The EuroBiotech Journal
RESEARCH ARTICLE
Bioinformatics

Machine Learning Approaches for Obsessive Compulsive Disorder Detection

Kabita Patel, Ajaya K. Tripathy

Abstract

Obsessive-Compulsive Disorder (OCD) is a psychiatric illness that produces significant psychological distress in patients. Individuals with OCD have recurring unwanted thoughts or sensations which make them obsessed with something and feel to do something repetitively as a compulsion. In general detection of OCD is performed by symptoms analysis. However, the symptoms are significantly visible at a later stage. Even individuals with OCD have less faith in the analysis of the symptoms as long as it is not affecting their life negatively. As a result, they start their treatment at a later stage and the treatment process becomes longer. However, it is observed that if the detection is performed through laboratory analysis through some biomarkers then the patients have more faith in the detection process and can start their treatment well in advance. Therefore laboratory detection of OCD can play a vital role in OCD treatment effectiveness. Most of the laboratory detection process proposed in the literature uses Machine Learning on related biomarkers. However, the prediction accuracy rate is not enough. This research aims to analyze the approaches to pediatric OCD based on machine learning using neuroimaging biomarkers and oxidative stress biomarkers. The challenges in OCD detection and prediction using neuroimaging biomarkers, oxidative stress biomarkers, and Machine Learning models have been described. Further, it analyzes the performance of different machine learning models that were used for OCD detection and highlights the research gap to improve prediction accuracy.

Keywords: Machine Learning , Obsessive Compulsive Disorder, Artificial Intelligence.

Introduction

Obsessive-Compulsive disorder (OCD) is one of the psychiatric disorders. Persons with OCD are having an unwanted and unpleasant feeling. Obsession turns a person anxious and fearful. As a protective measure against obsession, the individual with OCD involves in doing repeated activities called compulsion. In the latter stage OCD impact negatively individuals day to day life. As a treatment for OCD one can opt for Cognitive-Behavioral Therapy (CBT), repetitive Transcranial Magnetic Stimulation (rTMS), transcranial Direct Current Stimulation (tDCS), and Deep Brain Stimulation (DBS) [1].

The most commonly used method for OCD detection is behavioral symptoms analysis [2]. However, this method works well when the disease is already in an advanced stage. Mild symptoms are observable at an early stage. At the early stage, the individual feels normal and able to manage day-to-day activity smoothly, the individual with OCD (at the preliminary stage) is reluctant to believe the symptoms analysis detection result. As a result, they feel normal and not accepting the treatment at an early stage.

At a later stage when it affects negatively and the day-to-day activity gets affected negatively the OCD patient starts believing that they are mentally ill and start accepting treatment. In most cases, the patient becomes resistant to common OCD treatment and the treatment process becomes difficult and longer. However, laboratory detection brings more faith in the patient's mind and the probability of acceptance of treatment at an early stage may increase if laboratory detection is adopted. Therefore, early detection of OCD through laboratory tests can play a vital role in OCD treatment.

In the literature, several studies are performed on the prediction or detection of OCD

Kabita Patel
Dept. of Computer Sci. , Gangadhar Meher University, India
E-mail: patelkabita92@gmail.com
Ajaya K. Tripathy (Corresponding Author)
Dept. of Computer Sci. , Gangadhar Meher University, India
E-mail: ajayatripathy1@gmail.com

DOI: 10.2478/ebtj-2023-0012

© 2023 Kabita Patel et al., published by Sciendo
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License.

[3–6]. Most of these approaches use neuroimaging biomarkers or oxidative stress biomarkers for OCD analysis [5, 7–10]. The popular neuroimaging biomarkers used in OCD analysis are Electroencephalogram (EEG), Magnetic Resonance Imaging (MRI), Functional Magnetic Resonance Imaging (fMRI), and Diffusion Tensor Imaging (DTI). The popular oxidative stress biomarkers used in OCD analysis are superoxide Dismutase (SD), Glutathione Peroxidase (GP), Catalase (CAT), Malondialdehyde (MAL), and Serum Cortisol (SC). Most of the approaches proposed in the literature for OCD detection are based on machine learning models. These machine-learning models are applied to neuroimaging or oxidative stress biomarkers.

This research aims to analyse the approaches to paediatric OCD based on machine learning using neuroimaging biomarkers and oxidative stress biomarkers. The challenges in OCD detection and prediction using neuroimaging biomarkers, oxidative stress biomarkers, and Machine Learning models have been described. Further, it analyses the performance of different machine learning models that were used for OCD detection and highlights the research gap to improve prediction accuracy.

The rest of the article is organized as follows. Section 2 provides a brief background of Obsessive Compulsive Disorder. A brief introduction to OCD-relevant biomarkers is presented in Section 3. A brief introduction to machine learning and a brief description of the popular machine learning models used for OCD detection are presented in Section 4. A comprehensive related literature review is provided in Section 5. This section also provides a comparative analysis of OCD prediction or detection accuracy achieved by different machine learning models using different biomarkers. Section 6 highlights the research gap and future research directions. Finally, Section 7 concludes the article.

2. Obsessive Compulsive Disorder

Obsessive-Compulsive Disorder (OCD) is one of the psychiatric disorders. Persons with OCD are having an unwanted and unpleasant feeling. Obsession turns a person anxious and fearful. As a protective measure against obsession, the individual with OCD involves in doing repeated activities called compulsion. In the latter stage OCD impact negatively individual's day to day life. As a treatment for OCD one can opt for CBT, rTMS, tDCS, and DBS. Early detection of OCD through laboratory tests can play a vital role in OCD treatment.

In the literature, several studies are performed on the prediction or detection of OCD. Most of the approaches proposed in the literature for OCD detection are based on machine learning models. These machine learning models are applied to biological markers.

3. Biological Markers for OCD Analysis

There are two types of biomarkers that are used in OCD identification such as Neuroimaging Biomarkers and Oxidative Stress Biomarkers. The area of medical imaging known as neuroimaging focuses on the brain. Neuroimaging investigates: How the brain functions in complement to recognizing problems and measuring cognitive performance. Neuroimaging or brain scanning includes the use of various techniques to either directly or indirectly image the structure, function, or pharmacology of the brain. Several studies include the neuroimaging biomarkers EEG, DTI, MRI, and fMRI. On other hand, many kinds of literature describe oxidative stress biomarkers use in OCD patient, where oxidative stress include Superoxide Dismutase (SOD), Glutathione Peroxide (GPx), Malondialdehyde (MDA), Catalase (CAT), and serum cortisol. These biomarkers are used in identifying OCD patients in the pathological clinic.

4. Machine learning Approaches for OCD Analysis

Machine learning involves the development of models and algorithms that automatically learn from past patterns and predict the outcome of unknown patterns. The most commonly used approaches used in machine learning are classification, clustering, and regression.

Classification is a supervised learning task that aims to classify or categorize input patterns into specified classes. Classification models are trained on labeled patterns, which include both the input patterns and their corresponding class labels. A trained classification model can be used for OCD detection and can detect from the input features of an unknown pattern. The input features are extracted from the biomarkers used in OCD detection.

Clustering is an unsupervised learning task that groups data points based on intrinsic characteristics without predefined class labels. It identifies patterns and structures in data and assigns them to clusters based on proximity or similarity. These models can be used for segregating healthy control and OCD affected individuals.

Regression is a supervised learning task that predicts a continuous or numerical value based on input features. It estimates the relationship between input features and output variables, aiming to learn a mapping function for accurate predictions. These models can be used for OCD severity analysis.

A short description of the machine learning models used for OCD detection is presented as follows.

4.1 Logistic Regression

The logistic regression function gives a value between 0 and 1, which means that if the output is greater than 0.5 it is positive if it is less than 0.5 it belongs to the negative class. Suppose a logistic function $g(z) = 1/(1+e^{-z})$

If we have a function $h_{(\beta(X))} = g(\Sigma\, \beta_i X_i) = g(\Sigma\, \beta^T X) = 1/(1+e^{-\beta^T X})$

Suppose we have the input X and want to find out what is the probability of 'Y' given 'X', if Y = 1

$P(Y|X) = h(X)$ If $(Y = 1)$

$= h(X)^y\, (1-h(x))^{1-y}$

Where $h_{\beta(X)} = 1/(1+e^{-\beta^T X})$ Now, $P_y(X\beta)$ is our estimate of $P(y\beta)$.

We first define the likelihood and learn the optimal value of β

$L(\beta) = P(Y|X = \beta)$

$= \prod_{i=1}^{m} P(Y_i|X_i, \beta)$

$= \prod_{i=1}^{m} h(x_i)^{y_i} (1-h(x_i))^{1-y_i}$

$\log L(\beta) = \prod_{i=1}^{m} \wedge y_i \ \log h(x_i) + (1- y_i) \log (1 - h(x_i))$

Now we have to maximize this function

Where $\beta = \beta + \alpha \nabla_\beta l(\beta)$

Here α is learning rate

$\partial/\partial\beta \ (l(\beta)) = ((y \ (1/g(\beta^T X)) - (1-y)(1/(1-g(\beta^T X))\partial/(\partial\beta j) g(\beta^T X))$

$= (y(1 - g(\beta^T X)) - (1 - y) g(\beta^T X))X_j$

$= (y - h_\beta(x)) X_j$

Now we get, $\beta_j = \beta_j + \alpha(y - h_\beta (x)) X_j$

This is the stochastic gradient of logistic regression. The more illustration is in [11], and the algorithm is shown bellow in Algorithm1.

Algorithm 1: Logistic Regression Procedure

1. Initialize β with initial values.
2. Set learning rate α to a small positive value.
3. while not converged:
4. Initialize gradient vector $\nabla\beta$ to zeros.(a) for each training example (x_i, y_i):
   Compute sigmoid function value $h_{\beta(X)} = 1/(1+e^{-\beta^T X})$
   (b) Update log-likelihood gradient for each parameter j:
   $\nabla\beta_j = (y_i - h_{\beta(X)}) X_j$
   (c) for each parameter j: Update β_j:
   $\beta_j = \beta_j + \alpha(y - h_\beta(x)) X_j$
5. Repeat steps b-c for all training examples.
6. (Convergence criterion - optional) Check if the log-likelihood has converged or maximum iterations reached.
   Return the optimized parameter vector β.

4.2 K- Nearest Neighbors

K- Nearest Neighbors (KNN) algorithm uses feature similarity to predict the values of new data points which further means that the new data points will be assigned a value based on how closely it matches the points in the training set working steps are shown in Algorithm: 2.

One might read [12] for a comprehensive discussion of the K-nearest neighbor analysis.

Algorithm 2: K- Nearest Neighbor Classification Procedure.

1. For implementing any algorithm, we need a dataset so, during the first step of KNN, we must load the training as well as test data
2. Next, we need to choose the value of K i.e the nearest data points, K can be any integer.
3. For each point in the data do the following

(a) Calculate the distance between the last data and each row of training data with the help of any of the methods name ly: Euclidean distance. The most commonly used method to calculate the distance is Euclidean.

(b) Now, based on the distance value sort them in ascending order.

(c) Next, it will choose the top K rows from the sorted array.

(d) Now, it will assign a class to the test point based on the most frequent class of these rows.

4. End.

4.3 Decision Tree

The entropy of a data set 'D' is determined as

$$E(D) = -\sum x_i \log_2 x_i$$

Where x_i is the non-zero probability that an arbitrary temple in D belongs to class M_i

$x_i = |M_i D|/|D|$ (Probability of class ' i ')

Where E = 0 if and only if all the instances have the same class.

E is the maximum when instances are equally distributed among 'N' possible classes

Let assume the partitioning attribute having p distributed value $y_1, y_2, - - --, y_p$, as obsered from D in Fig. 1.

∴ D is partitioned to p parts as $D_1, D_2, ..., D_p$ where D_j contains tuples in D that have outcome y_i of Y.

i.e $D_1 = y_1, D_2 = y_2$.

The weighted entropy E_y (D) for all partitions of D concerning y is

$E_y(D) = \sum_{j=1}^{p} |D_j|/|D| \times E(D_j)$

∴ Information gain $G(Y) = E(D) - E_y(D)$

The attribute with the highest G(Y) is chosen as the selecting

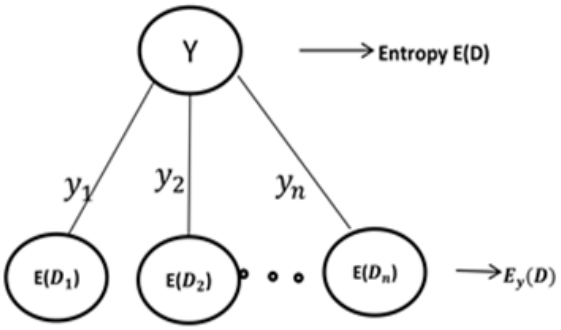


Figure 1. Decision Tree Attribute Selection for Splitting.

attribute at node N. One might read [13] for a comprehensive discussion of the Decision Tree. The total description of this Decision Tree algorithm shown in Algorithm. 3.

Algorithm 3: Decision Tree Procedure.

1. Define a class Node to represent a node in the decision tree:
2. If the node is a leaf node, store the class label.
3. Else, If the node is an internal node, store the splitting attribute and create child nodes.
4. Implement a function to calculate the entropy given a dataset D:
5. Initialize entropy E to 0.
6. For each class label M_i in the dataset: Calculate the probability $x_i = |M_i, D|/|D|$.
7. If X_i is non-zero, update the entropy:

$E(D) = -\sum x_i \log_2 x_i$

8. Return the entropy E
9. Implement a function to calculate the information gain given a dataset D and a splitting attribute A:
10. Calculate the entropy of the dataset: E(D) = entropy (D).
11. Initialize weighted entropy E(A) to 0.
12. For each possible value y_i of the attribute A: Create a subset D_j containing tuples in D with attribute value y_i.
13. Calculate the probability $x_i = |D_j|/|D|$
14. Update the weighted entropy: $E(A) += x_i * \text{entropy}(D_j)$.
15. Calculate the information gain: $G(Y) = E_D - E_A$
16. Return the information gain G(Y).

4.4 Support Vector Machine

Given the training dataset $\{A_i Z_i\}_{i=1}^N$, data point $A_i \in R^p$,where p is the number of predictors, class label $Z_i \in \{1, 2, ..., C\}$. The training dataset contains N number of samples and the number of class levels is C.

Let the equation of separability of hyperplane as

$\sum_{i=1}^n W_j A' + b = 0$ i.e $W^T A + b = 0$

Let "Â" be a point on the upper border of the margin. $\tilde{A}$ is a point on lower boarder of margin.

The projection of (" Â - $\tilde{A}$") along perpendicular to the optimal hyperplane equal to ("Â - $\tilde{A}$").

Where, Gradient will give vector perpendicular to that surface $\nabla A(W^T A+b)=W$

$\delta = (\hat{A} - \tilde{A})\, W/|W| = (W^T (\hat{A} - \tilde{A}))/|A|$ in matrix form.

$W^T A+b = \Delta$ for the sample in the upper border.

$W^T A+b = -\Delta$ for the sample in the lower border.

$W^T A+b = 0$ on the Optimal hyper plain.

Now, $(W^T A)/\Delta + b/\Delta = 1$

$(W^T A)/\Delta + b/\Delta = -1$

$(W^T A)/\Delta + b/\Delta = 0$

Let, $W^T/\Delta = W$ and $b/\Delta = b$

$W^T A+b = 1$ Lower

$W^T A+b = -1$ Upper

$W^T A+b = 0$ Optimal hyper plain.

Where, $W^T \tilde{A} + b = 1$

$W^T \tilde{A} + b = -1$

$W^T (\hat{A} - \tilde{A}) = 2$

Now, $\delta = 2 / ||W||$

$\delta/2 = 1/||W||$

$\therefore$ Maximize $1/||W||$ or minimize $||W||$ constant satisfied by all the data points.

i.e $W^T A_i + b = Z_i$

Now, minimize $1/2\, ||W||^2$

$W^T A_i + b \geq 1 \rightarrow$ for + ve Sample

$W^T A_i + b \leq 1 \rightarrow$ for - ve Sample

$\Rightarrow 1 - (W^T A_i + b) Z_i \leq 0$

By contrast the Lagrange multiplayer as

$L(W, b, \alpha) = 1/2\, ||W||^2 + \sum_{i=1}^N \alpha_i (1 - (W^T A_i + b) Z_i)$

Where the additional non-negative variables α_i are called lagrange multipliers. One might read [14] for a comprehensive discussion of the Support Vector Machine analysis and see the stapes in Algorithm 4.

Algorithm 4: Support Vector Machine Procedure

1. Input the training dataset $\{A_i, Z_i\}_{i=1}^N$,, where A_i is the feature matrix and Z_i is the corresponding target variable.
2. Initialize the weight vector w and bias term b to zeros.
3. Set the learning rate α and the regularization parameter C.
4. Set the maximum number of iterations or convergence criteria.
5. Define the upper border point "Â" and the lower border point "$\tilde{A}$" .
6. Repeat the following steps until convergence or the maximum number of iterations is reached:

(a) For each training example (A_i, Z_i) in the dataset: Compute the predicted output $\hat{Z}_i$ as: $\hat{Z}_i = W^T A_i + b$ Calculate the hinge loss $L = \max(0, 1 - Z_i \hat{Z}_i)$

(b) Update the weight vector and bias term: If $Z_i \hat{Z}_i \leq 1$ Calculate the projection of the vector ("Â - $\tilde{A}$ ") along the direction.

(c) perpendicular to the optimal hyperplane:

$\text{proj} = ((\hat{A} - \tilde{A}).W)/(||W||^2)$

(d) Update the weight vector and bias term:

$W = W + \alpha(Z_i - \hat{Z}_i).A_i - \alpha.C.proj.W$

$b = b + \alpha.(Z_i - \hat{Z}_i)$

(e) $(Z_i - \hat{Z}_i > 1)$: Update the weight vector based on the regularization term:

$W = W - \alpha.C.W$

7. Output the learned weight vector W and bias term b.

4.5 Random Forest

Random Forest is a popular ensemble learning algorithm in machine learning that combines the predictions of multiple individual decision to make more accurate and robust predictions. It is a versatile algorithm that can be used for both classification and regression tasks. The steps involved in the Random Forest Procedure Algorithm 5.

Algorithm 5: Random Forest Procedure

1. Start by choosing random samples from a pre-existing dataset.
2. A decision tree will then be built by this method for each sample. The forest outcome from each decision tree will then be obtained.
3. Voting will be done in this phase for each expected outcome.
4. Finally, choose the prediction result that received the most votes as the final forest result.

Fig. 2 gives a brief description of the graphical representation of Random Forest and One might read [15] for a comprehensive discussion of the Random Forest analysis.

4.6 Bayes Decision Theory

Bayesian decision theory is a fundamental statistical approach

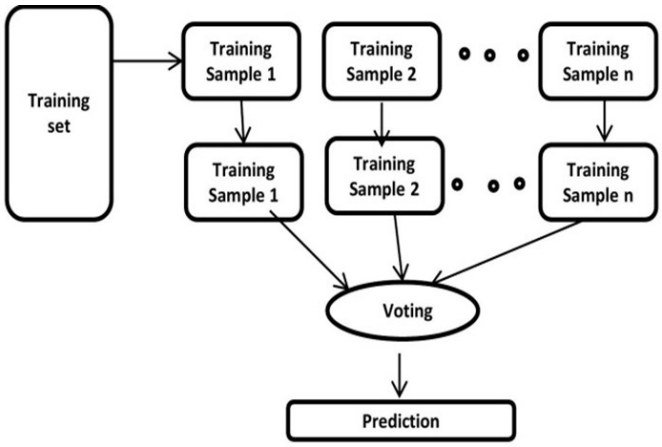


Figure 2. Working Procedure for Random Forest.

to the problem of pattern classification. It is considered the ideal case in which the probability structure underlying the categories (or class) is known perfectly.

Let X is a feature that we will use to predict class W.

We consider X to be a continuous random variable whose distribution depends on the class and it is expressed as P(X—W). This is the class conditional probability density function.

The joint probability density of finding a pattern that is in class W_j and has feature X can be written in two ways

i.e $P(W_j, X) = P(W_j|X)P(X)$

or, $P(W_j|X) = (P(W_j|X)P(W_j))/(P(X))$

Where, $P(X) = \sum_{j=1}^{c} (P(W_j|X)P(W_j))$

Where c is the number of classes.

This is the Baysed formula and one might read [16] for a comprehensive discussion of the Bayesian Decision Theory, Algorithm 6 is shown below.

Algorithm 6: Bayes Decision Theory Base Classification.

1. Input: Feature X and class labels W.
2. Estimate P (X|W) for each class W.
3. Calculate $P(W_j)$ for each class W_j.
4. Compute P(X) by summing $P(W_j|X)*P(W_j)$ for all classes.
5. For each class W_j, compute $P(W_j|X)$ as $(P(W_j|X)*P(W_j))/ P(X)$
6. Output the posterior probabilities $P(W_j|X)$ for each class W_j.

5. OCD Detection Literature Review and Comparative Analysis

In [17] this paper the author describes how a mental patient classifies by using the random forest machine learning algorithm. The drawback of this paper they do not predict the OCD patient they take all the mental patients. Use a combined decision tree, neural networks, and support vector machine in the random forest algorithm. The result of this gives 92.10% accuracy in the decision tree, 92.01% accuracy in the neural network, and 91.04% accuracy in the support vector machine. In [18] the author describes how drugs are responsible for OCD patients. For classifying, the patient uses the random forest as a classifier. They combine linear discriminant analysis and support vector machine in this random forest model. The limitation of this is they do not identify with OCD patients or healthy control.

In [19] the authors study about mental health from childhood to adulthood. The mental illness includes stress, social anxiety, depression, and obsessive-compulsive disorder. They accumulate the data from the online dataset. They use the Decision tree, Random forest, and Naive Bayes machine learning algorithm as classification techniques. By applying the three machine learning approach they get 82.2% accuracy in the decision tree, 79.3% accuracy in a Random forest, and 78.7% in the Naive Bayes algorithm out of these the decision tree gives the best result. The limitation of this study they do not use any biomarkers for identifying mental illness. They do not identify the patient with OCD or not. Summary of machine learning in mental health patient are represent in Table: 1.

In [20] the author discusses the treatment response of OCD patients. They expand an ensemble classification algorithm. They take continuous and discrete types of data. They combine the decision tree and Support Vector Machine(SVM) for batter ensemble classification, and they achieve 84% accuracy. The limitation of this is they only predict the treatment responder or non-responder of OCD patients. They do not identify the three class classifications like whether they are OCD or not and genetically linkage. They also compare the result of the accuracy, precision, sensitivity, specificity, and f-measure. In [9] the author study about the treatment response by using the EEG as biomarkers and random forest machine learning algorithm as a classification. They use the Approximate entropy (ApEn) method to excerpt the EEG complexity, then they use the random forests method to classify the ApEn complexity. They get 90.64% accuracy in treatment responsive of the patient The drawback of this study is they only predict the treatment response of OCD patients. They did not predict the genetic linkage and healthy control with OCD patients or not an OCD patient. In this study [21] the author predicts the early treatment response in patients with obsessive-compulsive disorder. They use the clinical data of resting-state functional magnetic response imaging (fMRI), and functional amplitude of low-frequency fluctuations (fALFF) as a biomarker, and they use a support vector regression (SVR) machine learning approach for data analysis. They get 90% accuracy by using the SVM classification approaches. In [22] the author predicts the treatment response within OCD patients by using Internet delivered Cognitive Behaviors Therapy (ICBT) as a biomarker. They use four machine-learning approaches for classification. The result of all four machine learning approaches is compared between each other, out of this the SVM technique gives the batter accuracy heaving with 83%. The limitation of this is they do not identify with OCD or healthy control.

In [8] the author predicts the severity of OCD patients.

Table 1. Studies relevant to mental health analysis using machine learning.

Publication	Approach and Analysis	Accuracy	Limitation
H. Abou-Warda, 2016	Take all type of mental patient and classify by using random forest machine leaning algorithm.	92%	Don't classify with OCD or Healthy control.
K. Sekaran and M. Sudha, 2020	Describe how drugs are responsible for OCD patient, Use Support Vector Machine as a classifier.	97.29%	Don't classify with OCD or Healthy control
V. Laijawala, 2020	Study about mental health from childhood to adulthood use decision tree is best classifier.	82.2%	Don't classify with OCD or Healthy control.

Table 2. Studies relevant to OCD treatment response prediction using machine learning.

Publication	Approach and analysis	Accuracy	Limitation
H. Hasanpour, 2019	Discusses the treatment response of OCD patients by using the ensemble classification.	84%	They don't identify with OCD or Healthy control.
T. B. Altu˘glu, 2020	They use EEG as a biomarker and random forest as a classifier.	90.64%	They don't identify with OCD or Healthy control.
H. Yan, 2022	For early treatment response use the fMRI as a biomarker and SVM as classifier	90%	They don't identify with OCD or Healthy control.
F. Lenhard, 2018	Use ICBT for treatment and SVM as a classifier	83%	They don't identify with OCD or Healthy control.

Table 3. OCD severity detection studies using machine learning.

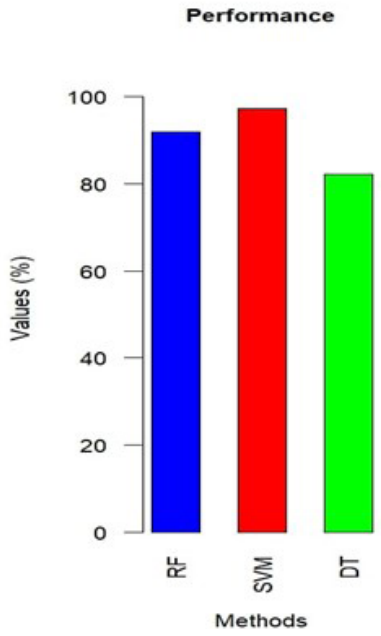
Publication	Approach and Analysis	Accuracy	Limitation
S. Mas, 2016	Use structural MRI, DTI as a biomarkers and SVM as a classifier.	90%	They don't identify with OCD or Healthy control.
M. Q. Hoexter, 2013	Use MRI as a biomarker and SVM as a classifier.	97%	They don't identify with OCD or Healthy control

Table 4. Studies relevant to OCD detection using machine learning.

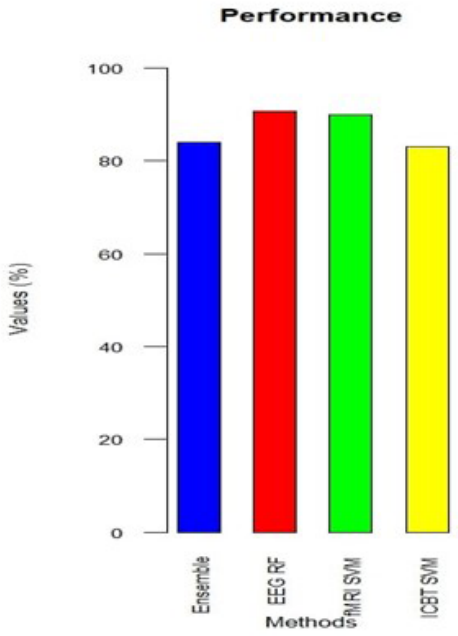
Publication	Approach and Analysis	Accuracy	Limitation
C. Zhou, 2018	Use MRI as a biomarker and SVM as a classifier with GMV, WMV, FA, AND MD	80.65%	They don't predict the genetic linkage of OCD patient.
X. Hu, 2016	Classify between the OCD patient and healthy control using MRI and SVM as a classifier	81.82%	They don't predict the genetic linkage of OCD patient.
A. Shrivastava, 2019	Describes the prediction of OCD patients by using oxidative stress biomarker and SVM as classifier.	96%	They do not identify the severity of the patient

Table 5. Machine learning methods for OCD treatment analysis.

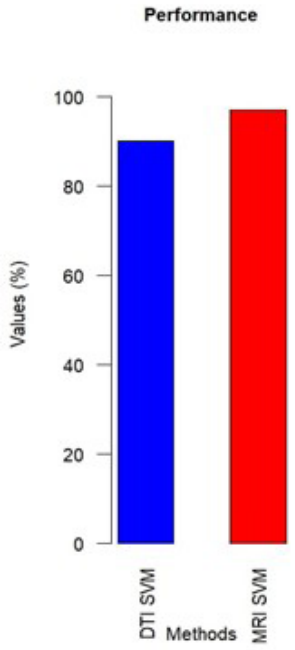
Publication	Approach and Analysis	Accuracy	Limitation
S. Aydin, 2015	Use the EEG signal as a biomarker and SVM as a classifier They conclude that OCD is a part of anxiety disorder.	85%	They don't predict with OCD or healthy control or genetic linkage.
T. Najafi, 2022	Discusses the obsessions, compulsions and healthy control use EEG signal and LDA as a classifier.	90%	They don't predict with OCD or healthy control or genetic linkage.



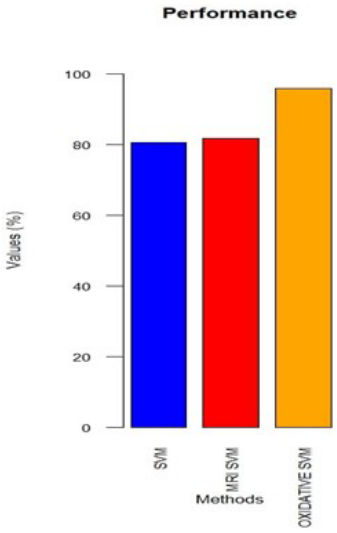
(a) Mental health relevant approaches.



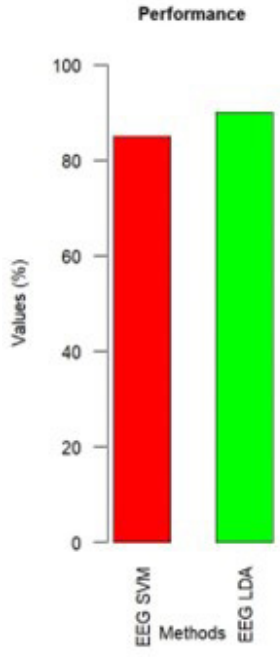
(b) OCD severity detection approaches



(c) Treatment response relevant approaches



(d) OCD detection analysis related



(e) OCD treatment analysis relevant approaches.

Figure 3. Performance of different approaches.

They collaborate the structural Magnetic Resonance Imaging (MRI), neuropsychological data, Diffusion tensor Imaging data (DTI), and genetic data to predict the severity of OCD patients they use a support vector machine and Naive Bayes two supervised machine learning algorithm as a classifier. They find 90% accuracy in child and adolescent patients with OCD severity. The limitation of this study is they do not predict the OCD patient or healthy control, they only predict the severity of OCD patient. In [7] the author predicts the severity of OCD patients by using Magnetic Resonance Imaging (MRI) as a bio-

marker and support vector machine learning approaches as a classifier, and heaving with 97% accuracy. The limitation of this study they only predict the severity of OCD they do not identify OCD or Healthy control.

In [3] the author predicts the Obsessive-Compulsive Disorder (OCD) patient and healthy control by using the MRI as a biomarker. They use a support vector machine (SVM) as a classification algorithm. They use four features within the SVM to get the batter result Gray Matter Volume (GMV), White Matter Volume (WMV), Fractional anisotropy (FA) and mean diffusivity (MD). They found 72.08%, 61.29%, 80.65%, and 77.42% accuracy by using GMV, WMV, FA, and MD. They concluded successfully whether the patient with OCD or healthy control. They do not identify whether the OCD patient is genetic linkage or not this is the main limitation of this study.

In this study, [4] the author classify between the OCD patient and Healthy Control using Magnetic Resonance Imaging (MRI) as a biomarker. Classify the OCD patient by including both Gray Matter (GM) and White Matter (WM), and get high classification accuracy i.e 81.82% by using the SVM as a classifier within the WM data. The limitation of this is the classify the genetically linkage with OCD patients. In [5] the author describes the prediction of OCD patients by using the oxidative stress biomarker. They predict the three-class classification of OCD patients such as those with Obsessive Compulsive Disorder (OCDP), non-affected First-Degree relatives (FDR), and Healthy Control (HC). They use different approaches of machine learning for classifying three-class classification such as Support Vector Machine (SVM), Random Forest (RF), Linear Discriminant Analysis (LDA), and K- Nearest Neighbor (KNN). The outcomes of all the machine learning approaches are compared with each other and get SVM is the best classifier technique for OCD patients heaving 96% accuracy. From this paper, we get all the classes are classified by using machine learning but they do not identify the severity of OCD patients.

Table 6. Prediction accuracy measures of OCD two class (OCD and non OCD) using SVM, RF, DT, LDA, and, KNN.

Measure	SVM	RF	DT	LDA	KNN
Accuracy	0.9644	0.9200	0.8220	0.8444	0.7711
Precision	0.9676	0.9255	0.8270	0.9310	0.8664
Recall	0.9810	0.9355	0.8280	0.8437	0.7887

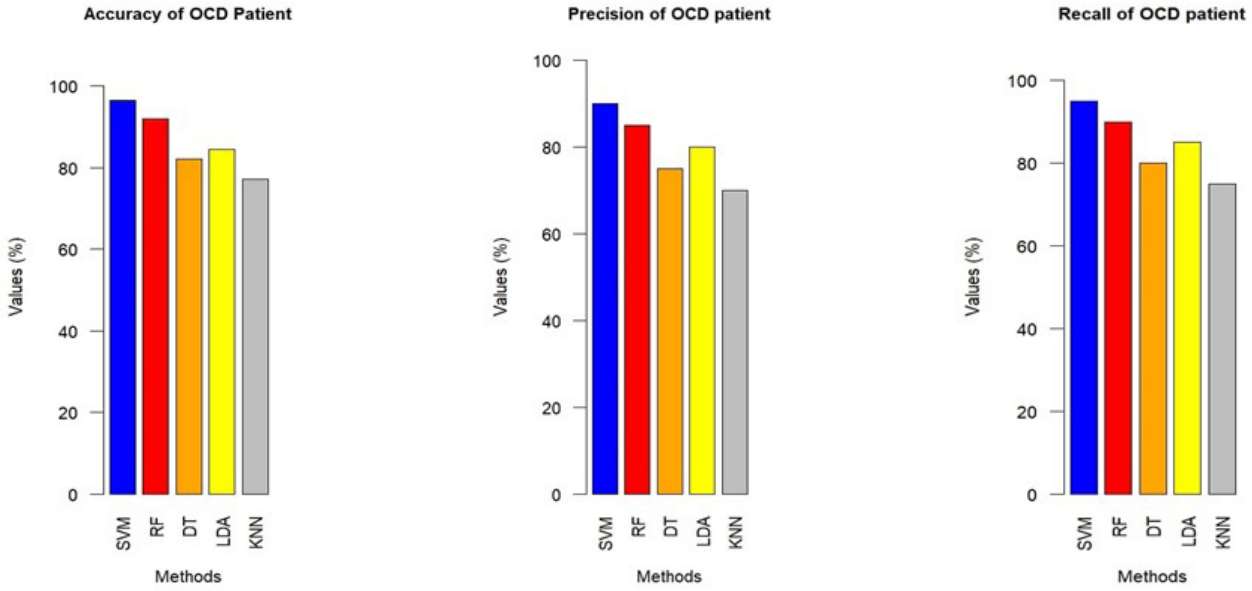
In this study, [23] the author predicts the OCD patient by using the neurobiology-aided feature and functional magnetic resonance. They use the Empaschiz algorithm and five-fold cross-validation heaving 80.3%. The limitation of this study is they only use the neuroimaging biomarkers, not the Oxidative stress biomarkers. In [24] the author discusses the obsessions, compulsions, and healthy control of the patient by using the EEG signal. The EEG signal is recorded from the severe obsession patient, severe compulsions, and healthy control patient. The classification of these three groups uses Linear Discrimination Analysis (LDA) for classification with 90% accuracy. The limitation of this study is they cannot classify the OCD patient, healthy control, and genetically link age with OCD patient. In this study, [25] the author uses the EEG signal as a biomarker and Support Vector Machine (SVM) as a classification technique. They use a two-fold cross-validation method for better results. They achieved 85% the highest classification accuracy. They conclude that OCD is a part of anxiety disorder. The disadvantage of this study is they did not predict the OCD patient like whether they are OCD patients, healthy control, or a genetic linkage. In this study [26] the author explore the correspondence between the range of OCD symptoms and those who are trusted in a clinical sample of outpatient. They use the principal components analysis as a classification technique. The limitation of this study they did not predict the OCD patient with healthy control.

In [27] the author predicts the OCD patient by applying the machine learning approaches. They used Random forest as an ensemble machine learning approach. They use the clinical data only and find the risk factor of the OCD patient. They get 95% accuracy to predict the OCD patient. The limitation of this is they do not identify whether the patient is OCD or not. The limitation of this paper is they only predict the treatment response not identifying the OCD or not OCD patient.

In [9] the author describes predicting the treatment response of the Obsessive-compulsive disorder patient by using Electroencephalography (EEG) as a biomarker. They use the random forest as a classification technique of the machine learning approach and they achieve 89.66% accuracy. The limitation of this study is they do not classify the OCD patient or healthy control and also do not identify the genetically linkage of OCD patients.

In various studies, different types of biomarkers and classification models, such as Random Forest, Support Vector Machines (SVM), Ensemble classification, Linear Discriminant Analysis (LDA), and Decision Tree, have been utilized. These studies employed diverse datasets for classification, including mental health patient data, functional Magnetic Resonance Imaging (fMRI), Electroencephalogram (EEG), and Internet-Based Cognitive Behavioral Therapy (ICBT) data for treatment response. Additionally, Magnetic Resonance Imaging (MRI) and Diffusion Tensor Imaging (DTI) were used to assess the severity of OCD patients, while MRI and oxidative stress data were employed to predict OCD patients.

Table 1 provides a summary of machine learning approaches in mental health patients, and a bar plot illustrating this information is presented in Figure 3(a). Similarly, Table 2 presents the different machine learning techniques utilized for predicting the treatment response of OCD patients, with the corresponding bar plot shown in Figure 3(b). The severity of OCD patients was assessed using various machine learning methods, as indicated in Table 3, and the summary is visualized in Figure 3(c). Prediction of OCD patients employed different machine learning techniques, detailed in Table 4, and the corresponding



(a) Prediction accuracy percentage bar plot.

(b) Precision bar plot.

(c) Recall bar plot.

Figure 4. OCD detection accuracy, Precision and Recall

bar plot is depicted in Figure 3(d). Lastly, Table 5 outlines the diverse machine learning approaches applied to OCD patient treatment, while Figure 3(e) represents a bar plot summarizing these findings. Again in Table 6 present the different accuracy, precision, and, recall of prediction of OCD patient, while Fig 4 represent a bar plot summarizing these findings.

6. Research Gap and Future Research Direction

This study analyses the OCD detection, prediction, treatment response, and severity analysis using machine learning models. To draw attention to research shortcomings with regard to the choice of the right model, the availability of datasets and signal pre-processing techniques. The following research gap has been identified, along with recommendations for future studies to fill it. These efforts can advance our understanding of OCD, improve diagnostic accuracy, personalize treatment approaches, and contribute to the development of more effective interventions for OCD patients.

1. In the literature, neuroimaging biomarkers got more importance for the detection of OCD using machine learning models. Collection of such biomarkers requires high-end types of equipment nearby the patients. Such facilities may not be available in rural areas. However, the collection of oxidative stress biomarkers can be estimated from blood samples which can be collected remotely and estimated in the appropriate laboratory. In the literature, very less studies are performed on OCD detection through machine learning using oxidative stress biomarkers.

2. A recent study [10], finds evidence of genetic linkage of OCD in first-degree relatives of OCD patients. Therefore, machine learning techniques can be explored more to segregate healthy individuals, ocd patients, and suspected individuals. This may lead to early detection of OCD.

To handle the issues raised in this section, hybrid machine learning models can be proposed to handle the nonlinearity nature of the biomarker set.

Conclusion

This article presents a comprehensive analysis of the machine learning models used for OCD detection. The approaches are categorized based on the biomarkers it uses. The machine learning models are experimented with, outcomes are recorded and a comparative analysis was performed. The study tabulates and classifies the approaches chronologically, based on the machine learning models used. It further highlights the research gap and future research direction to improve the OCD detection process.

Acknowledgment

This work is supported by the Odisha Higher Education Programme for Excellence and Equity (OHEPEE) World Bank [No. 6770/GMU].

References

1. S. K. Kar, B. Kumari, and A. Singh, “Multimodal augmentation approach with transcranial direct current stimu-

lation in management of obsessive-compulsive disorder with depression and comorbid seizure disorder: A case report," Indian Journal of Psychological Medicine, vol. 44, no. 6, pp. 621–623, 2022.
2. S. G. Ball, L. Baer, and M. W. Otto, "Symptom subtypes of obsessive-compulsive disorder in behavioral treatment studies: A quantitative review," Behaviour Research and Therapy, vol. 34, no. 1, pp. 47–51, 1996.
3. C. Zhou, Y. Cheng, L. Ping, J. Xu, Z. Shen, L. Jiang, L. Shi, S. Yang, Y. Lu, and X. Xu, "Support vector machine classification of obsessive-compulsive disorder based on whole-brain volumetry and diffusion tensor imaging," Frontiers in psychiatry, vol. 9, p. 524, 2018.
4. X. Hu, Q. Liu, B. Li, W. Tang, H. Sun, F. Li, Y. Yang, Q. Gong, and X. Huang, "Multivariate pattern analysis of obsessive–compulsive disorder using structural neuroanatomy," European Neuropsychopharmacology, vol. 26, no. 2, pp. 246–254, 2016.
5. A. Shrivastava, A. K. Tripathy, and P. K. Dalal, "A svmbased classification approach for obsessive compulsive disorder by oxidative stress biomarkers," Journal of Computational Science, vol. 36, p. 101023, 2019.
6. K. Patel, A. K. Tripathy, L. N. Padhy, S. K. Kar, S. K. Padhy, and S. P. Mohanty, "Accu-help: A machine learning based smart healthcare framework for accurate detection of obsessive compulsive disorder," arXiv preprint arXiv:2212.02346, 2022.
7. M. Q. Hoexter, E. C. Miguel, J. B. Diniz, R. G. Shavitt, G. F. Busatto, and J. R. Sato, "Predicting obsessive– compulsive disorder severity combining neuroimaging and machine learning methods," Journal of affective disorders, vol. 150, no. 3, pp. 1213–1216, 2013.
8. S. Mas, P. Gasso, A. Morer, A. Calvo, N. Bargallo, A. Lafuente, and L. Lazaro, "Integrating genetic, neuropsychological and neuroimaging data to model early-onset obsessive compulsive disorder severity," PLoS One, vol. 11, no. 4, p. e0153846, 2016.
9. T. B. Altu˘glu, B. Metin, E. E. T¨ulay, O. Tan, G. H. Sayar, C. Ta¸s, K. Arikan, and N. Tarhan, "Prediction of treatment resistance in obsessive compulsive disorder patients based on eeg complexity as a biomarker," Clinical Neurophysiology, vol. 131, no. 3, pp. 716–724, 2020.
10. A. Shrivastava, S. K. Kar, E. Sharma, A. A. Mahdi, and P. K. Dalal, "A study of oxidative stress biomarkers in obsessive compulsive disorder," Journal of ObsessiveCompulsive and Related Disorders, vol. 15, pp. 52–56, 2017.
11. D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, Logistic regression. Springer, 2002.
12. L. E. Peterson, "K-nearest neighbor," Scholarpedia, vol. 4, no. 2, p. 1883, 2009.
13. A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," Journal of Chemometrics: A Journal of the Chemometrics Society, vol. 18, no. 6, pp. 275–285, 2004.
14. J. C. Principe, W. Liu, and S. Haykin, Kernel adaptive filtering: a comprehensive introduction. John Wiley & Sons, 2011.
15. B. Xu, J. Z. Huang, G. Williams, Q. Wang, and Y. Ye, "Classifying very high-dimensional data with random forests built from small subspaces," International Journal of Data Warehousing and Mining (IJDWM), vol. 8, no. 2, pp. 44–63, 2012.
16. Y. Yao and B. Zhou, "Naive bayesian rough sets," in International conference on rough sets and knowledge technology. Springer, 2010, pp. 719–726.
17. H. Abou-Warda, N. A. Belal, Y. El-Sonbaty, and S. Darwish, "A random forest model for mental disorders diagnostic systems," in International Conference on Advanced Intelligent Systems and Informatics. Springer, 2016, pp. 670–680.
18. K. Sekaran and M. Sudha, "Predicting drug responsiveness with deep learning from the effects on gene expression of obsessive–compulsive disorder affected cases," Computer Communications, vol. 151, pp. 386–394, 2020.
19. V. Laijawala, A. Aachaliya, H. Jatta, and V. Pinjarkar, "Classification algorithms based mental health prediction using data mining," in 2020 5th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2020, pp. 1174–1178.
20. H. Hasanpour, R. G. Meibodi, K. Navi, and S. Asadi, "Dealing with mixed data types in the obsessive-compulsive disorder using ensemble classification," Neurology, Psychiatry and Brain Research, vol. 32, pp. 77–84, 2019.
21. H. Yan, X. Shan, H. Li, F. Liu, and W. Guo, "Abnormal treatment response in patients with obsessive–compulsive disorder," Journal of Affective Disorders, vol. 309, pp. 27–36, 2022.
22. F. Lenhard, S. Sauer, E. Andersson, K. N. M˚ansson, D. Mataix-Cols, C. R¨uck, and E. Serlachius, "Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: A machine learning approach," International Journal of Methods in Psychiatric Research, vol. 27, no. 1, p. e1576, 2018.
23. S. V. Kalmady, A. K. Paul, J. C. Narayanaswamy, R. Agrawal, V. Shivakumar, A. J. Greenshaw, S. M. Dursun, R. Greiner, G. Venkatasubramanian, and Y. J. Reddy, "Prediction of obsessive-compulsive disorder: importance of neurobiology-aided feature design and cross-diagnosis transfer learning," Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, vol. 7, no. 7, pp. 735–746, 2022.
24. T. Najafi, R. Jaafar, K. Najafi, and F. Eslamdoust Siahestalkhi, "Brain waves characteristics in individuals with obsessive-compulsive disorder: A preliminary study." International Journal of Online & Biomedical Engineering, vol. 18, no. 1, 2022.
25. S. Aydin, N. Arica, E. Ergul, and O. Tan, "Classification of obsessive compulsive disorder by eeg complexity and hemispheric dependency measurements," International journal of neural systems, vol. 25, no. 03, p. 1550010, 2015.
26. D. F. Tolin, R. E. Brady, and S. Hannan, "Obsessional be-

liefs and symptoms of obsessive–compulsive disorder in a clinical sample," Journal of Psychopathology and Behavioral Assessment, vol. 30, no. 1, pp. 31–42, 2008.
27. K. D. Askland, S. Garnaat, N. J. Sibrava, C. L. Boisseau, D. Strong, M. Mancebo, B. Greenberg, S. Rasmussen, and J. Eisen, "Prediction of remission in obsessive compulsive disorder using a novel machine learning strategy," International journal of methods in psychiatric research, vol. 24, no. 2, pp. 156–169, 2015.