Preprints.org

Article

A Proposed Explainable Artificial Intelligence-Based Machine Learning Model for Discriminative Metabolites for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome

Fatma Hilal Yagin [*] , Abedalrhman Alkhateeb , Ali Raza , Nagwan Abdel Samee , Noha F. Mahmoud , Cemil Colak , Burak Yagin [*]

Article

A Proposed Explainable Artificial Intelligence-Based Machine Learning Model for Discriminative Metabolites for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome

Fatma Hilal Yagin [1,*]**, Abedalrhman Alkhateeb** [2]**, Ali Raza** [3]**, Nagwan Abdel Samee** [4]**, Noha F. Mahmoud** [5]**, Cemil Colak** [1] **and Burak Yagin** [1]

[1] Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya, 44280, Türkiye; hilal.yagin@inonu.edu.tr; cemil.colak@inonu.edu.tr; burak.yagin@inonu.edu.tr

[2] Software Engineering Department, King Hussein School for Computing Sciences, Amman, Jordan; a.lkhateeb@psut.edu.jo

[3] Institute of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan; ali.raza.scholarly@gmail.com

[4] Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia; nmabdelsamee@pnu.edu.sa

[5] Rehabilitation Sciences Department, Health and Rehabilitation Sciences College, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia, Nfmahmoud@pnu.edu.sa

* Correspondence: Fatma Hilal Yagin; hilal.yagin@inonu.edu.tr

Abstract: Background: Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a complex and debilitating disease with a significant global prevalence of over 65 million individuals. It affects various systems, including the immune, neurological, gastrointestinal, and circulatory systems. Studies have shown abnormalities in immune cell types, increased inflammatory cytokines, and brain abnormalities. Further research is needed to identify consistent biomarkers and develop targeted therapies. A multidisciplinary approach is essential for diagnosing, treating, and managing this complex disease. The current study aims at employing explainable artificial intelligence (XAI) and machine learning (ML) techniques to identify discriminative metabolites for ME/CFS. **Material and Methods:** The present study used a metabolomics dataset of CFS patients and healthy controls, including 26 healthy controls and 26 ME/CFS patients aged 22-72. The dataset encapsulated 768 metabolites, classified into nine metabolic super-pathways: amino acids, carbohydrates, cofactors, vitamins, energy, lipids, nucleotides, peptides, and xenobiotics. Random forest-based feature selection and Bayesian Approach based-hyperparameter optimization were implemented on the target data. Four different ML algorithms [Gaussian Naive Bayes (GNB), Gradient Boosting Classifier (GBC), Logistic regression (LR) and Random Forest Classifier (RFC)] were used to classify individuals as ME/CFS patients and healthy individuals. XAI approaches were applied to clinically explain the prediction decisions of the optimum model. Performance evaluation was performed using the indices of accuracy, precision, recall, F1 score, Brier score, and AUC. **Results:** The metabolomics of C-glycosyltryptophan, oleoylcholine, cortisone, and 3-hydroxydecanoate were determined to be crucial for ME/CFS diagnosis. The RFC learning model outperformed GNB, GBC, and LR in ME/CFS prediction using the 1000 iteration bootstrapping method, achieving 98% accuracy, precision, recall, F1 score, 0.01 Brier score, and 99% AUC. **Conclusion:** RFC model proposed in this study correctly classified and evaluated ME/CFS patients through the selected biomarker candidate metabolites. The methodology combining ML and XAI can provide a clear interpretation of risk estimation for ME/CFS, helping physicians intuitively understand the impact of key metabolomics features in the model.

1. Introduction

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a complex and debilitating disease. It may come by broad heterogeneity and common symptoms, including severe fatigue, post-exertional malaise (PEM), restless sleep, cognitive impairment, and orthostatic intolerance [1]. The prevalence of ME/CFS is significant, with more than 65 million suffering individuals worldwide, indicating the significant impact of the disease on a global scale [2]. In addition, the true prevalence of the disease is difficult to determine due to factors such as underdiagnoses and misdiagnoses [3]. Although ME/CFS has been observed to be diagnosed more frequently in women, it is not a female-specific condition and approximately 35-40% of patients with ME/CFS are male [4]. The reasons behind the higher prevalence in women are not fully understood [4,5] and may be influenced by a variety of factors such as hormonal differences, genetic predisposition, and social and cultural factors.

Dysfunctions of various systems, including the immune, neurological, gastrointestinal, and circulatory systems, have been reported in individuals with ME/CFS [6–9]. Studies focusing on the immune system have revealed abnormalities in various immune cell types among ME/CFS patients, suggesting that the disease is an immune disorder [6,7]. Increased levels of inflammatory cytokines were also observed in the plasma of ME/CFS patients compared with healthy controls, indicating an increased inflammatory response [10].

Neuroimaging studies have identified abnormalities in the brains of ME/CFS patients, including changes in brain structure and function. These findings add to the understanding of cognitive impairment and other neurological symptoms experienced by individuals with ME/CFS [11]. Digestive problems are common among ME/CFS patients, with a significant proportion reporting symptoms consistent with irritable bowel syndrome (IBS). This suggests that the gastrointestinal tract plays a potential role in the pathophysiology of the disease [12,13].

The circulatory system plays a very important role in providing essential compounds and removing metabolic wastes from various organs [14]. Several studies have been conducted to characterize the blood metabolome of ME/CFS patients to gain insight into the underlying causes of the disease and to establish diagnostic strategies [14]. These studies have highlighted differences in amino acids, lipids, and imbalances in energy and redox metabolisms. However, it is important to note that no consistently altered metabolites were identified in all studies, which poses a challenge to a full understanding of the disease.

The surprising nature of ME/CFS, in which multiple organ systems are affected [6–14], underlines the complexity of the disease and the need for further research. ME/CFS is a heterogeneous condition, and individual variations in symptoms and underlying mechanisms may contribute to the difficulty in identifying consistent biomarkers or metabolic changes. More comprehensive and collaborative research efforts are required to uncover the underlying mechanisms for ME/CFS, identify reliable biomarkers, and develop targeted therapies. The involvement of multiple organ systems highlights the importance of a multidisciplinary approach in the diagnosis, treatment, and management of this complex disease. In this study, we comprehensively analyzed the metabolites of ME/CFS patients compared to normal controls to identify patterns of metabolites that could potentially serve as biomarkers for the disease. What makes our analysis comprehensive is that we examined metabolites belonging to nine different super pathways, aiming to address the heterogeneous nature of the disease and understand its mechanisms of development and progression. To achieve this, we employed a combination of explainable artificial intelligence (XAI) methodology combined with machine learning (ML). This methodology enabled us to identify discriminative metabolites for ME/CFS.

3

2. Materials and Methods

ME/CFS Metabolomics Dataset

The metabolomics data of CFS patients and healthy controls were utilized to perform the experiments in the study [2]. All of the participants were female and consisted of 26 healthy controls and 26 ME/CFS patients aged 22 to 72 years and with similar body mass index (BMI). Data for 768 different metabolites identified were obtained from the plasma sample used in the global metabolomics panel. According to the standards set by Metabolon®, the detected substances are further classified into nine different metabolic super-pathways. The distribution of identified compounds is as follows: amino acids 196, carbohydrates 25, cofactors and vitamins 29, energy 10, lipids 259, nucleotides 33, partially defined molecules 2, peptides 33, and xenobiotics 181 (Supplementary File 1 and 2).

Experimental Setup and Proposed Framework

The Python programming language was used to perform the research experiments. The experiments were conducted in an environment containing a graphics processing unit (GPU) backend with 16GB of RAM and 90GB of disk space. An architectural representation of the proposed methodology is depicted in Figure 1. Diagnosis and biomarker discovery of patients suffering from ME/CFS and healthy controls form the basis of the proposed study. Below is a step-by-step description of the proposed methodology:

- The first step involves obtaining metabolomics data to be used in experiments. Metabolomics data are based on results from a study of 26 healthy controls and 26 ME/CFS patients aged 22 to 72 years with similar body mass index (BMI).
- In the second step, artificial intelligence-based random forest (RF) feature selection is applied to identify biomarker candidate metabolites and to eliminate the high dimensionality problem in omic data. Because the metabolomics data has a large number of feature dimensions, the performance scores of the predicted models may be lower. Therefore, the twenty most important metabolites contributing to improved performance scores in ME/CFS prediction were identified.
- In the third step, 80%-20% split, 5-fold cross-validation (CV), and 1000 replicates Bootstrap approaches were used to validate the prediction models to be generated using the selected biomarker candidate metabolites, and the results were compared.
- In the fourth step, Bayesian hyper-parameter optimization was used to determine the optimal parameters.
- In the fifth step, predictive models were built to diagnose ME/CFS patients. For this purpose, Gaussian Naive Bayes (GNB), Gradient Boosting Classifier (GBC), Logistic regression (LR), and Random Forest Classifier (RFC) algorithms were constructed. Performance of the models was evaluated via area under (AUC) Receiver operating characteristic (ROC) Curve, Brier score, accuracy, precision, recall, and F1-score. While the primary purpose of the methodology is biomarker discovery and diagnosis of ME/CFS, an important secondary purpose is to provide users with indicative probability scores. Therefore, we evaluated the quality of the probabilities via a calibration curve and by calculating the Brier score.
- Finally, XAI approaches were applied to the proposed model to provide transparency and interpretability to the model and to explain the decisions made by the model. Through the use of XAI, we can grasp both the rationale and the process behind a particular decision made by the proposed model.
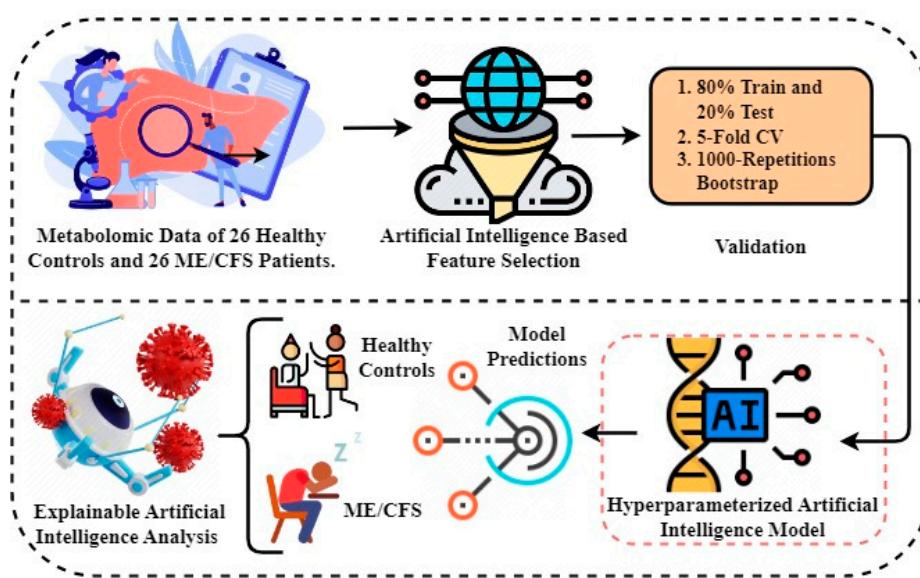
Figure 1. The proposed methodology architecture analysis for detecting healthy individuals and ME/CFS patients.

Feature selection

For feature selection and dimensional reduction from the utilized metabolomics data, the RF method, which is based on artificial intelligence, has been applied in this study. The mean decrease impurity (MDI) method is commonly utilized to carry out the process of choosing features that are included in the random forest model. The impurity of the decision trees in the forest is used as a factor in the calculation of the important score for each feature. This score is based on the average amount that each feature reduces the impurity of the decision trees. The feature importance score has been normalized in such a way that the total of all features important values is equal to 1. After that, the most important features with the highest scores are chosen to be used for training the models that are being applied. The RF method of feature selection can be mathematically represented as follows:

$$\text{Feature Importance}$$
$$= \frac{1}{n_{trees}} * \sum_{t=1}^{n_{trees}} \sum_{i=1}^{n_{nodes}} I_{v_i=f} * \left(\frac{N_t}{N}\right) * (impurity_{parent} - impurity_{childern})$$

Where:

ntrees is the number of decision trees in the random forest.

vi is the feature used for the split at node i of the t-th tree.

f is the feature being evaluated for importance.

Ivi = f is an indicator function that equals 1 if vi = f and 0 otherwise.

Nt is the number of samples in the t-th tree that reaches node i.

N is the total number of samples in the training set.

impurityparent is the impurity of the set of samples at the parent node i.

impuritychildren is the weighted impurity of the two sets of samples after the split based on feature f.

Validation Methods

To evaluate the distinctive performance and calibration quality of our ML methodology on metabolomics ME/CFS data, in addition to the two classical approaches, hold-out, and k-fold cross-validation, we also used the bootstrap sampling described by Steyerberg et al [15].

Hold-out Validation: It is the most basic validation method used for ML algorithms and is used to divide the dataset into two as training and test datasets. The training dataset, the training test dataset of the ML model, is used to evaluate the predictive performance of the model [16].

k-fold Cross-validation: A statistical method for evaluating and comparing learning algorithms, in k-fold cross-validation, the data is first divided into k folds, each of which has a size that is equal to or very close to being equal to the others. Following this, k iterations of training and validation are carried out in such a way that, within each iteration, a different fold of the data is hold-out for validation while the remaining k minus one folds are used for learning [17].

Bootstrap Validation: The Bootstrap resampling method is a way to predict the fit of a model to a hypothetical test set when an explicit test set is not available. It helps to avoid overfitting and improves the stability of ML algorithms. In this validation method, a set of artificial new datasets is "bootstrapped" by random sampling, replacing the original dataset. Each ML model was then trained on the sampled dataset and evaluated on the original dataset. This process was repeated 1000 times [18].

The Bayesian Approach for hyper-parameter optimization

The effectiveness of every ML model is determined by the hyper-parameters associated with that model. They have influence over the learning process or the structure of the statistical model that lies beneath the surface. On the other hand, there is no standard approach to selecting hyper-parameters in real experiments. As a substitute, practitioners frequently set hyper-parameters through a process of trial and error or occasionally allow them to remain at their default settings, both of which result in inadequate generalization. By recasting it as an optimization problem, hyper-parameter optimization gives a methodical approach to solving this issue. According to this line of thinking, a good set of hyper-parameters should (at the very least) minimize a validation error. When compared to the vast majority of other optimization problems that can arise in machine learning, hyper-parameter optimization is a nested problem. This means that at each iteration, an ML model needs to be trained and validated.   Many approaches have been developed to discover the optimal combination of ML model hyper-parameters. Grid search and random search are two optimization approaches that are often employed for this purpose. These strategies, however, have a few drawbacks. Grid searching is a time-consuming and inefficient strategy for the central processing unit (CPU) and graphics processing unit (GPU). The grid search strategy outperforms random search; nevertheless, the exact answer is more likely to be ignored. In comparison to these two strategies, Bayesian optimization is the best choice for searching for hyper-parameters. First, because the Gaussian process is involved, the Bayesian optimization technique may consider prior results. To put it another way, each step computation may be retrieved to assist in determining a better set of hyper-parameters. Second, compared to other methodologies (for example, grid search), Bayesian optimization takes fewer iterations and has a quicker processing time. Finally, even when working with non-convex issues, Bayesian optimization may be trusted [19–22].

Classification models

To identify patients into two categories, namely ME/CFS and healthy individuals, we made use of a variety of AI-based classification algorithms in this work. These included Gaussian Naive Bayes (GNB), Gradient Boosting Classifier (GBC), Logistic regression (LR), and Random Forest Classifier (RFC).

GNB: The GNB algorithm is a well-known classification method that is frequently utilized in the field of biomedical research to categorize various patient groups. GNB can be used to properly diagnose patients based on specific physiological traits or biomarkers in the case of healthy individuals as well as patients with ME/CFS. The Bayes theorem, which asserts that the probability of a hypothesis may be computed based on the probability of observing specific evidence, serves as the foundation for GNB's mathematical operation. This theorem underpins how GNB works. When using GNB, it is assumed that the conditional probability of each feature given the class is Gaussian, which indicates that the features are regularly distributed within each class. This is done to comply

with the requirements of the GNB algorithm. This assumption makes the computation of the posterior probability much easier, which in turn enables classification that is both more efficient and more accurate. The GNB method works by first determining the posterior probability of each class for a certain set of features, and then designating the class that has the highest probability as the class that will be predicted [23,24]. The following mathematical notations can be used to express GNB:

$$P(y|x) = P(x|y) * \frac{P(y)}{P(x)}$$

Where:

P(y|x) is the posterior probability of class y given input vector x.

P(x|y) is the likelihood of the input vector x given class y, modeled as a multivariate Gaussian distribution.

P(y) is the prior probability of class y, estimated as the relative frequency of y in the training set.

P(x) is the evidence or marginal likelihood of the input vector x, calculated as the sum of the joint probabilities of x and all possible classes y.

GBC: The GBC is a powerful ML algorithm that has shown great potential in the classification of healthy individuals and patients with ME/CFS. GBC operates by iteratively constructing an ensemble of weak prediction models, typically decision trees, and combining their outputs to make accurate predictions. During the training phase, GBC builds the ensemble by initially fitting a weak model to the training data. Subsequent models are then constructed in a way that each new model focuses on the instances that were previously misclassified by the ensemble [25,26]. The mathematical notations for the GBC model for classification are as follows:

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i)$$

Where:

$\hat{y}_i$ represents the predicted value for the i-th instance.

M denotes the number of weak classifiers (decision trees) used in the GBC.

$f_m(x_i)$ refers to the m-th weak classifier's prediction for the i-th instance.

LR: For binary classification problems such as disease categorization using patient data, LR is a common machine learning model. Given input data including patient demographics, symptoms, and laboratory test results, the LR model calculates the likelihood of a positive class. To maximize the likelihood of the positive class, the model learns the ideal set of weights or coefficients by minimizing the logistic loss function. To produce a probability between 0 and 1, the logistic function uses a linear combination of input features and their weights. Afterward, a threshold (such as 0.5) is used to the anticipated probability to determine the expected class; if the predicted probability is greater than the threshold, the positive class is predicted, and vice versa [27,28]. Here are some mathematical symbols for the LR model of binary classification:

$$p(y = 1|x, \theta) = \frac{1}{(1 + e^{-(\theta_0 + \theta_{x_1} + \theta_{x_2} + \cdots + \theta_{xp})})}$$

Where:

p(y=1|x,θ) is the predicted probability of the positive class given the input feature vector x and the model parameters θ.

e is the base of the natural logarithm (approximately 2.718).

θ0 is the intercept or bias term.

θx1, θx2, ..., θx0 are the coefficients or weights of the input features x1, x2, ..., xp

x = [x1, x2, ..., xp] is the input feature vector

RFC: The RFC, is a well-known technique for machine learning that is used for classification tasks, such as the classification of diseases based on patient data. Building an ensemble of decision trees that have been trained on random subsets of the input features and data samples is how RFC

goes about doing its work. Each decision tree in the ensemble makes a prediction based on a subset of the input features, and the final prediction is generated by aggregating the predictions of all of the trees in the ensemble. RFC can handle high-dimensional data with a large number of features and can also capture nonlinear correlations between the input features and the output classes [29–31]. The RFC equation can also be written as follows in its mathematical notation:

$$y = f(X)$$

Where:

X is an input data matrix with n samples and p features, where X = [x1, x2, ..., xn], and each xi is a vector of p features.

y is a vector of predicted class labels, where y = [y_1, y_2, ..., y_n].

f(X) is the function that maps the input data X to the predicted class labels y using a random forest model.

Performance Evaluation and Model Calibration

Performance Evaluation

Accuracy: Accuracy refers to the correct classification rate of a classification model. The accuracy score is calculated as the ratio of correctly guessed samples to the total number of samples. However, in the case of unbalanced classes or misclassification costs, the accuracy score alone may be insufficient and should be evaluated in conjunction with other metrics [32].

Precision: The precision score expresses how many of the positively predicted samples are actually positive. The precision score is calculated as the ratio of the number of false positives (False Positive) to the total number of positive predictions (True Positive + False Positive). The higher the precision score, the better the positive predictions of the model are [32].

Recall: The recall score expresses how many of the true positives (True Positive) are correctly estimated. The recall score is calculated as the ratio of the number of false negatives (False Negative) to the total number of true positives (True Positive + False Negative). The higher the recall score, the better the model captures true positives [32].

F1 score: The F1 score is calculated by taking the harmonic mean of the precision and recall scores. It is preferred to the harmonic mean because it provides the balance between precision and recall scores. The higher the F1 score, the higher the model classifies with both high precision and high recall [32].

ROC Curve and AUC: Evaluation of diagnostic tests is a topic of interest in contemporary medicine, and this is true not only for determining whether or not a disease is present in a patient but also for determining whether or not healthy people have the disease. The conventional method of diagnostic test evaluation employs sensitivity and specificity as measures of accuracy of the test in comparison with gold standard status. This method is used in diagnostic tests that have a binary outcome, such as positive or negative results from the test. In a scenario in which the test results are recorded on an ordinal scale (for example, a five-point ordinal scale: "definitely normal," "probably normal," "uncertain," "probably abnormal," and "definitely abnormal"), or in a scenario in which the test results are reported on a continuous scale, the sensitivity and specificity can be computed across all of the possible threshold values. Therefore, the sensitivity and specificity change throughout the different thresholds, and there is an inverse relationship between sensitivity and specificity. Then, the receiver operating characteristic (ROC) curve is called the plot of sensitivity versus 1-Specifity, and the area under the curve (AUC), as a reliable indicator of accuracy has been considered with relevant interpretations. ROC curves are plotted as sensitivity versus 1-Specifity. This curve is extremely important when determining how well a test can differentiate between different types of people and their actual conditions. A ROC curve is formed when the sensitivity vs specificity is plotted against each other across a range of cutoffs. This plot forms a curve in the unit square. In "ROC space," the ROC curves that correspond to diagnostic tests with progressively stronger discriminant capacity are situated gradually closer to the upper lefthand corner. The area under the curve is a statistic that provides a comprehensive overview of the ROC curve rather than focusing on

a single point of operation. AUC represents the area under the ROC curve and takes a value between 0 and 1. The AUC value measures the discrimination ability of the classification model. A high AUC value means that the model can discriminate well and has a high sensitivity and low false positive rate. The closer the AUC value is to 1, the better the model's performance [33–35].

Model Calibration

A well-calibrated model is one in which the estimated probability matches the true incidence of the outcome. For example, approximately 90% of patients with an estimated risk of ME/CFS of 0.9 would be classified as ME/CFS. This is critical for prediction models because clinical decision-makers need to know how confident the model is in making a particular prediction. Therefore, we calibrate the trained model to get the correctly predicted probability. In this article, we use the Brier score and Calibration curve for model calibration [36,37].

Brier Score: The Brier score is a metric used to evaluate the quality of probability estimates. It is especially used for probabilistic classification models. The Brier score provides a measure of the mean squared errors between the actual labels and the estimated probabilities. The lower the Brier score, the closer the predictions are to reality [36,37].

Calibration Curve: A calibration curve is a tool used to evaluate how close a classification model's estimates are to the true probabilities. This curve shows the accuracy of the probabilities predicted by the model. The calibration curve is important to determine the confidence level of the model and to evaluate the reliability of the predictions. A well-calibrated model means that high-probability predictions are more likely to happen, while low-probability predictions should be less likely to happen. It is verified that the probabilities predicted by a well-calibrated model are consistent with the realization rates [36,37].

XAI Approach

Interpretability is absolutely essential when using a complex ML model in a real-world environment such as the medical field. XAI is an emerging research area that aims to increase the interpretability and transparency of applied ML models. XAI ensures that decisions made by applied models are understood and trusted, especially in critical applications such as healthcare. XAI techniques can help users understand, validate, and trust the decisions made by these models in real-world applications [38,39]. In this research, Shapley values and the Treemap approach were used to interpret the estimation decision of the optimal ML model.

Shapley Additive Explanations (SHAP): It is an approach used to understand the contribution of each feature to the prediction to explain the predictions of SHAP ML models. This approach also takes into account the complexity of the ML model and the interactions between features that go into the model. It also measures the contribution of a feature to the prediction using Shapley values and thus produces graphical results for understanding the model's decisions [39].

TreeMap: TreeMap provides an intuitive description for tree-based ML models, showing the name of the property used for each decision level and the split value for the condition. If an instance satisfies the condition, it goes to the left branch of the tree, otherwise, it goes to the right branch. When purity is high in TreeMap, the knuckle/leaf has a darker color. The samples row at each node shows the number of samples examined at that node [40,41].

3. Results

In this section, firstly, the results of biomarker candidate metabolites are given and the performance results of the applied predictive artificial intelligence algorithms are evaluated using various evaluation metrics. Predictive models were constructed based on both the original data and the metabolites identified as biomarker candidates, and the results were compared. The model showing the final performance was used for ME/CFS estimation and the decision-making function of the model was tried to be explained using XAI approaches.

Feature Selection Results

Figure 2 depicts the features that were selected together with their respective relevance ratings, which were determined using a random forest method based on machine learning. The results reveal that the metabolomics of C-glycosyltryptophan, oleoylcholine, cortisone, and 3-hydroxydecanoate are extremely relevant for the diagnosis of ME/CFS patients.
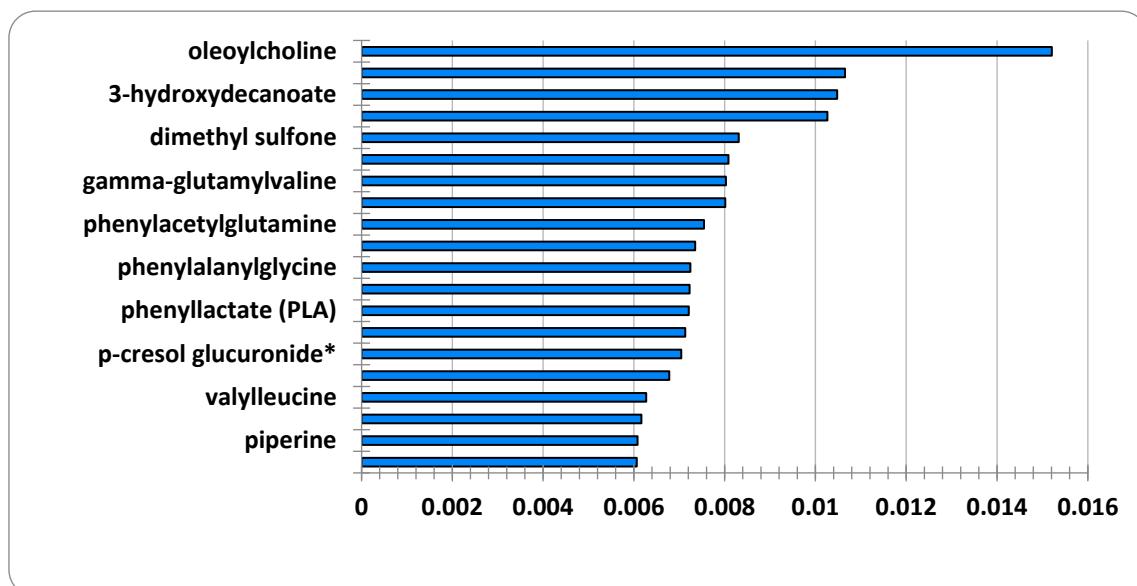

Figure 2. The histogram-based feature importance plot of selected features using the RF model.

Hyper-parameters Optimization Results

Table 1, optimal hyper-parameters of ML models according to the Bayesian optimization are given.

Table 1. The hyper-parameter tuning analysis of applied methods.

Technique	Optimized Parameter Value
GNB	var_smoothing=1e-9.
GBC	n_estimators=3, learning_rate=1.0, max_depth=1, random_state=0.
LR	random_state=0, max_iter=30, solver='liblinear'.
RFC	max_depth=26, min_samples_leaf=5, min_samples_split=3, n_estimators= 12.

The Model Performance Results

In this section, we show how selecting metabolic traits associated with ME/CFS can help learning models improve their performance. After training using all input features and a subset of them (significant features), the results of all used models (GNB, GBC, LR, RFC) are presented in Table 2. We also performed three experiments for model validation, in the first experiment the dataset was split into 80% and 20% to train and validate the learning models. For the second experiment, we used the cross-validation method during the training and validation of learning models. Finally, we used the 1000 iteration Bootstrap method in our last experiment. Bootstrap is a resampling method in which parts are changed at each iteration of the sampling process. This creates a randomly selected collection of samples from the set of input samples. This procedure can be performed k times. Models were trained on the sampled dataset and evaluated on the original dataset. In all experiments, we calculated the performance of all learning models with and without a feature selection step. After the three experiments outlined in this section, the results of each learning model were accuracy (A), precision (P), recall (R), F1-Score (F1), Brier Score (B), and AUC.

According to Table 2, in models using the original features in the dataset, it was determined that the GBC model achieved the lowest performance scores (accuracy: 36%; AUC: 33%; Brier score: 0.63) by dividing the data by 80-20. As a result of Bootstrap with the original features, the LR model had the best performance (accuracy: 96%; AUC: 95%; Brier score: 0.04). All results for models using biomarker candidate metabolites showed improved prediction performance when compared to models using the original features. The results of the investigation show that the performance metrics scores of all of the used machine learning approaches for diagnosing healthy controls and ME/CFS patients were much improved by applying selected biomarker metabolites. The interpretation was also more likely for these models, which took into account fewer risk factors. The Bootstrap validation method gave superior results compared to the first two experiments (80-20 split and 5-fold CV) both in experiments using the original metabolomics variables and in models using twenty biomarker candidate metabolites. It was determined that the RFC learning model outperformed the other three models (GNB, GBC, and LR) with the 1000 iteration bootstrapping method for ME/CFS prediction based on a few metabolite markers. The RFC learning model achieved 98% accuracy, 98% precision, 98% recall, 98% F1 Score, 0.01 Brier score, and 99% AUC.

Table 2. The comparative performance analysis of the applied artificial intelligence techniques with different approaches.

	Attained performance using all input features						Attained performance using feature selection					
Technique	**A (%)**	**P (%)**	**R (%)**	**F1 (%)**	**B**	**AUC (%)**	**A (%)**	**P (%)**	**R (%)**	**F1 (%)**	**B**	**AUC (%)**
	80%-20% split validation						**80%-20% split validation**					
GNB	73	72	73	72	0.27	67	73	72	73	72	0.27	67
GBC	36	39	36	37	0.63	33	73	75	73	73	0.27	73
LR	64	64	64	64	0.36	60	73	72	73	72	0.27	67
RFC	45	56	45	44	0.54	51	82	86	82	80	0.18	75
	Results with 5-folds cross validation						**Results with 5-folds cross validation**					
GNB	52	36	94	62	0.26	59	82	77	92	84	0.15	91
GBC	48	47	35	37	0.34	52	95	94	99	95	0.05	98
LR	58	46	71	54	0.45	46	95	95	96	96	0.03	98
RFC	56	68	38	56	0.28	64	97	96	97	98	0.04	99
	Results with 1000 repetition bootstrap						**Results with 1000 repetition bootstrap**					
GNB	63	70	63	60	0.36	63	83	84	83	83	0.17	91
GBC	92	92	92	92	0.07	92	96	96	96	96	0.03	92
LR	**96**	**96**	**96**	**96**	**0.04**	**95**	96	96	96	96	0.04	99
RFC	90	90	90	90	0.09	90	**98**	**98**	**98**	**98**	**0.01**	**99**

GNB: Gaussian Nave Bayes; GBC: Gradient Boosting Classifier; LR: Logistic Regression; RFC: Random Forest Classifier; A: accuracy; P: precision; R: recall; B: Brier score; AUC: Area under the ROC Curve.

The ROC area reached by each learning model after training on the selected biomarkers is shown in Figure 3. The better the performance of the prediction model, as measured by the ROC curve, the closer the value of the AUC is to one. As can be seen in Figure 3, the RFC model reached its highest AUC value of 99%.
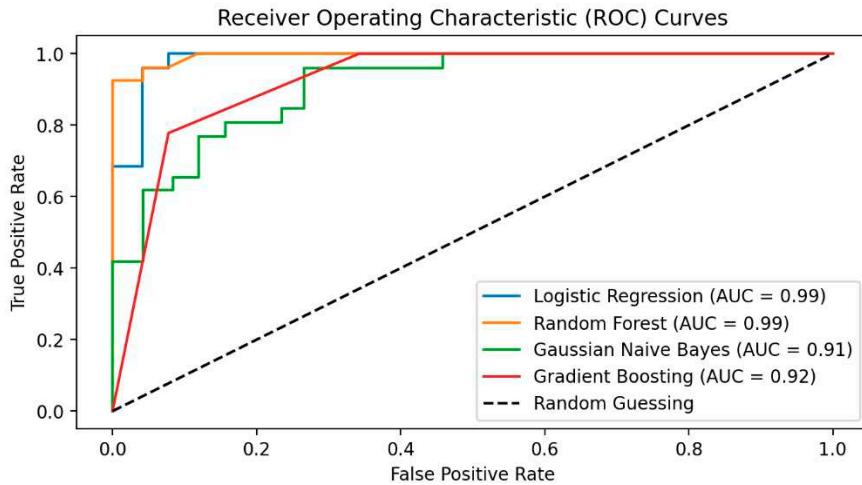
Figure 3. The attained AUC of all ML models after being trained/validated using the biomarker metabolites.

It is prevalent in classification to seek out both a prediction of the class label and the probability of that label. By examining these possibilities, the diagnostic decision of the learning model can be more relied upon. Therefore, we have drawn the RFC model's calibration curve, as depicted in Figure 4, to guarantee its accuracy. To calibrate the accuracy of predictions, the calibration process compares the actual label frequency to the expected label probability. A closer alignment of dots along the major diagonal of the plot indicates a more accurate calibration or more trustworthy prediction.
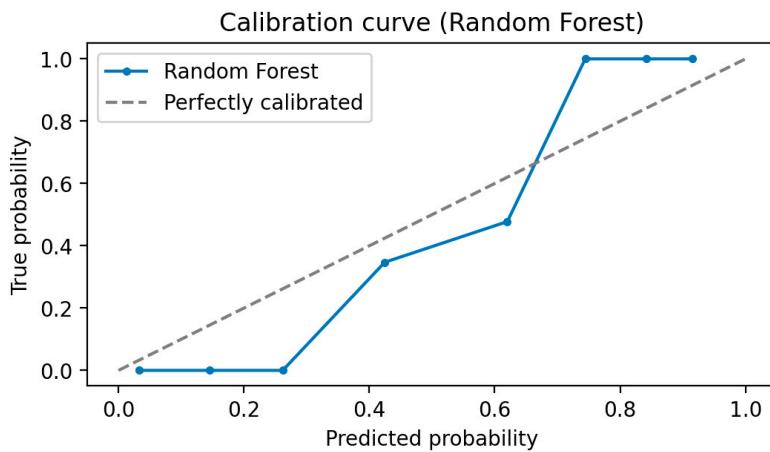
Figure 4. The calibration curve analysis of outperformed RFC model.

XAI Results

The SHAP was used to identify metabolomics biomarkers according to their importance or contribution to the prediction of ME/CFS and to explain the prediction decisions of the model. The RFC-trained model was subjected to SHAP annotation and identified the most important trait metabolites responsible for the prediction of ME/CFS. The results pointed to a list of metabolites with importance scores. Metabolite biomarkers are arranged in decreasing order of importance. Oleoylcholine, phenylactate (PLA), octanoylcarnitine (CB), hydroxyasparagine**, and piperine are among the most prominent metabolite biomarkers important in the diagnosis of ME/CFS. Figure 5 also visualizes the relationships between the relative value of biomarker candidate metabolites and the SHAP values for these metabolites. In each row of the graph, each patient is marked as a dot. The horizontal position of the dot reflects the SHAP values, and the color of the dot encodes the relative value of the metabolites and their mean in the dataset. For example, low values (relatively blue) of

Oleoylcholine and phenylactate (PLA) metabolites contribute positively to ME/CFS, thus increasing disease risk (Figure 5).
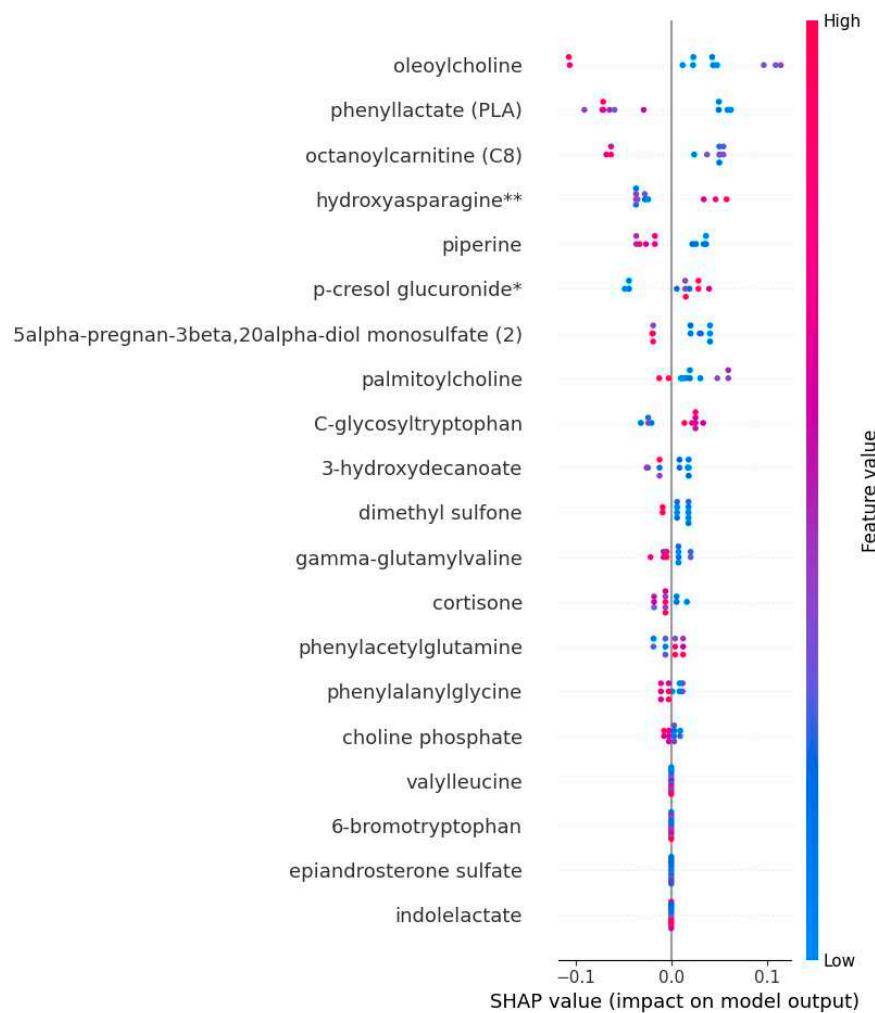

Figure 5. The explainable impact of the proposed RFC model output on biomarker metabolite features**.**

In addition to that, to gain an understanding of how the RFC model behaves, we have employed a method that is known as Treemap analysis. Figure 6 is an illustration of the Treemap that is included in the RFC. The analysis explains how the proposed model came to its conclusion about the classification of patients as healthy controls and those with ME/CFS.
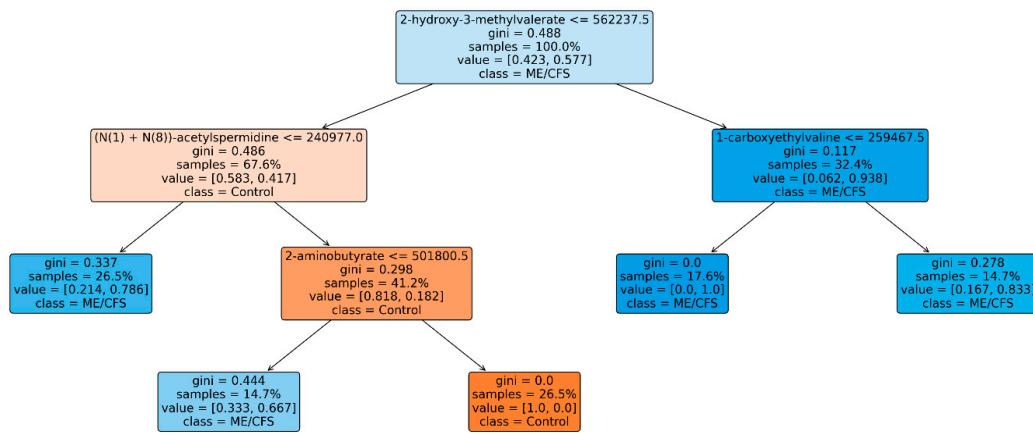
Figure 6. The treemap space analysis of the proposed RFC model.

4. Discussion

Fatigue is a common occurrence in human beings and serves as an indicator of disrupted homeostasis within the body, resulting from either excessive physical and mental exertion or illness [42]. In addition to being one of the most significant social concerns, chronic fatigue additionally constitutes the most significant economic losses [43]. Pain, cognitive dysfunction, autonomic dysfunction, sleep disturbance, and neuroendocrine and immune symptoms are just some of the many symptoms associated with ME/CFS [44]. A patient must have a symptom from neurological impairments, an immune/gastrointestinal/genitourinary impairment, and an energy metabolism/transport impairment to be diagnosed with ME/CFS and meet the criteria for post-exertional neuroimmune exhaustion. However, the strength and severity of such symptoms in a patient vary and are heterogeneous, from moderate to severe, with some patients even becoming bed-bound [44]. Because it is challenging to identify the typical abnormal elements for this disorder utilizing general and conventional medical examination, artificially intelligent-based automated methods may aid in improving the diagnosis of ME/CFS. In recent years, a growing number of research have explained the pathology of ME/CFS and have established biomarkers for the same by employing a metabolome analysis technique [45–47]. This has allowed for the development of a variety of diagnostic studies [48].

The present investigation of the effectiveness of methodology combining ML and XAI techniques to investigate biomarkers of ME/CFS and develop an interpretable predictive model for disease diagnosis. Metabolomics data from patients diagnosed with ME/CFS and healthy controls were used. The classification algorithms employed include GNB, GBC, LR, and RFC. The classifiers' performance has been evaluated both with and without the implementation of the feature selection algorithm (RF). In addition to classical hold-out validation, cross-validation, and bootstrap approaches were also used to evaluate the performance of classification learning algorithms in the validation stage, and the effectiveness of these three validation approaches was also examined. Shapely values, an explainable AI system, have been utilized to interpret the classification models' predictions and decisions. After being trained and validated on the significant selected features using the Bootstrapping method, the RFC model is found to be superior to the other four models (GNB, GBC, and LR). Accuracy, precision, recall, F1 Score, and the AUC were all at or above 98% for the RFC model. The higher the values that attain for precision and sensitivity, the higher the proportion of correct diagnoses, also known as True Positives (TP), and the lower the value of false negatives (FN). Errors both positive and negative, known as false positives (FP) and FN, are widespread in comparative biology research. In addition, we demonstrated that our method was capable of demonstrating the main features as well as the interpretations of ML findings by utilizing SHAPley values and SHAP plots. The SHAP method's findings indicated that oleoylcholine, phenyllactate,

octanoylcarnitine, hydrooxyasparagine, piperine, p-cresol glucuronide, and palmitoylcholine are all chemicals associated with ME/CFS and crucial to the model's final decision. The use of the SHAP technique revealed that the indolelactate, which has low Shapley values, is the least significant of all the features. On the other hand, the feature with the highest Shapley value is the oleoylcholine, which is also the one that contributes the most significant information for the diagnosis of ME/CFS. Oleoylcholine is a member of the class of chemical compounds known as acyl cholines. Germain et al. [2] researched the metabolic pathways that influence the diagnosis of ME/CFS patients by performing statistical analysis in conjunction with pathway enrichment analysis. They found that acyl cholines, which are part of the sub-pathway of lipid metabolism known as fatty acid metabolism, are consistently reduced in two different patient cohorts that suffer from ME/CFS. Nagy-Szakal et al. [49] have gained insights into ME/CFS phenotypes through comprehensive metabolomics. Biomarker identification and topological analysis of plasma metabolomics data were performed on a sample group consisting of fifty ME/CFS patients and fifty healthy controls. They have demonstrated that patients with ME/CFS have higher plasma levels of ceramide and observed that there is a variation in the level of carnitine, choline, and complex lipid metabolites. The study of plasma metabolomics data attained a more accurate prediction model of ME/CFS (AUC = 0.836).

A comprehensive metabolomics was conducted by Naviaux et al. [49] to better understand the biology of CFS. They have investigated 612 plasma metabolites across 63 different metabolic pathways. Twenty metabolic pathways were revealed to be abnormal in patients with Chronic Fatigue Syndrome. Sphingolipid, phospholipid, purine, cholesterol, microbiota, pyrroline-5-carboxylate, riboflavin, branch chain amino acid, peroxisomal, and mitochondrial pathways were all disrupted. Diagnostic accuracies of 94% were found using AUC characteristic curve analysis. In our experiment utilizing the ML-based model, we were able to achieve a greater level of accuracy (98%) for our proposed prediction model, the RFC model. Petrick, and Shomron [50] has been discussed how well the ML-based model performs. They highlighted how AI and ML have permitted important breakthroughs in untargeted metabolomics workflows and key findings in the fields of disease diagnosis. In conclusion, the proposed model (RFC) was successful in correctly diagnosing ME/FCS patients. The findings indicate that ML, when paired with the Shapely analysis, is able to explain the ME/FCS classification model and offer physicians a basic knowledge of the main metabolic chemicals that influence the model decision. Clinicians can get benefit from individual explanations of the important metabolic compounds in order to gain a better grasp of why the model yields certain diagnoses for individuals with ME/CFS.

5. Conclusions

Although research into the causes and mechanisms of ME/CFS continues, the exact underlying factors are not yet fully understood. It has been reported to result from a complex interaction of biological, genetic, environmental, and psychological factors. Advances in research are crucial to better understanding the disease, improving diagnosis and treatment options, and ultimately finding a cure. Based on this information, the RFC model proposed in this study correctly classified and evaluated ME/CFS patients through the selected biomarker candidate metabolites. The methodology combining ML and XAI can provide a clear interpretation of risk estimation for ME/CFS, helping physicians intuitively understand the impact of key metabolomics features in the model.

6. Limitations and future works

This study lacked a third-party verification by an independent biologist, which may have provided more explanation of the collected results, vital metabolic chemicals, and their significance to the diagnosis of patients with ME/FCS. It is vital to broaden the present investigation further by incorporating multicenter experiments in subsequent research or to make use of the associated data from multiple locations for external validation. The size of the metabolomics dataset might be increased by collecting additional samples from patients. This would be an improvement for this line of investigation. The performance of patient diagnosis will be improved by the development of advanced transfer learning-based methodologies.

15

References

1. Toogood PL, Clauw DJ, Phadke S, Hoffman D. Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS): Where will the drugs come from? Pharmacological Research. 2021;165:105465.
2. Germain A, Barupal DK, Levine SM, Hanson MR. Comprehensive circulatory metabolomics in ME/CFS reveals disrupted metabolism of acyl lipids and steroids. Metabolites. 2020;10:34.
3. Malato J, Graça L, Sepúlveda N. Impact of imperfect diagnosis in ME/CFS association studies. medRxiv. 2022:2022.06. 08.22276100.
4. Valdez AR, Hancock EE, Adebayo S, Kiernicki DJ, Proskauer D, Attewell JR, et al. Estimating prevalence, demographics, and costs of ME/CFS using large scale medical claims data and machine learning. Frontiers in pediatrics. 2019;6:412.
5. Faro M, Sàez-Francás N, Castro-Marrero J, Aliste L, de Sevilla TF, Alegre J. Gender differences in chronic fatigue syndrome. Reumatología clínica (English edition). 2016;12:72-7.
6. Marshall-Gradisnik S, Eaton-Fitch N. Understanding myalgic encephalomyelitis. Science. 2022;377:1150-1.
7. Malkova A, Shoenfeld Y. Autoimmune autonomic nervous system imbalance and conditions: Chronic fatigue syndrome, fibromyalgia, silicone breast implants, COVID and post-COVID syndrome, sick building syndrome, post-orthostatic tachycardia syndrome, autoimmune diseases and autoimmune/inflammatory syndrome induced by adjuvants. Autoimmunity reviews. 2022:103230.
8. Dehhaghi M, Panahi HKS, Kavyani B, Heng B, Tan V, Braidy N, et al. The role of kynurenine pathway and NAD+ metabolism in myalgic encephalomyelitis/chronic fatigue syndrome. Aging and disease. 2022;13:698.
9. Nunes JM, Kell DB, Pretorius E. Cardiovascular and haematological pathology in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS): A role for viruses. Blood reviews. 2023:101075.
10. Hornig M, Montoya JG, Klimas NG, Levine S, Felsenstein D, Bateman L, et al. Distinct plasma immune signatures in ME/CFS are present early in the course of illness. Science advances. 2015;1:e1400121.
11. Shan ZY, Barnden LR, Kwiatek RA, Bhuta S, Hermens DF, Lagopoulos J. Neuroimaging characteristics of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS): a systematic review. Journal of translational medicine. 2020;18:1-11.
12. Navaneetharaja N, Griffiths V, Wileman T, Carding SR. A role for the intestinal microbiota and virome in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)? Journal of clinical medicine. 2016;5:55.
13. Maes M, Leunis J-C, Geffard M, Berk M. Evidence for the existence of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) with and without abdominal discomfort (irritable bowel) syndrome. Neuroendocrinol Lett. 2014;35:445-53.
14. Germain A, Giloteaux L, Moore GE, Levine SM, Chia JK, Keller BA, et al. Plasma metabolomics reveals disrupted response and recovery following maximal exercise in myalgic encephalomyelitis/chronic fatigue syndrome. JCI insight. 2022;7.
15. Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans M, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of clinical epidemiology. 2001;54:774-81.
16. Levman J, Ewenson B, Apaloo J, Berger D, Tyrrell PN. Error Consistency for Machine Learning Evaluation and Validation with Application to Biomedical Diagnostics. Diagnostics. 2023;13:1315.

17. Zhang X, Liu C-A. Model averaging prediction by K-fold cross-validation. Journal of Econometrics. 2023;235:280-301.

18. Diniz MA. Statistical methods for validation of predictive models. Journal of Nuclear Cardiology. 2022;29:3248-55.

19. Zhang J, Wang Q, Shen W. Hyper-parameter optimization of multiple machine learning algorithms for molecular property prediction using hyperopt library. Chinese Journal of Chemical Engineering. 2022;52:115-25.

20. Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H. Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology. 2019;17:26-40.

21. Jones DR. A taxonomy of global optimization methods based on response surfaces. Journal of global optimization. 2001;21:345-83.

22. Yagin FH, Gülü M, Gormez Y, Castañeda-Babarro A, Colak C, Greco G, et al. Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique. Applied Sciences. 2023;13:3875.

23. Mansourian P, Zhang N, Jaekel A, Zamanirafe M, Kneppers M. Anomaly Detection for Connected Autonomous Vehicles Using LSTM and Gaussian Naïve Bayes. International Conference on Wireless and Satellite Systems: Springer; 2023. p. 31-43.

24. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM, Suri HS, Biswas M, et al. Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. Computer methods and programs in biomedicine. 2019;176:173-93.

25. Iqbal A, Barua K. A real-time emotion recognition from speech using gradient boosting. 2019 international conference on electrical, computer and communication engineering (ECCE): IEEE; 2019. p. 1-5.

26. Alshboul O, Shehadeh A, Almasabha G, Almuflih AS. Extreme gradient boosting-based machine learning approach for green building cost prediction. Sustainability. 2022;14:6651.

27. Shah K, Patel H, Sanghvi D, Shah M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research. 2020;5:1-16.

28. Muharemi F, Logofătu D, Leon F. Machine learning approaches for anomaly detection of water quality on a real-world data set. Journal of Information and Telecommunication. 2019;3:294-307.

29. Ilyas H, Ali S, Ponum M, Hasan O, Mahmood MT, Iftikhar M, et al. Chronic kidney disease diagnosis using decision tree algorithms. BMC nephrology. 2021;22:1-11.

30. Sattari MT, Apaydin H, Shamshirband S. Performance evaluation of deep learning-based gated recurrent units (GRUs) and tree-based models for estimating ETo by using limited meteorological variables. Mathematics. 2020;8:972.

31. Daneshvar D, Behnood A. Estimation of the dynamic modulus of asphalt concretes using random forests algorithm. International Journal of Pavement Engineering. 2022;23:250-60.

32. Yacouby R, Axman D. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. Proceedings of the first workshop on evaluation and comparison of NLP systems2020. p. 79-91.

33. Bowers AJ, Zhou X. Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. Journal of Education for Students Placed at Risk (JESPAR). 2019;24:20-46.

34. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean journal of anesthesiology. 2022;75:25-36.

35. Muschelli III J. ROC and AUC with a binary predictor: a potentially misleading metric. Journal of classification. 2020;37:696-708.

36. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association. 2020;27:621-33.

37. Liu J, Wang C, Yan R, Lu Y, Bai J, Wang H, et al. Machine learning-based prediction of postpartum hemorrhage after vaginal delivery: combining bleeding high risk factors and uterine contraction curve. Archives of Gynecology and Obstetrics. 2022;306:1015-25.

38. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, et al. Explainable AI in medical imaging: An overview for clinical practitioners–Beyond saliency-based XAI approaches. European journal of radiology. 2023:110786.

39. Yagin FH, Cicek İB, Alkhateeb A, Yagin B, Colak C, Azzeh M, et al. Explainable artificial intelligence model for identifying COVID-19 gene biomarkers. Computers in Biology and Medicine. 2023;154:106619.

40. Khanna VV, Chadaga K, Sampathila N, Prabhu S, Bhandage V, Hegde GK. A distinctive explainable machine learning framework for detection of polycystic ovary syndrome. Applied System Innovation. 2023;6:32.

41. Chatterjee J, Dethlefs N. Scientometric review of artificial intelligence for operations & maintenance of wind turbines: The past, present and future. Renewable and Sustainable Energy Reviews. 2021;144:111051.

42. Tanaka M, Tajima S, Mizuno K, Ishii A, Konishi Y, Miike T, et al. Frontier studies on fatigue, autonomic nerve dysfunction, and sleep-rhythm disorder. The Journal of Physiological Sciences. 2015;65:483-98.

43. Yamano E, Watanabe Y, Kataoka Y. Insights into metabolite diagnostic biomarkers for myalgic encephalomyelitis/chronic fatigue syndrome. International Journal of Molecular Sciences. 2021;22:3423.

44. Group ICFSS. The chronic fatigue syndrome: A comprehensive approach to its definition and study. Annals of Internal Medicine. 1994;121:953-9.

45. Armstrong CW, McGregor NR, Lewis DP, Butt HL, Gooley PR. The association of fecal microbiota and fecal, blood serum and urine metabolites in myalgic encephalomyelitis/chronic fatigue syndrome. Metabolomics. 2017;13:1-13.

46. Tomas C, Newton J. Metabolic abnormalities in chronic fatigue syndrome/myalgic encephalomyelitis: a mini-review. Biochemical Society Transactions. 2018;46:547-53.

47. Huth TK, Eaton-Fitch N, Staines D, Marshall-Gradisnik S. A systematic review of metabolomic dysregulation in chronic fatigue syndrome/myalgic encephalomyelitis/systemic exertion intolerance disease (CFS/ME/SEID). Journal of translational medicine. 2020;18:1-14.

48. Jason LA, Boulton A, Porter NS, Jessen T, Njoku MG, Friedberg F. Classification of myalgic encephalomyelitis/chronic fatigue syndrome by types of fatigue. Behavioral Medicine. 2010;36:24-31.

49. Nagy-Szakal D, Barupal DK, Lee B, Che X, Williams BL, Kahn EJ, et al. Insights into myalgic encephalomyelitis/chronic fatigue syndrome phenotypes through comprehensive metabolomics. Scientific reports. 2018;8:10056.

50. Petrick LM, Shomron N. AI/ML-driven advances in untargeted metabolomics and exposomics for biomedical applications. Cell Reports Physical Science. 2022;3.