JESPAR

Journal of Exercise Sciences
& Physical Activity Reviews

ORIGINAL ARTICLE

Prediction of myalgic chronic fatigue syndrome disorder with machine learning approach

Fatma Hilal Yagin[1*] | Georgian Badicu[2]

[1]Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya, Türkiye
[2]Department of Physical Education and Special Motricity, Transilvania University of Brasov, 500068 Brasov, Romania

ABSTRACT

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a complex disorder characterized by unexplained fatigue, post-exertional malaise, unrefreshing sleep, and cognitive impairment or orthostatic intolerance. Due to the absence of a recognized laboratory diagnostic test, diagnosis relies on patient history and physical examination. This study aimed to identify significant metabolomic markers and employ machine learning techniques for the classification of ME/CFS. Utilizing open-access metabolomics data from 26 ME/CFS patients and 26 controls, we implemented a comprehensive data preprocessing and modeling framework. Feature selection was performed using Random Forest, and data normalization was achieved through standardization. A Gaussian Naive Bayes model was trained and validated using 5-fold cross-validation. The model exhibited an accuracy of 0.786, sensitivity of 0.952, specificity of 0.619, and an F1 score of 0.816. These results indicate a high efficacy in identifying positive cases of ME/CFS.

Keywords: Metabolomics, machine learning, random forest, gaussian naive bayes

***Corresponding:** Fatma Hilal Yagin; hilal.yagin@inonu.edu.tr
Journal home page: www.e-jespar.com
Academic Editor: Dr. Mehmet Gülü
https://doi.org/10.5281/zenodo.12601089

ARTICLE HISTORY
Received: 01 May 2024
Accepted: 27 May 2024
Published: 01 July 2024

INTRODUCTION

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is characterized by unexplained fatigue, post-exertional malaise, unrefreshing sleep, and either cognitive impairment or orthostatic intolerance (Encephalomyelitis, 2015). Since there isn't a recognized laboratory diagnostic test, the diagnosis is made based on the patient's medical history, physical examination, and rule out other conditions. A sore throat and cervical lymphadenopathy are common prodromes associated with infection that patients with ME/CFS often report (Encephalomyelitis, 2015). Irritable bowel syndrome (IBS) is thought to affect 35–90% of patients (Aaron et al., 2001; Hausteiner-Wiehle & Henningsen, 2014), as opposed to 10–20% of the general population (Canavan, West, & Card, 2014).

Given that chronic fatigue is a multifactorial disorder, it is unlikely that an objective measure for fatigue conditions will be found, such as a verified single biomarker or biosignature made up of a small number of biomarkers unless the essence of the disorder, its causes, and its pathophysiology are identified. To classify a cohort of patients into eleven categories, ranging from mild to extreme fatigue, however, it has recently been demonstrated (Erasmus, Steffens, Van Reenen, Vorster, & Reinecke, 2019) that data from a battery of conventional tests gave some objective markers to complement clinical and lifestyle data. Furthermore, studies on metabolomics have shed light on conditions similar in complexity to chronic fatigue, such as irritable bowel syndrome (Fourie et al., 2016), fibromyalgia syndrome (Hackshaw et al., 2019), and chronic widespread pain (Freidin et al., 2018). Similarly, the application of metabolomics techniques through intervention research yielded significant insights into experimental investigations concerning acute alcohol intake (Irwin et al., 2018), and nutrition (Wittwer et al., 2011).

Recent developments in computer technology have greatly increased the range of fields in which machine learning is applied, most notably the healthcare industry (Adams, Sepich-Poore, Miller-Montgomery, & Knight, 2022). In order to estimate new data outcomes, machine learning typically entails building predictive models and spotting data patterns (Radakovich, Nagy, & Nazha, 2020). The goal is to derive insights from the data that already exists (Camacho, Collins, Powers, Costello, & Collins, 2018). New approaches to disease categorization and diagnosis can be developed by combining metabolomics data with machine learning (Wu et al., 2022; Zhang et al., 2021). In general, substantial potential for machine learning applications in healthcare has been opened by advances in big data and artificial intelligence technologies. In this study, we aimed to examine the importance of metabolomics factors in ME/CFS patients and to distinguish ME/CFS using these factors.

METHODS

Participant and Data

Open-access metabolomics data from controls and ME/CFS patients were used in this investigation [2]. There were 26 ME/CFS patients and 26 controls among the all-female

participants. Data from plasma samples utilized in the metabolomics panel were acquired for 768 identified compounds.

Data Preprocessing

To preprocess and model the data, we utilized Random Forest for feature selection, normalization techniques, Naive Bayes for modeling, and 5-fold cross-validation for validation, with performance evaluated using accuracy, sensitivity, specificity, and F1 score. Random Forest classifier was trained to identify feature importances, and the top features were selected based on these scores. The data was then normalized using standardization to ensure consistency across features. A Gaussian Naive Bayes (Kamel, Abdulah, & Al-Tuwaijari, 2019) model was subsequently trained on the normalized data. To assess the model's generalizability, 5-fold cross-validation was employed, where the dataset was divided into five parts, and the model was iteratively trained and tested (Fushiki, 2011). The performance of the model was measured using accuracy, sensitivity, specificity, and F1 score, calculated as follows: accuracy as the ratio of correctly predicted instances to total instances, sensitivity as the ratio of true positives to the sum of true positives and false negatives, specificity as the ratio of true negatives to the sum of true negatives and false positives, and the F1 score as the harmonic mean of precision and recall. This comprehensive methodology ensured robust feature selection, data normalization, model training, and performance evaluation.

Statistical Analyses

Univariate statistical analysis was conducted to compare metabolite levels between ME/CFS patients and healthy controls. Normal distribution was examined by the Shapiro-Wilk test. Metabolite levels were summarized as mean ± standard deviation (SD). Differences in metabolite levels were assessed using independent samples t-tests. For each metabolite, a p-value was calculated to determine the statistical significance of the difference between the two groups. A p-value of less than 0.05 was considered statistically significant. Analyses were performed using Python 3.9 software and SPSS 28.0 (IBM Corp., Armonk, NY, United States) package program.

RESULTS

After random forest-based feature selection, the most significant metabolomic features were alphaketobutyrate, hydroxy asparagine, indole-lactate, sarcosine, arachidoyl carnitine(c20), dihomolinolenoylcholine, linoleoylcholine, oleoylcholine, stearoyl choline, gamma-glutamyl valine, leucylglycine, phenylalanylalanine, valylleucine, and dimethyl sulfone. The results of group comparisons for these metabolites are presented in Table 1. In this study, the univariate statistical analysis revealed significant differences in the levels of several metabolites between ME/CFS patients and healthy controls.

The levels of alphaketobutyrate were significantly higher in ME/CFS patients (5246503.808±1772631.799) compared to controls (4017571.269±2185763.392) with a p-value of 0.031. Hydroxyasparagine levels were also elevated in ME/CFS patients

(573425.692±137075.289) versus controls (481279.5±86458.084) with a p-value of 0.006. Conversely, indolelactate levels were lower in ME/CFS patients (2572602.769±884853.405) compared to controls (3347740.808±1084248.796), with a p-value of 0.007. Other metabolites with significant differences included sarcosine, arachidoylcarnitine (c20), dihomolinolenoylcholine, linoleoylcholine, oleoylcholine, stearoylcholine, leucylglycine, phenylalanylalanine, and valylleucine, all showing p-values less than 0.05. The leucylglycine and oleoylcholine were notably decreased in ME/CFS patients, with p-values of 0.008 and 0.006, respectively. These findings highlight distinct metabolic alterations associated with ME/CFS, suggesting potential biomarkers for diagnosis and therapeutic targets.

Metabolite name*	ME/CFS outcome		p value
	No	Yes	
alphaketobutyrate	4017571.269±2185763.392	5246503.808±1772631.799	0.031
hydroxyasparagine	481279.5±86458.084	573425.692±137075.289	0.006
indolelactate	3347740.808±1084248.796	2572602.769±884853.405	0.007
sarcosine	11734957.308±3355435.095	9880927.5±3252648.976	0.048
arachidoylcarnitine(c20)	112077.154±73138.417	155622.115±75991.163	0.04
dihomolinolenoylcholine	143066.231±165173.693	62686.692±33381.024	0.022
linoleoylcholine	1069924.885±1032038.38	490354.885±304684.426	0.01
oleoylcholine	497184.346±463913.821	214579.885±127072.909	0.006
stearoylcholine	455832.731±438699.554	211923.077±138906.689	0.011
gammaglutamylvaline	1118706.731±552336.404	1452248.692±713351.638	0.065
leucylglycine	83965.385±73754.139	40506.077±23791.63	0.008
phenylalanylalanine	207109.115±64803.251	156876.115±47899.172	0.003
valylleucine	151425.346±85363.49	101872.577±43210.332	0.012
dimethylsulfone	1462964.885±2836362.33	549959.231±1307674.775	0.142

*: Metabolite levels are summarized as mean ± SD (standard deviation).

Table 1. Univariate statistical analysis results

The performance metrics results for the model developed for ME/CFS prediction using these biomarker metabolites are presented in Table 2. The model achieves an accuracy of 0.786, indicating that it correctly predicts the ME/CFS approximately 78.6%. The sensitivity, or true positive rate, is 0.952, demonstrating that the model is highly effective at correctly identifying positive instances. However, the specificity is lower at 0.619, meaning the model is less effective at correctly identifying negative instances, with a higher rate of false positives. The F1 score, which balances precision and recall, is 0.816, reflecting a good balance between precision and recall despite the lower specificity. Overall, the model performs well in identifying true positive cases but has room for improvement in reducing false positives.

Metric	Value
Accuracy	0.786
Sensitivity	0.952
Specificity	0.619
F1-score	0.816

Table 2. Performance of machine learning model on ME/CFS prediction

DISCUSSION

The findings of this study underscore the potential of integrating metabolomics data with machine learning techniques to enhance the diagnostic process for ME/CFS. The results of our univariate statistical analysis reveal significant alterations in the metabolite profiles of ME/CFS patients compared to healthy controls, underscoring potential biomarkers and therapeutic targets for this debilitating condition. Elevated levels of alphaketobutyrate and hydroxyasparagine in ME/CFS patients suggest disruptions in amino acid metabolism, which may be linked to the pathophysiology of ME/CFS. Conversely, reduced levels of indolelactate indicate a possible impairment in tryptophan metabolism. The significant differences in choline-containing metabolites, such as dihomolinolenoylcholine, linoleoylcholine, oleoylcholine, and stearoylcholine, point towards dysregulation in lipid metabolism, which could impact cellular membrane integrity and signaling pathways. Additionally, the observed changes in leucylglycine and phenylalanylalanine levels highlight disruptions in peptide metabolism. These metabolic disturbances provide insight into the biochemical underpinnings of ME/CFS and support the hypothesis that ME/CFS is associated with a unique metabolic signature. Further research is necessary to explore the mechanistic pathways involved and to validate these metabolites as reliable biomarkers for ME/CFS diagnosis and treatment monitoring. The application of Random Forest for feature selection proved effective in isolating key metabolomic markers, such as alphaketobutyrate and hydroxy asparagine, which significantly contributed to the model's predictive capability. The high sensitivity of 0.952 indicates that the model is adept at identifying true positive cases, which is crucial for early and accurate diagnosis of ME/CFS. However, the lower specificity of 0.619 points to a higher rate of false positives, highlighting a need for further refinement of the model to better distinguish between ME/CFS and other conditions. The use of 5-fold cross-validation provided a robust framework for model validation, ensuring that the model's performance metrics are generalizable and not overly fitted to the training data. The F1 score of 0.816 reflects a good balance between precision and recall, reinforcing the model's reliability in practical diagnostic settings. Metabolomics has been studied in recent ME/CFS studies to identify possible biomarkers and metabolic anomalies in patients. Numerous metabolites, including sphingomyelins and short-chain fatty acids in plasma, as well as abnormalities in the brain's N-acetylaspartate system, amino acid route, choline, myo-inositol, and lactate, have been linked to ME/CFS in studies. These metabolites contribute to our understanding of the pathophysiology of this intricate illness by drawing attention to metabolic anomalies and possible diagnostic biomarkers for ME/CFS (R Xiong et al., 2021; Ruoyun Xiong et al., 2021; Yamano, Watanabe, & Kataoka, 2021). In the findings of this study, we determined that alphaketobutyrate, hydroxy asparagine, indole-lactate, sarcosine, arachidoyl carnitine(c20), dihomolinolenoylcholine, linoleoylcholine, oleoylcholine, stearoyl choline, gamma-glutamyl valine, leucylglycine,

phenylalanylalanine, valylleucine, and dimethyl sulfone metabolites may be biomarker candidates for ME/CFS. In conclusion, this study demonstrates the feasibility of using machine learning and metabolomics for the classification of ME/CFS, offering a potential pathway towards more objective and reliable diagnostic tools. Continued advancements in computational methods and the availability of larger, more diverse datasets will be critical in refining these models and improving diagnostic accuracy for ME/CFS.

CONCLUSIONS

This study identified metabolomic markers distinguishing Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) from healthy controls, highlighting significant metabolic differences. A Gaussian Naive Bayes model achieved an accuracy of 0.786, demonstrating strong capability in predicting ME/CFS. High sensitivity (0.952) indicates effective detection of ME/CFS cases, albeit with lower specificity (0.619). These findings suggest promising avenues for developing diagnostic biomarkers and personalized treatment strategies for ME/CFS.

Author Contributions

Conceptualization, F.H.Y. methodology, F.H.Y, G.B.; formal analysis, G.B.; investigation, G.B.; data curation, F.Y.H.; writing—original draft preparation, F.H.Y, G.B.; writing—review and editing, F.H.Y, G.B.

Informed Consent Statement:

The research was conducted in line with the Declaration of Helsinki.

Acknowledgments:

We would like to thank all participants who took part in the research.

Funding:

This research was not funded by any institution or organization.

Conflicts of Interest:

The authors declare that no conflicts interest.

REFERENCES

Aaron, L. A., Herrell, R., Ashton, S., Belcourt, M., Schmaling, K., Goldberg, J., & Buchwald, D. (2001). Comorbid clinical conditions in chronic fatigue: a co-twin control study. *Journal of general internal medicine, 16*(1), 24-31.

Adams, E., Sepich-Poore, G. D., Miller-Montgomery, S., & Knight, R. (2022). Using all our genomes: Blood-based liquid biopsies for the early detection of cancer. *View, 3*(1), 20200118.

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell, 173*(7), 1581-1592.

Canavan, C., West, J., & Card, T. (2014). The epidemiology of irritable bowel syndrome. *Clinical epidemiology*, 71-80.

Encephalomyelitis, I. B. M. (2015). Redefining an Illness In: Io Medicine, editor. Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Redefining an Illness. In: Washington (DC): The National Academies Press.

Erasmus, E., Steffens, F. E., Van Reenen, M., Vorster, B. C., & Reinecke, C. J. (2019). Biotransformation profiles from a cohort of chronic fatigue women in response to a hepatic detoxification challenge. *Plos one, 14*(5), e0216298.

Fourie, N. H., Wang, D., Abey, S. K., Sherwin, L. B., Joseph, P. V., Rahim-Williams, B., . . . Henderson, W. A. (2016). The microbiome of the oral mucosa in irritable bowel syndrome. *Gut microbes, 7*(4), 286-301.

Freidin, M. B., Wells, H. R., Potter, T., Livshits, G., Menni, C., & Williams, F. M. (2018). Metabolomic markers of fatigue: Association between circulating metabolome and fatigue in women with chronic widespread pain. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1864*(2), 601-606.

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing, 21*, 137-146.

Hackshaw, K. V., Aykas, D. P., Sigurdson, G. T., Plans, M., Madiai, F., Yu, L., . . . Rodriguez-Saona, L. (2019). Metabolic fingerprinting for diagnosis of fibromyalgia and other rheumatologic disorders. *Journal of Biological Chemistry, 294*(7), 2555-2568.

Hausteiner-Wiehle, C., & Henningsen, P. (2014). Irritable bowel syndrome: relations with functional, mental, and somatoform disorders. *World journal of gastroenterology: WJG, 20*(20), 6024.

Irwin, C., Van Reenen, M., Mason, S., Mienie, L. J., Wevers, R. A., Westerhuis, J. A., & Reinecke, C. J. (2018). The 1H-NMR-based metabolite profile of acute alcohol consumption: A metabolomics intervention study. *Plos one, 13*(5), e0196850.

Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019). *Cancer classification using gaussian naive bayes algorithm.* Paper presented at the 2019 international engineering conference (IEC).

Radakovich, N., Nagy, M., & Nazha, A. (2020). Machine learning in haematological malignancies. *The Lancet Haematology, 7*(7), e541-e550.

Wittwer, J., Rubio-Aliaga, I., Hoeft, B., Bendik, I., Weber, P., & Daniel, H. (2011). Nutrigenomics in human intervention studies: current status, lessons learned and future perspectives. *Molecular Nutrition & Food Research, 55*(3), 341-358.

Wu, J., Xu, M., Liu, W., Huang, Y., Wang, R., Chen, W., . . . Zhou, M. (2022). Glaucoma characterization by machine learning of tear metabolic fingerprinting. *Small Methods, 6*(5), 2200264.

Xiong, R., Gunter, C., Fleming, E., Vernon, S., Bateman, L., Unutmaz, D., & Oh, J. (2021). Multi-'omics of host-microbiome interactions in short-and long-term Myalgic Encephalomyelitis. *Chronic Fatigue Syndrome (ME/CFS), 27*, 2021.

Xiong, R., Gunter, C., Fleming, E., Vernon, S. D., Bateman, L., Unutmaz, D., & Oh, J. (2021). Multi-'omics of host-microbiome interactions in short-and long-term Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS). *bioRxiv*, 2021.2010. 2027.466150.

Yamano, E., Watanabe, Y., & Kataoka, Y. (2021). Insights into metabolite diagnostic biomarkers for myalgic encephalomyelitis/chronic fatigue syndrome. *International journal of molecular sciences, 22*(7), 3423.

Zhang, M., Huang, L., Yang, J., Xu, W., Su, H., Cao, J., . . . Qian, K. (2021). Ultra-fast label-free serum metabolic diagnosis of coronary heart disease via a deep stabilizer. *Advanced Science, 8*(18), 2101333.