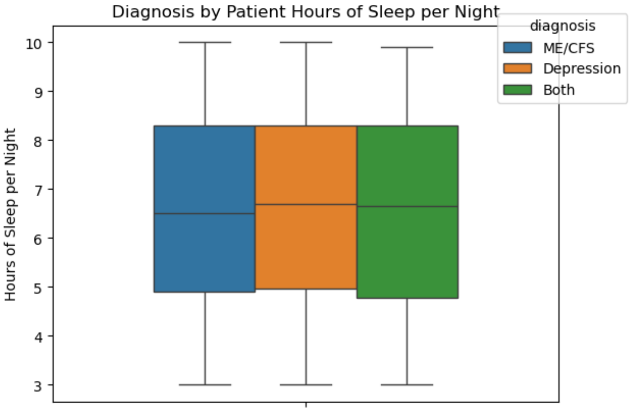

Figure 6 Box plot of Hours Sleep by Diagnosis

Historical clinical records show that hours of sleep per night are similar across patients with ME/CFS, Depression, or both, differing from real world clinical patterns where ME/CFS patients usually sleep for longer duration due to persistent fatigue while depression patients sleep less. *“Sleep studies note that ME/CFS patients can have extended total sleep times, but the sleep is non-restorative*” [7].

In addition, By Figure 7, illustrate that similar sleep duration, sleep quality levels are poor across all groups. This reflects how ME/CFS patients often feel unrefreshed despite longer sleep, while depression patients also experience low sleep quality linked to shorter or disrupted sleep. It highlights the need to evaluate both sleep quality and quantity in these patients.

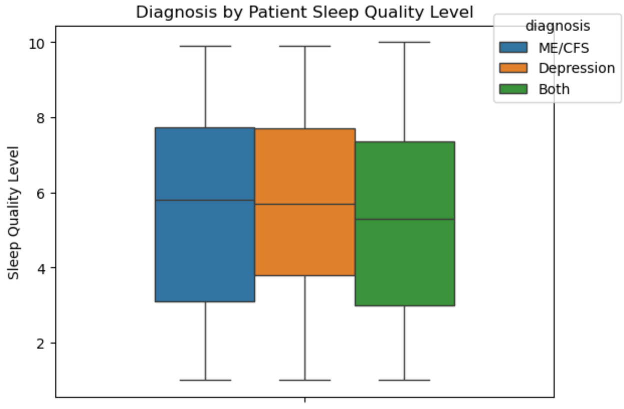

Figure 7 Box plot of Sleep Quality by Diagnosis

Figure 8 illustrates the distribution of depression (PonHQ-9) scores among patients based on diagnosis. Patients diagnosed with both ME/CFS and Depression exhibit the highest median depression scores, likely due to the strong influence of depression on the overall score. Those with depression alone also show relatively high scores, but slightly lower than the group with both conditions. In contrast, ME/CFS-only patients show notably lower depression scores, with a sharp peak around 8-10 range, as seen in the density plot, suggesting most of these patients experience mild depression. However, a few outliers suggest some ME/CFS patients may still report unusually low depression scores. By figure 9, the accompanying gender-based analysis show no significant differences in depression scores between male and female patients across all diagnostic categories, indicating gender does not substantially influence the observed pattern.

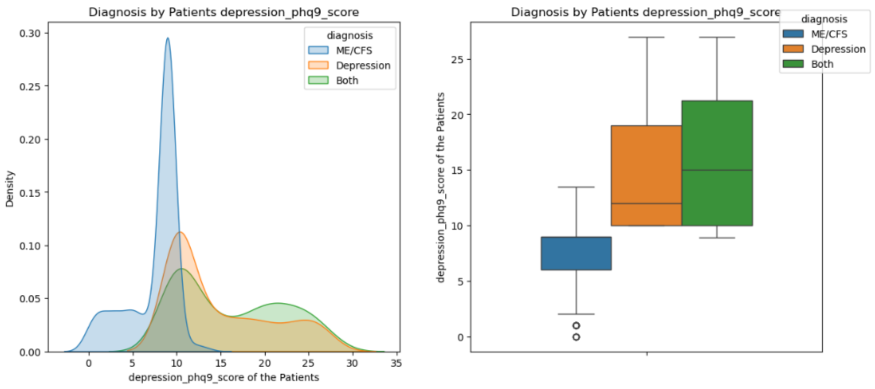

Figure 8 Box plot, Density plot of Depression score by Diagnosis

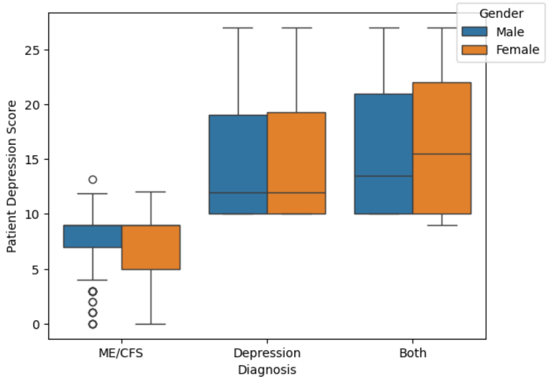

Figure 9 Box plot Depression score, Gender by Diagnosis

Figure 10 shows the distribution and spread of fatigue severity scores across diagnostic groups. Patients with ME/CFS and those with both ME/CFS and depression exhibit the highest median fatigue scores, with a strong density concentration between scores of 6 and 8, suggesting consistently high fatigue levels

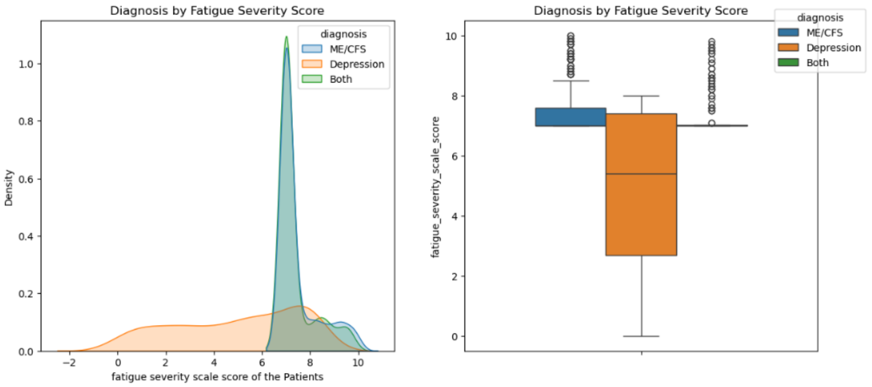

Figure 10 Box plot, Density plot Fatigue score by Diagnosis

These two groups also share nearly identical distribution patterns as shown in the left-hand side. Several outliers above this range indicate unusually high fatigue scores within these groups. In contrast, patients diagnosed with depression alone show a much lower median fatigue score and a wider spread, indicating greater variability in fatigue levels. In Figure 11, the accompanying bar chart demonstrates that ME/CFS patients are more likely to be working or partially working compared

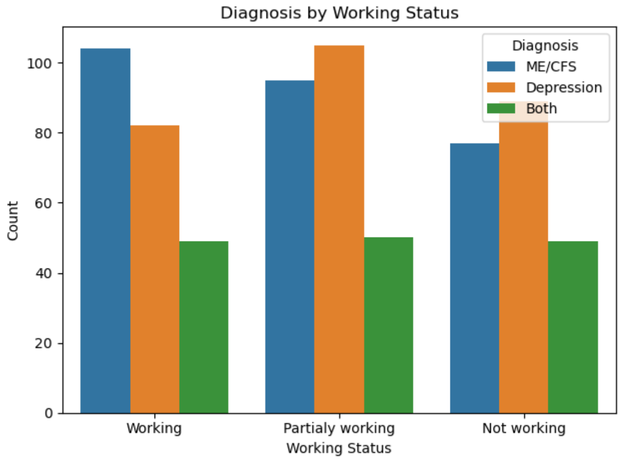

Figure 11 Bar chart of Working status by Diagnosis

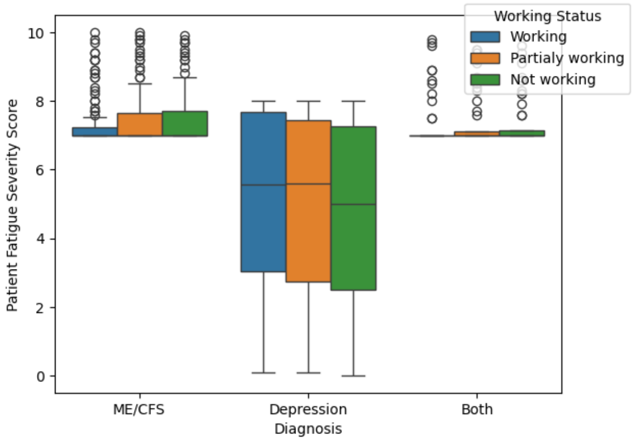

Figure 12 Box plot of Fatigue score, Working status by Diagnosis

to those with depression or both conditions. This may initially seem counterintuitive given their high fatigue levels but could reflect differences in coping strategies or support systems. To explore this further, Figure 12 presents fatigue severity scores by diagnosis and working status. It shows that within the ME/CFS group, working status has a minimal impact on fatigue scores, as all three employment categories reflect similarly high median fatigue levels.

This suggests that fatigue severity in ME/CFS patients is consistently high regardless of work status, supporting the notion that employment does not significantly influence fatigue levels in this group.

Factor Analysis for Mixed Data

After completing the graphical exploration each variable, we applied a dimensionality reduction technique known as Factor Analysis of Mixed Data (FAMD) to compress the dataset while retaining as much variance as possible from both numerical and categorical variables.

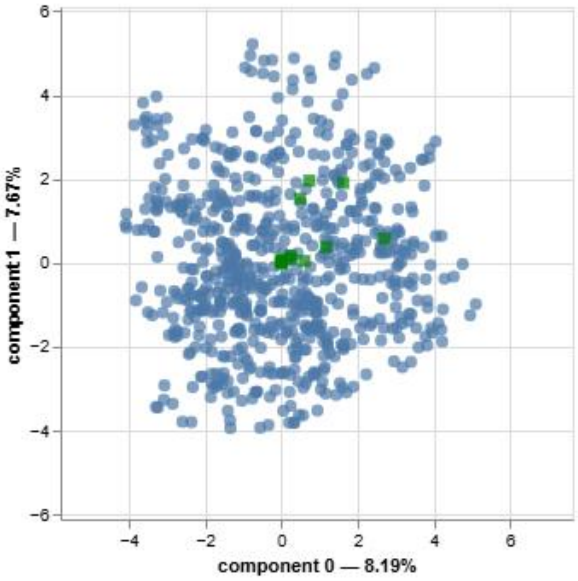

Figure 13 Scree plot

Figure 13 shows the first two components, which together explain a small portion of total variance around 16%. The points are widely scattered without any visible clusters, indicating that no clear group structure is present in the lower-dimensional space.

Figure 14 highlights multivariate outliers identified using the robust mahalnobis distance. Outliers (shown in red) are spread across space, suggesting that unusual patterns do not exist in the dataset.

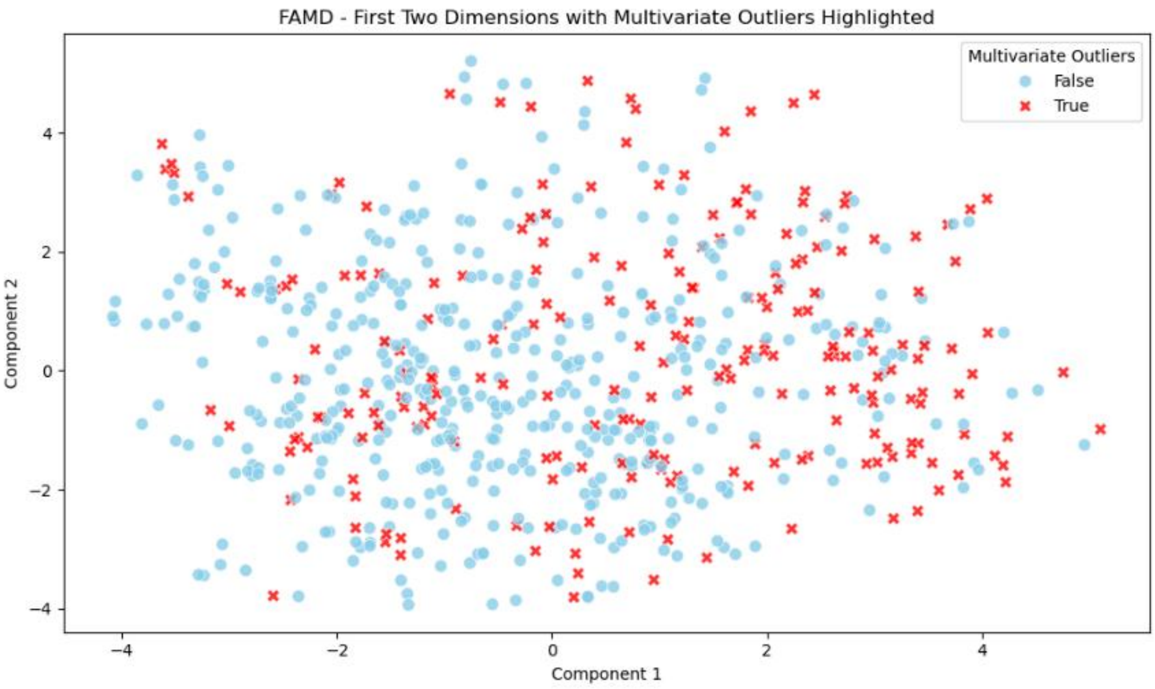

Figure 14 Outlier Detection plot

Suggestions for Advanced Analysis

Objective 1-Build a machine learning model to classify whether a person has ME/CFS or Depression using available data.

- The diagnosis variable is imbalanced (i.e. the “Both” category is underrepresented). Use SMOTE or class weighting during model training to ensure fair performance across all classes.
- Use cross validation techniques to evaluate performance metrics (accuracy, precision, recall, F1-score) fairly across all three classes.

Objective 2-Find out which symptoms, survey answers, or personal details (like age or gender) are most helpful in telling the two conditions apart.

- Use feature importance scores from tree-based models (e.g. Random Forest,XGBoost) to rank variables.
- Confirm findings with permutation importance for robustness.

Objective 3-Use Explainable AI tools (SHAP or LIME) to understand why the model makes certain predictions, so doctors and researchers can trust and use it.

- Use SHAP for Global and Local Interpretability.
- Use LIME for instance-Level Explanations.

- Use SHAP beeswarm plots to showcases how variables contribute positively or negatively to each diagnostic class.

References

[1] Background - Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome - NCBI Bookshelf. (n.d.). Retrieved July 8, 2025, from https://www.ncbi.nlm.nih.gov/books/NBK284897/

[2] Chronic Fatigue Syndrome or Depression? Similarities, Differences, and Diagnosis. (n.d.). Retrieved July 8, 2025, from https://www.webmd.com/depression/cfs-vs-depression

[3] Chronic Fatigue Syndrome vs. Depression - DMA. (n.d.). Retrieved July 8, 2025, from https://discoverymood.com/blog/chronic-fatigue-syndrome-versus-depression/

[4] Factor analysis of mixed data | Prince. (n.d.). Retrieved July 8, 2025, from https://maxhalford.github.io/prince/famd/

[5] Jackson, M. L., & Bruck, D. (2012). Sleep Abnormalities in Chronic Fatigue Syndrome/Myalgic Encephalomyelitis: A Review. *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, *8*(6), 719. https://doi.org/10.5664/JCSM.2276

[6] MICE imputation - How to predict missing values using machine learning in Python - Machine Learning Plus. (n.d.). Retrieved July 8, 2025, from https://www.machinelearningplus.com/machine-learning/mice-imputation/

[7] Sleep Abnormalities in Chronic Fatigue Syndrome/Myalgic Encephalomyelitis: A Review - PMC. (n.d.). Retrieved July 8, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC3501671/

Appendix

Python Code and Dataset:https://github.com/hemalmewan/Data-Analysis-Project.git