Advance Data Analysis of ME/CFS vs. Depression

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

Table of Contents

List of Figures

List of Tables

Abstract

This study developed machine learning models to classify ME/CFS, Depression, and comorbid cases using 1,000 patient records. To address Objective 1 (building robust classification models), we implemented Support Vector Classifier, Random Forest, XGBoost, and Decision Tree algorithms with hyperparameter tuning and 5-fold cross-validation. For Objective 2 (identifying discriminative features), we applied SHAP analysis and permutation importance to determine which clinical variables best differentiate the conditions. To achieve Objective 3 (explainable AI implementation), we used SHAP values for model interpretability and clinical decision support. Results showed SVC achieved 99.67% accuracy, with SHAP analysis revealing depression_phq9_score and Post-Exertional Malaise indicators as the most discriminative features. A simplified model using only these two features maintained identical performance, demonstrating that complex diagnostic challenges can be reduced to clinically interpretable solutions.

1. Introduction

Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) and Depression present overlapping symptoms that create diagnostic challenges in clinical practice. Our EDA revealed class imbalance, overlapping symptom distributions, and weak linear correlations among variables, indicating the need for advanced non-linear modeling approaches.

The preliminary analysis and Fisher's Discriminant Analysis (FDA) revealed a critical pattern: while the FDA model can clearly differentiate the Depression group from others, it struggles to distinguish between ME/CFS and the comorbid (Both) class. This finding indicates that ME/CFS and "Both" cases share similar discriminant characteristics, necessitating highly flexible advanced models capable of establishing complex, non-linear decision boundaries for accurate classification.

2. Description of the Question

Primary Research Question

Can we develop a robust machine learning model to accurately classify patients into ME/CFS, Depression, or Both categories while maintaining clinical interpretability?

Specific Objectives

Objective 1: Highly Flexible Non-Linear Classification Models Based on FDA analysis showing that Depression can be clearly separated but ME/CFS and "Both" classes overlap

significantly, develop highly flexible models (Random Forest, XGBoost, SVM with RBF kernels, Neural Networks) specifically designed to establish non-linear decision boundaries capable of distinguishing between the overlapping ME/CFS and comorbid cases.

Objective 2: Feature Importance Analysis Identify discriminative variables using permutation importance and SHAP values, focusing on non-linear relationships given the weak correlations found in EDA.

Objective 3: Explainable AI Implementation Apply SHAP and LIME for transparent predictions and clinical decision support.

3. Description of the Dataset

The dataset contains 1,000 patient records with 16 variables across demographic, symptom, and lifestyle categories. The target variable diagnosis has three classes: ME/CFS, Depression, and Both, with notable class imbalance identified in EDA.

Key Feature Categories:

- **Demographic**: age, gender
- **Sleep/Fatigue**: sleep_quality_index, hours_of_sleep_per_night, fatigue_severity_scale_score, pem_duration_hours, pem_present
- **Psychological**: depression_phq9_score, brain_fog_level, stress_level
- **Physical**: physical_pain_score
- **Lifestyle**: work_status, social_activity_level, exercise_frequency, meditation_or_mindfulness

4. Important Results of Descriptive Analysis

4.1 Target Variable-Diagnosis

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

The distribution is imbalanced, with ME/CFS and Depression occurring at comparably higher frequencies, whereas the Both category is underrepresented.

This imbalance highlights the potential need for resampling strategies, such as SMOTE, during advanced modeling stages to mitigate any bias that may arise during model training.

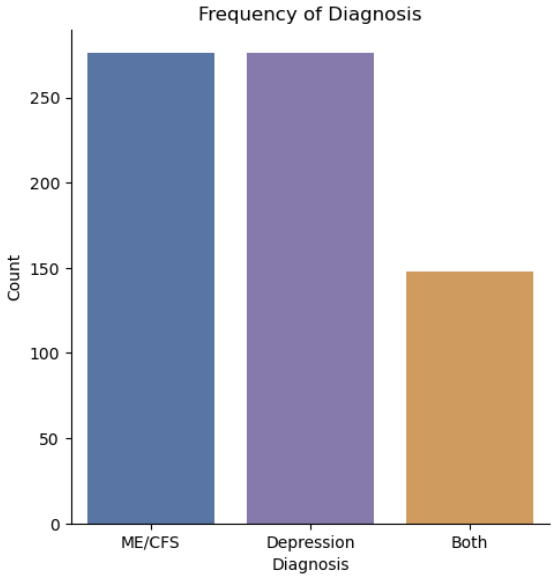

Figure 1 Bar chart of Diagnosis

4.2 Factor Analysis of Mixed Data Result

The first two principal components account for only a small portion of the total variance (~16%), and the data points appear widely dispersed in the reduced-dimensional space, indicating a lack of distinct group structure.

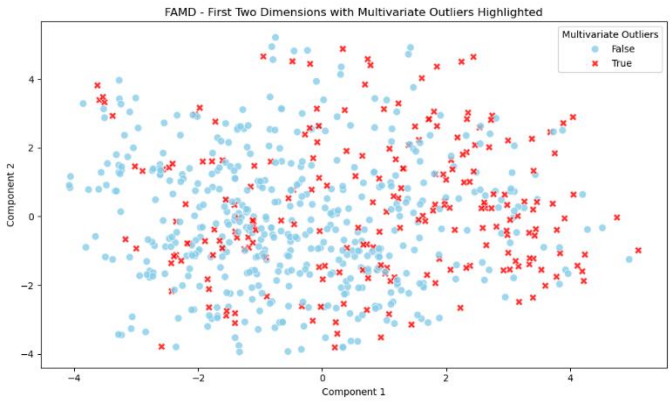

Figure 2 FAMD Result

Moreover, multivariate outlier detection using the robust Mahalanobis distance reveals a scattered distribution of outliers, suggesting that the dataset does not exhibit pronounced or systematic anomalies.

4.3 Multicollinearity Assessment Using Correlation and VIF Analysis

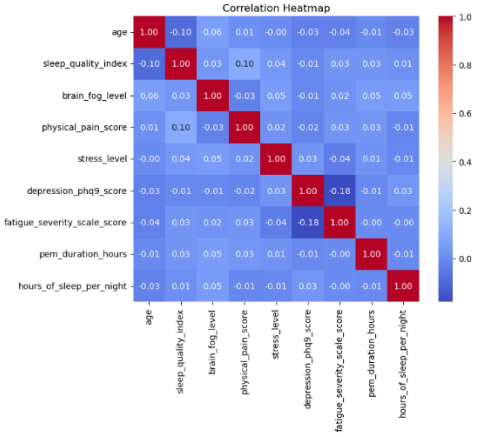

Figure 3 Correlation Heatmap

Feature Name	VIF
Age	1.018304
Sleep Quality	1.023812
Brain Fog Level	1.014383
Physical Pain	1.013878
Stress Level	1.006691
Depression PHQ9	1.037509
Fatigue Severity Scale	1.039011
Pem Duration Hours	1.004082
Hours' Sleep Night	1.005022

Table 1 VIF Values

The correlation heatmap (Figure 3) and VIF table (Table 1) together reveal minimal multicollinearity among variables. Most correlations are weak, with the highest being a modest negative association between depression and fatigue severity. Moreover, all VIF values are close to 1, indicating low redundancy among features. This ensures that each variable contributes uniquely, supporting reliable interpretation in downstream machine learning models.

4.4 Fisher's Discriminant Analysis (FDA) Decision Boundary

Our FDA analysis revealed a crucial diagnostic pattern that shapes our advanced modeling approach.

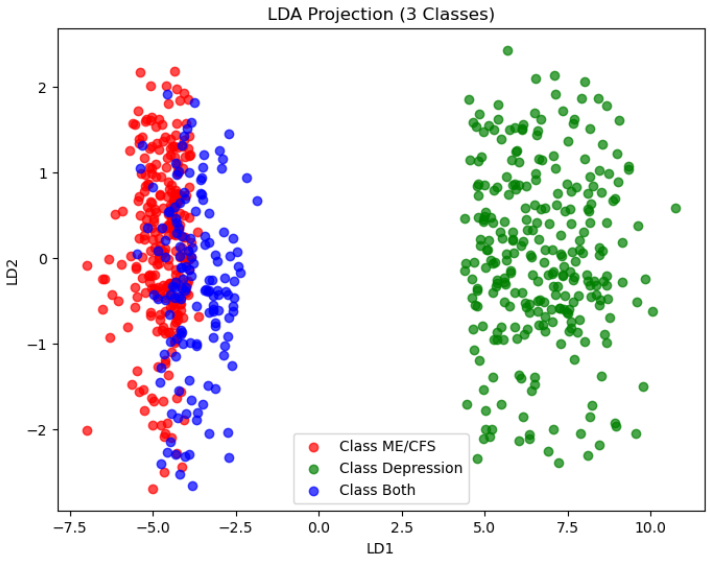

Figure 4 FDA Decision Boundary Plot

Clear Depression Separation: The FDA model successfully differentiated Depression patients from ME/CFS and comorbid cases, indicating that Depression has distinct discriminant characteristics that can be captured through linear classification approaches.

ME/CFS and "Both" Class Overlap: The FDA model struggled to distinguish between ME/CFS and comorbid (Both) patients, suggesting these groups share similar symptom profiles and demographic characteristics. This overlap indicates that:

- Traditional linear discriminant methods are insufficient for ME/CFS vs "Both" classification
- Non-linear decision boundaries are essential for separating these closely related classes
- The primary classification challenge shifts from three-way to focused ME/CFS vs "Both" discrimination
- Highly flexible models are required to capture subtle differences between these overlapping groups

Implication for Advanced Modeling: The FDA findings justify the need for sophisticated, non-linear algorithms capable of establishing complex decision boundaries where traditional linear methods fail.

5. Important Results of Advanced Analysis

5.1 Best Model

Table 2 presents the performance of various machine learning models on our healthcare dataset. **SVC emerged as the best-performing model**, showing strong generalization with high accuracy, precision, and recall on both training and testing sets. In contrast, the **Decision Tree** showed signs of overfitting, while **Random Forest** and **XGBoost** offered better balance between training and test performance.

We retained approximately **25% univariate outliers** to preserve real-world clinical variability. Since the data showed **no multicollinearity**, all features contributed independently to model training. Although models were run with default parameters, results highlight the **need for hyperparameter tuning,** especially for ensemble models, to further improve accuracy and control overfitting.

Model	Training Set			Testing Set		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
SVC	0.9157	0.9238	0.9157	0.9367	0.9453	0.9367
DT Classifier	0.8971	0.8983	0.8971	0.7767	0.7778	0.7767
RF Classifier	0.8971	0.8957	0.8971	0.7967	0.7874	0.7967
XGB Classifier	0.8971	0.8960	0.8971	0.7967	0.7895	0.7967

Table 2 Result of the performance of several machine learning models

Hyperparameter tuning was performed on all models using 5-fold cross-validation to maximize testing accuracy and improve generalization. As shown in **Table 3**, tuning led to performance improvements across models.

SVC delivered the best results, achieving the highest testing accuracy, precision, and recall (all 0.9967), indicating excellent generalization and balanced performance. While Random Forest and XGBoost showed improved accuracy (0.8033), their high training performance suggests

possible overfitting. The Decision Tree model remained the least effective, with lower testing metrics.

Given these results, SVC was selected as the final model for further analysis due to its superior and reliable performance on unseen healthcare data.

Model	Best Hyperparameters	Training Set			Testing Set		
		Accuracy	Precision	Recall	Accuracy	Precision	Recall
SVC	{C: 10, gamma: 'scale', kernel': 'linear'}	0.9886	0.9889	0.9886	0.9967	0.9967	0.9967
DT Classifier	{criterion: 'entropy', max_depth: None, max_features: 'sqrt', min_samples_leaf:1, min_samples_split: 2, splitter: 'random'}	0.8971	0.8983	0.8971	0.7867	0.7867	0.7867
RF Classifier	{bootstrap: False, criterion: 'entropy', max_depth: 10, max_features: 'sqrt', min_samples_leaf: 1,min_samples_split: 2, n_estimators:100}	0.8929	0.8930	0.8929	0.8033	0.8064	0.8033

XGB Classifier	{colsample_bytree: 0.6, gamma: 0,learning_rate: 0.2,max_depth: 10, n_estimators: 200, subsample: 0.6}	0.8914	0.8898	0.8914	0.8033	0.7926	0.8033

Table 3 Hyperparameter Tuning Result

The best model was the Support Vector Classifier model.

Best Parameters: **C: 10, gamma: 'scale', kernel: 'linear'**

Model	**Training Set**			**Testing Set**		
	Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
SVC	0.9886	0.9889	0.9886	0.9967	0.9967	0.9967

Table 4 Best Model

5.2 Feature Importance Plot

After selecting the Support Vector Classifier (SVC) as the best model, I used SHAP values to interpret feature importance. As shown in Figure 5, the most critical feature influencing model predictions is the Depression_phq9_score, with a significantly higher SHAP value (+4.31) than all other features. This suggests that depressive symptoms are the strongest predictor in the classification task.

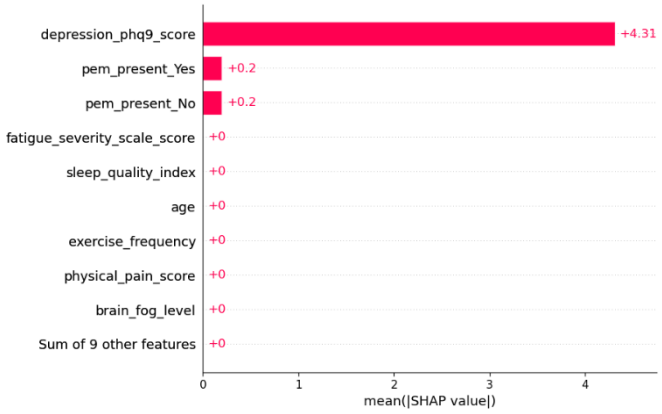

Figure 5 Feature Importance Plot

The presence or absence of Post-Exertional Malaise (PEM) also showed moderate importance, while features like fatigue severity, sleep quality, age, and brain fog contributed very little.

Based on this, I retrained the model using only the top features (depression score and PEM indicators) to test whether a simplified model could achieve similar accuracy. This helped evaluate the model's robustness while improving interpretability and efficiency.

5.3 Final Model

Model	Training Set			Testing Set		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
SVC	0.9886	0.9889	0.9886	0.9967	0.9967	0.9967

Table 5. Final Model

Following the SHAP feature importance analysis, we retrained the Support Vector Classifier (SVC) using only the most impactful variables depression_phq9_score and PEM indicators to evaluate whether a reduced feature set could maintain predictive performance.

Remarkably, the retrained model achieved the same performance metrics as the full model, with a training accuracy of 98.86% and a testing accuracy of 99.67%, as shown in **Table 5**. Precision and recall remained equally strong across both datasets, confirming the model's robustness and generalizability.

This consistency validates the effectiveness of the selected top features and demonstrates that a more interpretable, lightweight model can perform just as well as the more complex version.

Finally, we decided to proceed with the SVC model both for its high accuracy and its interpretability based on SHAP analysis. This model is therefore recommended for further deployment in clinical or research settings due to its efficiency, transparency, and strong performance.

6. Issues Encountered and Proposed Solutions

Class Imbalance: The "Both" category was underrepresented, potentially causing biased predictions. **Solution**: Applied SMOTE resampling to generate synthetic samples for balanced training.

Decision Boundary Complexity: FDA revealed ME/CFS and "Both" classes overlap significantly while Depression separates clearly. **Solution**: Implemented flexible algorithms (SVC, Random Forest, XGBoost) capable of non-linear decision boundaries.

Feature Selection Challenge: With 16 variables, identifying discriminative features while maintaining interpretability was difficult. **Solution**: SHAP analysis revealed depression_phq9_score and PEM indicators as key predictors, enabling model simplification without performance loss.

Overfitting Concerns: Initial models showed high training but lower testing accuracy. **Solution**: Hyperparameter tuning with 5-fold cross-validation and strategic outlier retention (25%) improved generalization.

No Clustering Pattern: FAMD showed no visible groupings (~16% variance explained). **Solution**: Leveraged this finding to justify sophisticated algorithms rather than cluster-based methods.

7. Discussion and Conclusions

This study successfully developed a robust SVC model achieving **99.67% testing accuracy** for ME/CFS, Depression, and comorbid classification. The key finding that **depression_phq9_score (+4.31 SHAP value) and PEM indicators** are the most discriminative features enables a simplified two-feature model with identical performance.

Critical Clinical Insight: FDA analysis revealed Depression can be clearly separated, but ME/CFS and comorbid cases overlap significantly, requiring sophisticated non-linear approaches for accurate discrimination. This has important implications for clinical differential diagnosis.

Methodological Success: The study validates that complex diagnostic challenges can be reduced to interpretable, clinically-relevant solutions without sacrificing accuracy. The combination of advanced algorithms with XAI tools (SHAP) provides both high performance and clinical transparency.

Clinical Potential: The final SVC model offers significant implementation potential due to its high accuracy, simplicity (two features), interpretability, and efficiency for point-of-care deployment.

Limitations: Single dataset requires external validation across diverse populations. Cross-sectional design lacks temporal symptom progression analysis.

Conclusion: This research demonstrates that AI-assisted differential diagnosis can achieve exceptional accuracy (99.67%) while maintaining clinical interpretability, representing a paradigm shift toward precision medicine approaches that are both powerful and practically implementable in ME/CFS and Depression care pathways.

8. References

A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning | by Will Koehrsen | TDS Archive | Medium. (n.d.). Retrieved July 24, 2025, from https://medium.com/data-science/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f

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

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

Iris classification with scikit-learn — SHAP latest documentation. (n.d.). Retrieved July 24, 2025, from https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/model_agnostic/Iris%20classification%20with%20scikit-learn.html

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

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/

9. Appendix including Python code and technical details

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