ARTIFICIAL NEURAL NETWORK BASED PREDICTION OF TREATMENT RESPONSE TO REPETITIVE TRANSCRANIAL MAGNETIC STIMULATION FOR MAJOR DEPRESSIVE DISORDER PATIENTS

by

Dana Bazazeh

B.Sc., Khalifa University, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis/dissertation entitled:

Artificial Neural Network based prediction of treatment response to repetitive Transcranial Magnetic Stimulation for Major Depressive Disorder patients

submitted by Dana Bazazeh in partial fulfillment of the requirements for

the degree of Master of Applied Science

in Electrical and Computer Engineering

Examining Committee:

Dr. Jane Wang
Co-supervisor

Dr. Rabab Ward
Co-supervisor

Dr. Shahriar Mirabbasi
Supervisory Committee Member

Additional Examiner

Additional Supervisory Committee Members:

Supervisory Committee Member

Supervisory Committee Member

Abstract

Major Depressive Disorder (MDD) is a severe medical condition that affects thousands of people every year. Therapy in MDD includes medication and psychotherapy, and is prescribed on the basis of the type and severity of depressive episodes. Treatment-resistance is common among MDD patients. Repetitive Transcranial Magnetic Stimulation (rTMS) is a form of deep brain stimulation used for relieving depressive symptoms. Due to its high cost and lengthy procedure, it's reserved for patients showing treatment-resistance to at least 2 trials of antidepressants. Of all MDD patients, only 50% show response to rTMS, which leads to unnecessary patient frustration and additional costs. Prediction of resistance to rTMS treatment can thus help physicians decide on the best treatment course for each patient. This thesis presents a machine-learning based clinical assistive tool that predicts the probability of a patient to respond to rTMS treatment and if so, predict the probability whether they are likely to achieve remission. The most relevant clinical and sociodemographic variables associated with predicting treatment outcomes were selected on the basis of importance scores ranked using a Random Forest (RF) algorithm, and an elaborative analysis of their significance was presented. The most important variables were fed into a Deep Artificial Neural Network (DANN) for classification of patients who will respond to rTMS treatment. Two DANN variants were designed, trained, optimized and tested to predict each of rTMS treatment response and remission outcomes. Our model is based on the pre-treatment clinical and sociodemographic data which had been collected from 414 patients diagnosed with treatment-resistant MDD. Results show that our DANN model outperforms existing clinical procedures and yields an accuracy of 84.4% in predicting remission and 73.8% in distinguishing responders form non-responders. Additionally, a thorough evaluation and comparison with other

methods that have used machine learning algorithms to predict rTMS treatment outcome was carried and discussed in detail. Findings in this thesis signify the potential of individual-based assessments that can improve rTMS treatment procedure.

Lay Summary

Major Depressive Disorder (MDD) is a serious condition that affects people of all ages across the world. Several treatments are available for reducing depressive symptoms among patients, but unfortunately many patients demonstrate resistance to therapeutic treatments. Recently, physicians have started directing patients with treatment-resistance to antidepressants towards undergoing deep brain stimulation therapy such as repetitive Transcranial Magnetic Stimulation (rTMS). Albeit beneficial to some, 50% of patients fail to respond to a complete course of rTMS treatment. Clinical assistive tools can be utilized to provide aid to physicians in planning adequate treatment plans for patients on a case-by-case basis. In this thesis, we propose a clinical tool that takes in a patient profile consisting of sociodemographic and clinical characteristics, and outputs a prediction of the patient's most likely outcome to rTMS treatment. The findings in this thesis were evaluated and highlight the benefits of integrating machine-learnt models into clinical decision making.

Preface

This thesis is the outcome of collaborations with my co-supervisors Dr. Jane Wang and Dr. Rabab Ward and our medical collaborator Dr. Fidel Vila-Rodriguez. All chapters in this work have been written by me and guided with input from my supervisors. The research task was proposed by my supervisors and I was responsible for designing solutions, conducting tests and evaluating all the results. Data used in this work was obtained from Dr. Fidel Vila-Rodriguez as part of a THREE-D study in collaboration between University of British Columbia Hospital, Vancouver, BC, Centre for Addiction and Mental Health, Toronto, ON and Toronto Western Hospital, Toronto, ON. Clinical Research Ethics Board (CREB) approval was obtained by Dr. Fidel Vila-Rodriguez with CREB number H13-02340. All supervisors; Dr. Jane Wang, Dr. Rabab Ward and Dr. Fidel Vila-Rodriguez have provided constant feedback and guidance during the design, implementation and review phases in this work. Fellow colleague Xiang Liu has also contributed with ideas during brainstorming meetings with supervisors.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AUC	Area Under the Curve
AI	Artificial Intelligence
ANN	Artificial Neural Network
BSI-A	Brief Symptom Inventory-Anxiety subscale
DANN	Deep Artificial Neural Network
DL	Deep Learning
ECT	Electroconvulsive Therapy
EEG	Electroencephalogram
FDA	U.S. Food and Drug Administration
HRSD	Hamilton Rating Scale for Depression
Hz	Hertz
IDS	Inventory of Depressive Symptoms
iTBS	intermittent Theta Burst Stimulation
LR	Logistic Regression
LSTM	Long Short Term Memory
MADRS	Montgomery-Asberg Depression Rating Scale
MDD	Major Depressive Disorder
MFA	Mixture of Factor Analysis
ML	Machine Learning
mRmR	minimal-Redundancy maximal-Relevance
NPV	Negative Predictive Value

PCA	Principal Component Analysis
PPV	Positive Predictive Value
QEEG	Quantitative Electroencephalography
QIDS	Quick Inventory of Depressive Symptoms
RF	Random Forest
RMT	Resting Motor Threshold
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
rTMS	repetitive Transcranial Magnetic Stimulation
SMOTE	Synthetic Minority Over-sampling Technique
STAR*D	Sequenced Treatment Alternatives to Relieve Depression
STD	Standard Deviation

Acknowledgements

I would like to express my deepest gratitude to my supervisors, *Dr. Jane Wang* and *Dr. Rabab Ward* for providing me with this great opportunity and for their continuous support and guidance throughout my program. Without them, the work presented here would not have been possible.

Also, many thanks to *Dr. Fidel Vila-Rodriguez* for his help and support throughout this work and for providing me with continuous and valuable feedback.

I would like to also thank my mentor *Dr. Raed Shubair* for advising me throughout this time and mentoring me to become the best version of myself. His work ethic and passion were inspirational.

To all my *colleagues* who helped me improve myself and push my boundaries, and all my *friends* who made me feel loved and supported during this challenging chapter of my life.

Last but not the least, to my *family* who have made many sacrifices so that I can fuel my ambitions. I simply cannot repay all the nights my *parents* spent nurturing me to become the person I am today and for giving me the uttermost love and support that helped me grow. I owe everything to them.

Dedication

To my beloved parents

Chapter 1: Introduction

1.1 Overview

Major Depressive Disorder (MDD) is the fifth most leading cause of disability around the world, as assessed based on years lived with disability, and its adverse effect disrupt the daily lives of hundreds of millions of people every year [1]. Studies have shown that most MDD patients (70-90%) are unable to achieve remission given an initial treatment course [2-5]. This outlines the need for designing methods to assist physicians in the decision making process when it comes to depression treatment plans. Repetitive Transcranial Magnetic Stimulation (rTMS) is a non-invasive brain stimulation technique approved by Health Canada and FDA (U.S. Food and Drug Administration), and used for the therapeutic treatment of MDD [6]. rTMS treatment is commonly used when a patient is deemed unresponsive to an adequate dose and trial of antidepressants. Response to treatment is standardly defined as a 50% reduction in score from an initial baseline using one of the known rating scales [7-9]. Recent studies have showed response rate of 50-55% and remission rate of 30-35% to rTMS treatment on average [10-12]. rTMS treatments are costly and time-consuming, which has pushed researchers to determine ways to predict early on the treatment outcome to rTMS for each patient. Different studies have attempted to predict response to rTMS using biomarkers based on neurochemical, neurophysiological, and neuroimaging measures [13]. However, obtaining biomarker-related data is both expensive and complicated. Alternatively, clinical data can be obtained quickly, economically and only require a minimal amount of preprocessing.

Machine Learning (ML) is the science of enabling computers to self-learn, by giving them examples in the form of data samples to initiate this learning process [14]. The ML field consists

of many algorithms that can be used for different learning tasks. Deep Learning (DL) is a subfield of ML that emulates the way a human brain learns new information. A typical DL model consists of several layers of increasing abstraction stacked in sequence, and usually requires large amounts of data to learn a given task [15]. ML and DL algorithms have been rigorously applied in the medical field due to their ability to detect complex patterns in large and noisy data. This had led the introduction of a surge of successful applications ranging from image processing and segmentation to diagnosis and treatment outcome prediction. Similarly in studies of depression, applications of ML algorithms have seen wide use for treatment outcome analysis and symptom significance [16].

1.2 Motivation & Scope

Assistive tools can help guide a physician's decision towards the most appropriate treatment course, which would eventually improve the efficacy of rTMS treatment. Meanwhile, patients with treatment-resistance to rTMS can be directed to other forms of therapy that can be deemed more effective. This will help cut-down time and costs associated with ineffective treatment that usually expands over multiple weeks. Recent studies that aimed to predict treatment outcome to rTMS relied mainly on electroencephalogram (EEG) pre-treatment data with small datasets [17]. EEG data is more expensive and cumbersome to collect and analyze when compared to clinical data. Based on our extensive research and knowledge, there has been no work focused on predicting rTMS treatment outcome using only clinical and/or sociodemographic data collected at baseline. Prediction of treatment outcome in general could be approached in two different ways. One such way is using standard statistical techniques that require constant input from domain experts. Here, the underlying correlation between predictive variables and the treatment outcome variable is

investigated sequentially, and such techniques are only feasible when variable relationships are linear in nature. Future predictions using this approach call for an in-depth analysis of a patient profile made by a skilled physician, which is costly and time consuming, as well as prone to high levels of human error. On the other hand, ML based prediction has shown huge success in the medical field, as it allows for multiple and simultaneous analysis of predictor-outcome relationships. In many medical tasks such as diagnosis and prognosis, prior assumptions regarding relationships between variables can't be made, due to the heterogeneity of a disease. Therefore ML algorithms can be used to model such high complex relationships and approximate functions to a high degree of accuracy. Performance of ML models are quantifiable and allow for easy comparison with different models. In addition, a well-trained model can make new predictions instantly, with a high degree of accuracy and without any additional analysis or input from specialized physicians. Given all these advantages, we determined that investigating different ML algorithms would be the best plausible approach to solve the problem of predicting treatment outcome in depression.

In this thesis, we present a novel ML-based framework that aims to provide high accuracy predictions of outcome to rTMS treatment using only a small number of patient characteristics that can be collected through assessment forms within 10 minutes.

1.3 Problem at Rest

Our assumption is that a combination of carefully selected pre-treatment sociodemographic and clinical data can sufficiently be used to design and drive a ML model to accurately predict treatment outcome to rTMS therapy. This thesis explores the following research questions:

- RQ1. *What are the most important features that could be associated with rTMS treatment outcome for each of response and remission states?*

  We use ML to perform an analysis of all collected features and symptoms relevant to rTMS and independently analyze the effect of each feature on treatment outcome.

- RQ2. *How can we design a clinical predictive tool for predicting rTMS outcome?*

  We use a combination of ML and DL techniques to develop an end-to-end framework that is trained and tested to accurately predict rTMS treatment outcome using a refined feature set of only the most predictive features.

1.4 Contributions

This thesis makes the following main contributions:

- Provide a refined, cleaned dataset consisting of relevant sociodemographic and clinical features, labeled with appropriate response and remission rates for 414 patients diagnosed with MDD.
- Design and develop a novel prediction framework that selects important features associated with rTMS treatment outcome and uses them as input to drive a DL model, which is trained and tested using multiple evaluation metrics.
- Offer an elaborative analysis of most important predictors of rTMS treatment outcomes and justifications based on recent research studies.
- Provide a systematic comparison of several ML algorithms used to drive rTMS treatment outcome predictions based on our private dataset.

1.5 Thesis Organization

The rest of the thesis is structured as follows: Chapter 2 provides a detailed description of rTMS treatment, as well as the most popular ML algorithms used in the medical field. In addition, Chapter 2 discusses related work that also utilize ML algorithms to make predictions about depression treatment outcome. Chapter 3 explains our proposed approach in detail and includes a description of data collection, model design and architectural aspects of our method. Chapter 4 reports our results using a comprehensive set of evaluation metrics and provides a thorough discussion of the limitations and constraints of our approach. Finally, Chapter 5 concludes our work and describes potential future work.

Chapter 2: Background & Related Work

2.1 Depression Treatment Overview

Several treatment options are in place to relieve the symptoms of depression, and with the help of a qualified physician, patients with MDD can explore these possibilities based on their unique profile and treatment history. Psychotherapy is a popular treatment given to patients with depression, and is informed by psychiatrists stirring cognitive, behavioral and interpersonal talks with patients in order to achieve mindfulness. Many physicians encourage the combination of psychotherapy with other forms of treatment, such as antidepressants, to improve efficacy [18]. Antidepressants are among the most commonly prescribed medication for depression, and consist of a large family of drugs with different strength and effects. The main group of antidepressants include Tricyclic Antidepressants (TCAs), Selective Serotonin Reuptake Inhibitors class (SSRIs), Mono-Amine Oxidase Inhibitors (MAOIs) and 'atypical' antidepressants [19]. Patients who develop treatment-resistance to antidepressants seek other effective forms of therapy. Transcranial Magnetic Stimulation (TMS) is a therapeutic procedure that works by positioning magnetic fields to penetrate a specific region of the brain, which has been associated with depression, in order to stimulate nerve cells that can help reduce depressive symptoms [20][21]. Albeit non-invasive, TMS is usually reserved for when antidepressants deem ineffective, mainly because of its high cost. Lastly, Electroconvulsive Therapy (ECT) is a technique in which electric currents are sent through the patient's brain, causing a small seizure to change the biochemical brain state and relieve depressive symptoms. Although highly effective, ECT is considered risky considering its need for general anesthesia and serious side effects such as amnesia and disorientation, which is why physicians withhold it for only the most severe, treatment-resistant cases of depression [22].

2.2 Transcranial Magnetic Stimulation

2.2.1 Repetitive Transcranial Magnetic Stimulation

Repetitive Transcranial Magnetic Stimulation (rTMS) is a non-invasive therapeutic procedure for treatment-resistant MDD [23] It works by influencing the brain's cerebral electrical activity using a repetitive pulsed magnetic field that transports across the scalp, skull and into the prefrontal cortex part of the brain [24]. rTMS has been extensively researched and approved by the US Food and Drug Administration (FDA) in 2008 for clinical practice as a safe treatment for MDD [25]. A complete treatment course using rTMS can add up to $12,000, which makes it more costly than traditional antidepressants treatments [26]. As a result, rTMS is usually recommended to a patient only after they fail to respond to at least 2 antidepressant trials of adequate dose and strength [23]. The most widely adopted protocol uses 10 Hz stimulation frequency with 4s on and 26s off bursts of stimulation for a total of 3,000 pulses per session, which takes around 37.5 mins to deliver [27]. This is done with a stimulation intensity of 120% resting motor threshold (RMT). Since rTMS is based on magnetic fields penetrating the brain, people with implanted metals inside the head or neck should avoid this treatment. rTMS has proved successful in patients with treatment-resistant depression and has minimal side effects, however, its efficacy is highly dependent on parameters such as type of coil, placement position, stimulation frequency, strength and length [23]. The NeuroStar TMS Therapy System (Neuronetics, Malvern, Pennsylvania) was the first device approved by the FDA to perform rTMS, and was later followed by several other manufactured devices [28].

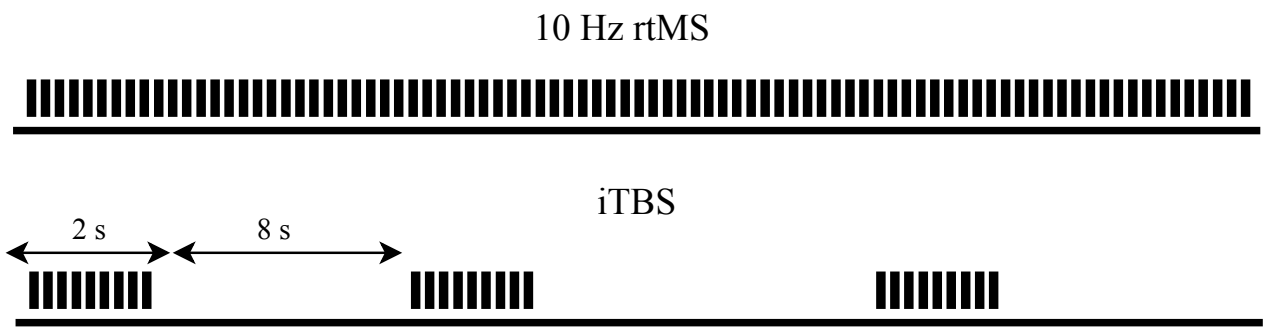


Figure 2.1 Difference in burst transmission rate between rTMS and iTBS

2.2.2 Intermittent Theta Burst Stimulation

Intermittent Theta Burst Stimulation (iTBS) is a variation of the standard rTMS protocol that has seen a lot of popular use in clinical practice. This was the result of several attempts to improve the treatment efficacy of rTMS. The main difference between rTMS and iTBS is the length of the protocol. iTBS is proved to be much faster with better tolerability [29][30]. In iTBS, triplet bursts are emitted at a frequency of 50 Hz for every 200ms time interval, which is repeated at 5 Hz, 2s on and 8s off for a total of 600 pulses [29]. The total duration for a single session of iTBS is around 3 mins, which improves upon patient tolerability and comfort level. This great reduction in procedural length means that treatment centers can facilitate more number of people per day, which makes it more cost-effective, when compared to standard rTMS. A recent study has shown the non-inferiority of iTBS when compared with standard 10 Hz rTMS treatment [31].

2.3 Machine Learning

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that combines algorithms capable of learning from observed data instances and making predictions about new unseen data. The learning process in ML is driven by the availability of large amounts of data samples, and a

model is designed to extract relevant information and relations from within the data, which is then used in the form of a general model that forecasts the behavior of new similar data. The architecture of a ML model is highly dependent on the complexity of a given problem and the characteristics of the data associated with it. ML algorithms can be divided into two main categories; supervised [32] and unsupervised learning [33]. Supervised learning is the most common type of ML and consists of input features X = {x1,x2,…,xi} extracted from a data set and paired with labeled output variables Y, and the ML algorithm works toward learning the mapping between X and Y. Classification and regression problems come under the family of supervised learning [34][35]. On the other hand, in unsupervised learning, data consists of only the input features X, without any labeled output variables Y. An unsupervised model tries to learn the underlying structure of the data to extract information. Data clustering is one of the most popular unsupervised learning problems, where data is grouped together based on shared similarities [36]. ML is widely applied in many domains today, and extensively used within the medical field. Its applications ranges from image processing and segmentation to disease diagnosis and prognosis. We will further discuss in more detail a number of ML algorithms that are under investigation within this thesis.

2.3.1 Random Forest

Random Forest (RF) is a supervised learning algorithm that uses a simple decision tree as its building block. It is an ensemble learning technique based on combining several decision trees together, where generally a large number of trees is preferred to yield good results [37]. Every decision tree in the ensemble is independently constructed using a subset of the data instances with the bootstrapping technique, where data is sampled with replacement from the main dataset. An individual decision tree may also only include a subset of the variables in the main data set, and

this parameter can be optimized when designing a RF. Assume we have a data set with n samples $\{(X_1, y_1),\ (X_2, y_2),\ldots,\ (X_n, y_n)\}$, where X_i is an input vector having v variables/features and represented as $X_i = \{\ x_{i1}, x_{i2}, \ldots, x_{iv}\}$ and $y_{i=}\{0,1\}$ represent a binary output class label. To build T number of trees using bootstrapping technique means we need T number of bootstrap datasets, one per tree [38]. These datasets should contain n samples, drawn from the original dataset with replacement, which means duplicate samples are possible. Each of the bootstrap datasets may only have m variables/features, obtained as a square root of the full list of v variables ($m = \sqrt{v}$). Once all the trees T are trained and optimized using a training set, each individual decision tree gives out its own predicted output variable when given new data, and the final output of RF is a majority vote across $Y = \{\ y_1, y_2, \ldots, y_T\ \}$ representing all the individual trees outputs. For every sample input-output pair $(X_\mathrm{i}, y_\mathrm{i})$, there exists K bootstrap datasets that do not have the sample $(X_\mathrm{i}, y_\mathrm{i})$, and these K datasets are known as the out-of-bag sets. The out-of-bag error (OOBE) is then estimated by the classification of all possible n samples $(X_\mathrm{i}, y_\mathrm{i})$, each using only their own of out-of-bag sets K_i. This OOBE is used to reflect the generalization error of the RF classifier, and an optimizer can help reduce this error. Due to its speed, intuition and flexibility in combining categorical and continuous data, RF have seen popular use in applied ML for both classification and regression problems. A detailed description of the algorithm along with its parameters can be found in [37].

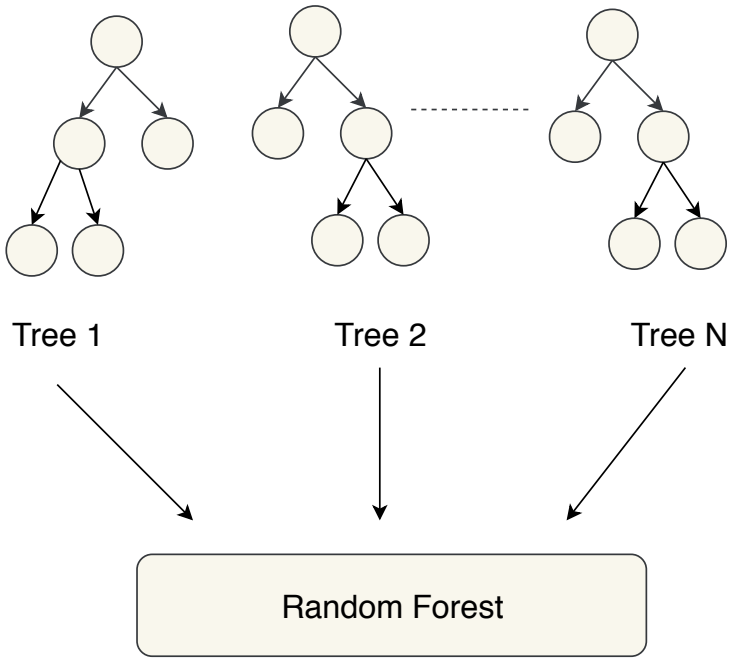


Figure 2.2 Visualization of Random Forest architecture

Simplified Random Forest Pseudocode
Procedure RF **for t in Trees** Sample n data points D_n from D with replacement Randomly select V_m variables from variable set V_o, where m = sqrt(o) Using D_n samples with V_m features, construct decision tree Cast output vote Minimize out of bag error **end loop** **return majority vote across all Trees** **end procedure**

Table 2.1 Random Forest simple pseudocode

2.3.2 Logistic Regression

Logistic Regression (LR), unlike its name implies, is a classification ML algorithm based on a statistical analysis approach. In binary LR, a collection of continuous independent variables is used to predict a dichotomous dependent variable. It differs from linear regression in that it is a probabilistic based approach, where the output is in the form of a probability representing which output class a given data point belongs to. Mathematically, this is represented using the steps below as described in [39]

For simplicity, we assume a binary dependent variable Y with two independent variables $x1, x2$. Given that p is the probability that the binary output class for a specific data sample with feature space X is 1.

$$p = P(Y = 1 \mid X = x1, x2)$$

The logistic response function is defined below where β_n represent the coefficients of the logistic function, found using maximum-likelihood estimation.

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}$$

To transfer the response function to linearity, we introduce the odds ratio term $\frac{p}{1-p}$

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

Finally, a logistic model can be estimated with the below logistic equation

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

LR performs well with noisy data and is fairly simple to use. Variable coefficients estimated using maximum-likelihood, albeit hard to interpret, can provide some insight to the way features interact to predict the outcome variable. For this reason, LR has seen increasing use in the medical field, and in depression related studies [40][41].

2.3.3 Artificial Neural Network

Artificial Neural Network (ANN) [41] is a ML-based framework that was initially inspired by the way the neural system works in the human brain. It consists of many nodes that are interconnected together in a hierarchical layered architecture. A standard ANN has an input layer, hidden layer and an output layer, all of which are fully connected such that every node is connected to all the nodes in the succeeding layer. There are usually as many nodes in the input layer as there are variables in the variable space X. Nodes are computational units that take in weighted inputs and depending on a given threshold, will sometimes activate releasing an output value. The activation function is modeled as below:

$$f(z) = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

The activation function $f(z)$ usually has a limit between 0 and 1 for a sigmoid activation or -1 and 1 for a tanh function. Here n represents the number of inputs to the node, w_i is the weights

associated with each input, x_i the input value and b the added bias. The linear activations are fed into the succeeding layer in the architecture, where the weighted sum of these activations are fed into the second layer's activation function and so on until it propagates to the output layer. The weights of the system are initialized and then adjusted by backpropagating the classification error calculated at the output layer for every data instance or epoch that the model inputs. This is iteratively done for all samples in the training set, with the aim of minimizing a given cost function to achieve higher classification accuracy. Deep Artificial Neural Networks (DANN) is the term given to a specific variant of an ANN where there is more than a single hidden layer [15]. Number of hidden layers are increased to account for increased abstraction and complexity. ANNs have several user defined hyperparameters which should be tuned to reduce error and increase accuracy. Parameters to be tuned include number of epochs, learning rate, step size, tolerance, momentum and alpha variable. ANNs can approximate most functions with a high generalization capacity.

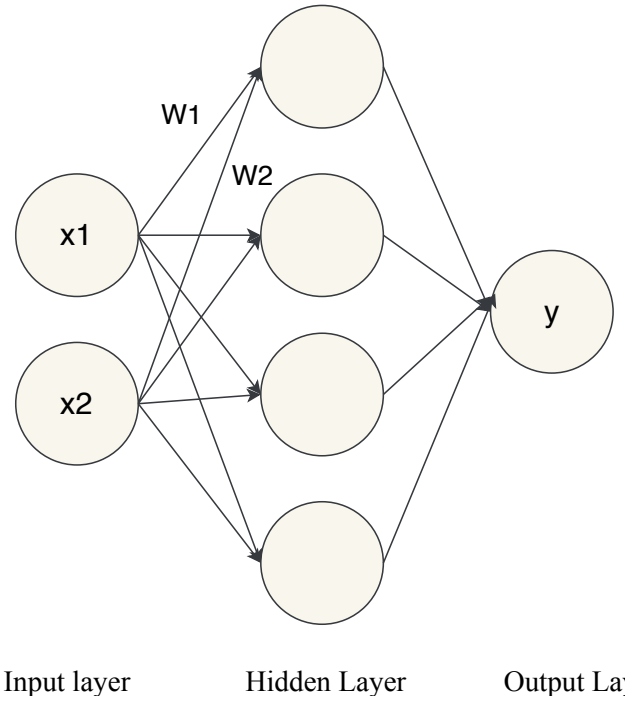


Figure 2.3 Simple multiple layer perceptron neural network

2.4 Related Work

There has been few studies that explored the use of ML algorithms in predicting the outcome of different depression therapies. These studies can be divided into two main categories. The first cluster of work relates to predicting therapy outcome of MDD patients undergoing antidepressant treatment courses using sociodemographic and clinical data, including self-reported depression symptom inventories [43-47]. In [41] data from level 1 and 2 of the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study (n = 2,876) was utilized to construct a LR model that predicts remission rates, which showed promising results with an Area Under the receiver operating characteristic Curve (AUC) of 0.719. Similarly in [43], data from level 1 of the STAR*D trial (n = 1,949) was used and the most predictive variables/features consisting of sociodemographic and diagnostic clinical features were identified using an elastic net model. The features were then fed into a gradient boosting machine, which is an ensemble ML approach that combines weakly trained decision trees together. Their external validation results using the Combining Medications to Enhance Depression Outcomes (COMED) cohort data showed an accuracy of 59.6%. Another study [44] focused on independently collecting clinical and demographic data from 552 patients diagnosed with treatment resistant MDD using the Montgomery-Asberg Depression Rating Scale (MADRS) and fitted a random forest classifier to predict treatment outcome, which achieved an overall accuracy of 75% using internal 10-fold cross validation.

The second cluster of work mainly focused on using electroencephalogram (EEG) pre-treatment data to predict outcome of rTMS treatment [17][47][48] . In the most recent and largest of those studies [47], 147 MDD subjects were recruited and pre-treatment quantitative

electroencephalography (QEEG) data was extracted to fit a Support Vector Machine (SVM) classifier which had an overall accuracy of 86.4% and an AUC of 0.92 in predicting response to rTMS treatment. In [17], 55 patients with MDD were recruited and pre-treatment QEGG cordance data was collected and used to construct an ANN with 10 hidden layer neurons and a sigmoid activation function, using the trainlm function. The model achieved an overall accuracy of 87.27% in identifying responders from non-responders, using internal 10-fold cross validation. Finally in [48], pre-treatment resting EEG data was initially collected from 27 participants and consisted of 142 features. A dimensionality reduction technique was performed using the minimal-redundancy-maximal-relevance criterion (mRmR) that reduced the number of features to only four with the highest relevance score. The top four features were then used to fit a response prediction model using the Mixture of Factor Analysis (MFA) technique. The model was internally validated using a Leave-2-out cross-validation technique and achieved an overall accuracy of 80%. Table 2.2 below summarizes all the related work, describing in more detail their objective, data type, data size, approach used to solve their task and their final results. The main challenge in this problem is attempting to approach a complex task such as predicting rTMS treatment outcome using only clinical and sociodemographic data at baseline. As can be seen in Table 2.2, all other attempts to predict rTMS outcome consisted of collecting EEG data, which is expensive and has a large overhead preprocessing cost. Additionally, these two studies included a small sample size (147 and 27 samples), which are too small to draw meaningful generalizations from. Based on our thorough research, we did not identify any work that attempts to sufficiently use clinical and/or sociodemographic data at baseline to predict response or remission rates for rTMS based therapeutics, which in turn highlights the novelty of this work.

Related Work	Objective	Data	Data Dimension (samples x features)	Data Type	Machine Learning Technique/Model	Results
Antidepressant treatment outcome prediction						
Chekroud et. al 2016 [43]	Predict antidepressant (citalopram) treatment remission outcome	Clinical + Sociodemographic	1949 x 25	Continuous + Categorical	Gradient Boosting Machine	Accuracy 64·6%
Kautzky et. al 2018 [44]	Predict antidepressant treatment response outcome	Clinical + Sociodemographic	552 x 47	Continuous + Categorical	Random Forest	Accuracy 75.0%
Iniesta et. al 2016 [45]	Predict antidepressant treatment remission outcome	Clinical + Demographic	793 x 41	Continuous + Categorical	Elastic Net	AUC 72.0%
Perlis 2012 [46]	Predict antidepressant treatment remission outcome	Clinical + Sociodemographic	4023 x 15	Continuous + Categorical	Logistic Regression	AUC 71.2.%
Repetitive Transcranial Magnetic Stimulation (rTMS) treatment outcome prediction						
Erguzel et al. 2015 [47]	Predict rTMS treatment response outcome	EEG	147 x 6	Continuous	Support Vector Machine	Accuracy 86.4%
Khodayari-Rostamabad et al. 2011 [48]	Predict rTMS treatment response outcome	EEG	27 x 4	Continuous	Mixture of Factor Analysis	Accuracy 80.0%
Bazazeh et al. 2018 (our work)	Predict rTMS treatment response + remission outcome	Clinical + Sociodemographic	414 x 15	Continuous + Categorical	Deep Artificial Neural Network	Response: Accuracy 73.8% Remission Accuracy 84.4%

Table 2.2 Summary of related work

Chapter 3: Proposed Approach

3.1 Overview

In this thesis, we propose a clinical assistive tool, with high accuracy and low cost for predicting treatment outcome to rTMS treatment. The core of our approach is built on machine learning algorithms that is centered around learning to differentiate between responders and non-responders in a pre-treatment setting. Figure 3.1 shows an overview of our proposed approach. Given a database of raw patient characteristics, the first task consisted of cleaning and reformatting the database structure to make it more readable and machine ready. An initial filtration was done to eliminate noisy and irrelevant variables, using the opinion of medical experts to clarify and eliminate unnecessary variables as well as remove patient records with missing data. The reduced dataset was then fed into a RF model, where ensemble learning allowed for variable ranking based on how each variable/feature improved the purity of a tree node. Using two RF models, top 15 features for each of response and remission outcomes were extracted into two refined datasets. Due to imbalances in the datasets with low occurrence of response and remission instances, both datasets were inserted into an over-sampling function to avoid hindering the final model performance. Subsequently, the balance-adjusted datasets were each inserted into a stand-alone DANN. Each DANN had a repetitive tuning process, where selected model hyperparameters underwent performance optimization. Once the final hyperparameter settings was found, models were trained and tested using a cross-validation technique. The datasets were also analyzed using RF and LR models for providing a comparative view. Performance analysis consisted of reporting on evaluation metrics such as accuracy, sensitivity and specificity, and examining ROC curves.

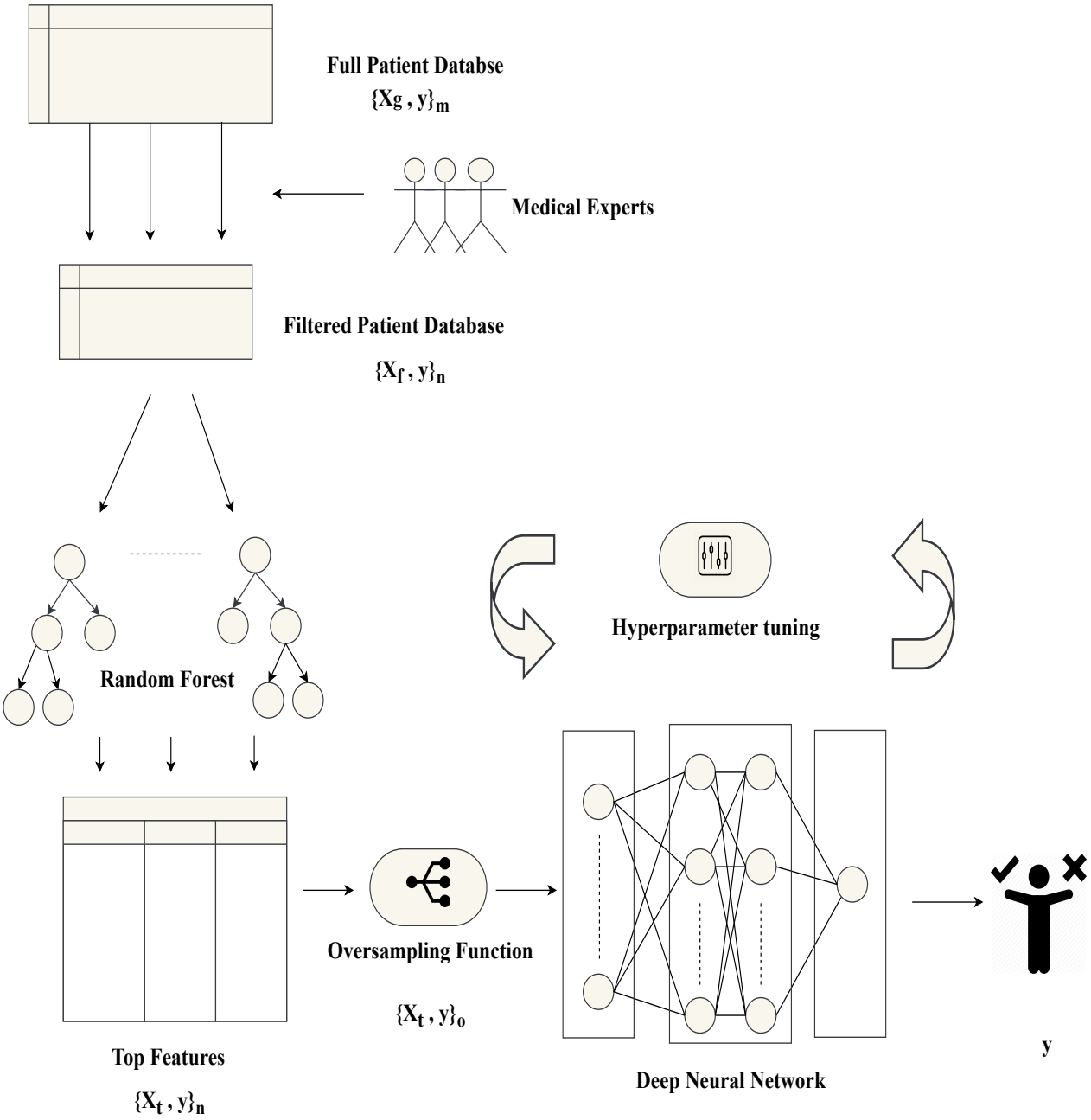


Figure 3.1 Our proposed system framework

3.2 Participant Data

A total number of 501 patients aged 18-65 were recruited at 3 Canadian Hospitals; Centre for Addiction and Mental Health, Toronto, ON, Toronto Western Hospital, Toronto, ON and University of British Columbia Hospital, Vancouver, BC. All selected patients were diagnosed with MDD using the Mini-International Neuropsychiatric Interview [49]. Inclusion criteria is a list of primary participant features or characteristics needed to answer certain study questions. The inclusion criteria here consisted of the following; inability to withstand two antidepressant courses with adequate dosage and length, showed at least a score of 18 for depression severity using the 17-item Hamilton Rating Scale for Depression (HRSD-17) [9] and having to undergo a fixed antidepressant course at least 4 weeks before rTMS treatment and throughout the treatment. Exclusion criteria is a list of characteristics found among participants who have successfully met a study's inclusion criteria, but have additional undesirable characteristics that can hinder the success of the study. The exclusion criteria here consisted of the following; previous diagnosis of a personality, bipolar or psychotic disorder or any existing substantial neurological or medical illness or anomalous serology. Patients were also excluded if they had undergone drug abuse within a 3 month period prior to treatment initiation, demonstrated suicidal behavior, pregnancy, undergone previous rTMS therapy, shown treatment resistance to electroconvulsive therapy (ECT) or at least 3 antidepressant trials, or were consuming 2 mg or more of any kind of anticonvulsant such as lorazepam. For safety reasons, patients with any kind of metallic substances in their head or neck such as a pacemakers or metallic implants were also excluded.

3.3 TMS Procedure & Outcome Measurement

A total of 414 patients were identified after evaluation of the inclusion and exclusion criteria. Patients were then equally assigned to treatment groups based on a randomization process that either placed the patient to receive standard 10 Hz rTMS treatment or iTBS treatment, and these two groups were balanced (1:1) according to the degree of resistance to antidepressants. This randomization process consisted of creating a randomization table using randomly generating permutation blocks obtained with a computer-based algorithm and grouped according to study site. The rTMS therapy was delivered using either the magnetic stimulator MagPro X100 or the R30 stimulator (MagVenture, Farum, Denmark) which targeted the left dorsolateral prefrontal cortex region of the brain. The visor neuronavigation system (ANT Neuro, Enschede, Netherlands) was used for an accurate positioning of the stimulator coil with the prefrontal cortex of each patient. Depression severity was assessed for each patient at baseline before commencing the treatment trial using three different scales; the 17-item Hamilton Rating Scale for Depression (HRSD) [9], the self-rated 16-item Quick Inventory of Depressive Symptoms (QIDS) (27) [8] and the 30-item Inventory of Depressive Symptoms (IDS-30) [7]. Anxiety has been previously correlated with depression treatment response and was hence also assessed using the Brief Symptom Inventory Anxiety Subscale (BSI-A) [50]. The delivery of standard rTMS using 120% RMT stimulation intensity consisted of 3,000 pulses per session at a 10 Hz frequency with 4s on and 26s off, and an average of 37.5 mins per session [51][52]. The iTBS treatment using 120% RMT stimulation intensity was completed at a significantly faster pace than that of rTMS, with an average of 3.15 mins per session and consisted of 600 total pulses of 50 Hz burst triplets. These triplets were repeated with a 5 Hz frequency, 2s on and 8s off [29]. Each patient initially received 20 treatment sessions, one per day for a total of 5 days/sessions per week, which was continued through a period

of 4 weeks. Depression severity was measured using IDS-30, BSI-A, HRSD and QIDS scales and the results were assessed for each patient at the end of the treatment course by the same trained staff member who had initially assessed their baseline scores. Patients who were not able to achieve remission but had 30% or more drop in their depression severity scores as reported by any of the three depression inventories, were given 10 treatment sessions in addition to the initial 20 sessions. Treatment outcomes were recorded at the end of the treatment trial for each patient and consisted of the dichotomous variables representing response and remission. Remission is achieved with a score of HRSD-17 <8, IDS-30 <14, or QIDS <6, and response was defined as a 50% decrease in depression severity based on scores from HRSD-17, IDS-30 or QIDS.

Feature	**Mean**	**Standard Deviation**
Age	42.4	11.5
Years of education	16.3	3.2
Age of onset	21.1	11.3
Episode duration in months	23.4	27.3
Baseline HRSD-17 score	23.7	4.4
Baseline QIDS-SR score	17.2	4.6
Baseline IDS-30 score	39.6	10.1
Baseline BSI-A score	10.2	5.4
Stimulation intensity	52.3	48.0
Antidepressants Treatment History Form: Strength Score	6.3	3.4
Years of education	16.3	3.2

Feature	Total	%
Male	168	40.5%
Level of education		
Grade 6 or less	2	0.5%
Grade 7-12	8	1.9%
High school graduate	31	7.5%
College dropout	71	17.1%
2 year college degree	53	12.8%
4 year college degree	125	30.2%
Graduate program dropout	34	8.2%
Graduate school degree	90	21.7%
Financial status		
Employed	150	36.2%
Unemployed	41	9.9%
Ontario disability support program	39	9.4%
Ontario works financial aid program	15	3.6%
Insurance disability	85	20.5%
Spouse support	35	8.5%
Family support	49	11.8%
Standard 10Hz rTMS	205	49.4%
Received pharmacotherapy during treatment		
Benzodiazepine	139	34%

Antidepressant	318	77%
Antidepressant combination	91	22%
Antipsychotic augmentation	77	19%
Lithium augmentation	13	3%
Antidepressants Treatment History Form: Trials		
One failed antidepressant	185	45%
Two failed antidepressants	116	28%
Three failed antidepressants	81	20%
Unable to tolerate two trials	32	8%
Antidepressants Treatment History Form: High		
0	1	0.24%
1	15	3.8%
2	19	4.6%
3	149	36%
4	199	48%
5	31	7.5%
Treatment outcomes		
Response rate	152	36.7%
Remission rate	97	23.4%

Table 3.1 Baseline sociodemographic and clinical data

3.4 Data Description & Predictor Selection

Data collected from patients included a combination of sociodemographic and diagnostic clinical features, this was in addition to the depression severity scores deduced from the three depression-symptom inventories (IDS-30, HRSD-17, QIDS) and the anxiety score measured by BSI-A, all of which were collected at baseline. The initial patient dataset can be represented as $\{X_g, y\}_m$, where X_g is the input space, having $g = 117$ variables/features, y is the binary output response and $m = 414$ is the total number of patient samples. Due to the overlap of scale descriptors found common between the three depression-symptom inventories, we decided to only include the IDS-30 descriptor set in our finalized dataset. Our choice of the IDS-30 scale is justified by the findings of a previous study that revealed a high degree of correlation between IDS-30 and QIDS scale descriptors and expressed limitations of the HRSD-17 scale in including all possible depression-related symptoms [53]. Eighteen features were pre-selected from an initial large pool containing 50 features, which was done using recommendations from medical experts. Treatment variant was an important feature that reflects whether the patient had undergone standard rTMS or iTBS treatment. The feature space included both continuous, binary and categorical variables. A description of all included features can be found in Table 3.1. Across the entire data, 27 samples had missing data and were hence simply excluded from the dataset. This filtration of variables and removal of samples with missing entries lead to a new dataset $\{X_f, y\}_n$, where $f = 52$ variables/features and $n = 387$ patient samples. In this model, we define response and remission as the two binary output variables y to be modeled, where 0 represents non-response and non-remission, as reported at the end of each patient trial.

Inventory of Depressive Symptoms Descriptors *Each descriptor is rated on a scale of 0 - 3*
Falling Asleep
Sleep During the Night
Waking up Too Early
Sleeping Too Much
Feeling Sad
Feeling Irritable
Feeling Anxious
Response of Your Mood to Good or Desired Events
Mood in Relation to the Time of Day
The Quality of Your Mood
Appetite change
Weight change
Concentration
View of Myself
View of My Future
Thoughts of Death or Suicide
General Interest
Energy Level
Capacity for Pleasure or Enjoyment
Interest in Sex

Feeling slowed down
Feeling restless
Aches and pains
Other bodily symptoms
Panic/Phobic symptoms
Constipation/diarrhea
Interpersonal Sensitivity
Physical Energy
Energy Level
Capacity for Pleasure or Enjoyment
Interest in Sex
Feeling slowed down
Feeling restless

Table 3.2 IDS-30 scale description

Brief Symptom Inventory – Anxiety subscale (BSI-A) *Each descriptor is rated on a scale of 0 – 4* *(based on past 7 day experience)*
Nervousness or shakiness inside
Suddenly scared for no reason
Feeling fearful
Feeling tense or keyed up

Spells of terror or panic
Feeling so restless you couldn't sit still

Table 3.3 BSI-A scale description

3.5 Proposed Model Framework

3.5.1 Variable Selection

The 18 clinical and sociodemographic variables, the 6 BSI-A descriptors and the 28 IDS-30 descriptors were combined together to form a data set where each patient was represented by 52 independent variables. Variable reduction is an important pre-processing step in ML. It is leads to a decrease in the computational complexity of an algorithm as well as the development of a more refined variable set that can reduce noise and improve model performance. For our specific problem task, variable reduction can allow us to design a more practical ML model capturing only the most relevant characteristics from the dataset. Variable transformation is a type of variable reduction that works by completely transforming the original variables into a new set of variables. The aim of this method is to combine similar characteristics seen across variables into a smaller set of variables that explain most of the data set variance. Principal Component Analysis (PCA) is a popular variable transformation procedure that linearly projects a variable set into a set of 'principal components', which is a term given for the constructed linearly uncorrelated variables [54]. To investigate this, we show how PCA projection can be applied to our data set by constructing different components and visualizing how each component contributes to the amount of total data variance. Figure 3.2 demonstrates the fraction of the total variance explained by each component constructed using all 52 variables in our data set. Generally, components are combined until their variances sum up to around 80-90% of the total variance. However, as shown in Figure

3.2, the first and second components contribute only 12% and 6% of the total variance respectively, with most of the other components having less than 2% variance contribution. This means that PCA is not suited for our data set, in which the variables are highly uncorrelated. Moreover, in order to preserve variable interpretability, it is best to deploy a variable selection procedure, rather than f variable transformation. Several variable selection techniques exist, but we mainly focused on classifier-embedded variable selection, as it explores the relationship between the independent variables and the dependent output. It works by identifying strongly relevant variables that highly contribute to a classifiers performance. RF is classifier constructed with a large number of independent decision trees. In addition, RF has been widely used as a variable selector, and proven to outperform other embedded methods in terms of variable robustness [55]. For this reason, we have chosen to use RF as our algorithm of choice to perform variable selection. Nodes within a single decision tree represent a split condition for a specific variable. The split parameter is optimized using nodal Gini impurity [56]. For each variable during the classifier training, the Gini decrease in impurity is found, and this corresponds to how well a specific variable can help improve the classifier's accuracy. This impurity decrease of each variable is then averaged across all the decision trees and each variable receives an importance score, represented in the form of a fraction of the total variable importance. At this stage, no preprocessing or standardization of the data was needed as the RF is insusceptible to it. Two different dichotomous RF classifiers were used to identify the most important variables related to each of the two outcomes; response and remission. In order to optimize the performance of the RF, the model parameters were obtained using a grid search algorithm [57] which works by sweeping through a pre-defined set of parameters and testing the classifier using 4-fold cross validation until the best set of parameters is found. The parameters under consideration were the number of decision trees and the max tree

depth, which were set to 25 trees with a max depth of 5 levels. The classifier was trained using bootstrapping to reduce out of bag error estimate and validated using internal 4-fold cross validation. In 4-fold cross validation, the data is split into 75% training set constituting the 3 folds and 25% validation set for 4th fold, and this is repeated 4 times such that each data instance appears in the training set 3 times, and in the validation set once [58]. Variable importance was then extracted post classifier training and the top 15 most relevant variables/features for each of response and remission models were recorded. The aim of this step is to reduce the variable space in an attempt to promote practicality of the predictive tool and to make it easier for the physician to collect and predict treatment outcome based on a smaller set of variable. Variable selection lead to a smaller dataset $\{X_t, y\}_n$ where $t = 15$ variables/features and $n = 387$ patient samples. Additionally, neural networks work best when noise and redundancy is reduced from the input data. Therefore, the 15 top variables /features for each treatment outcome will then be scaled and loaded into a deep neural network model as described in a later section.

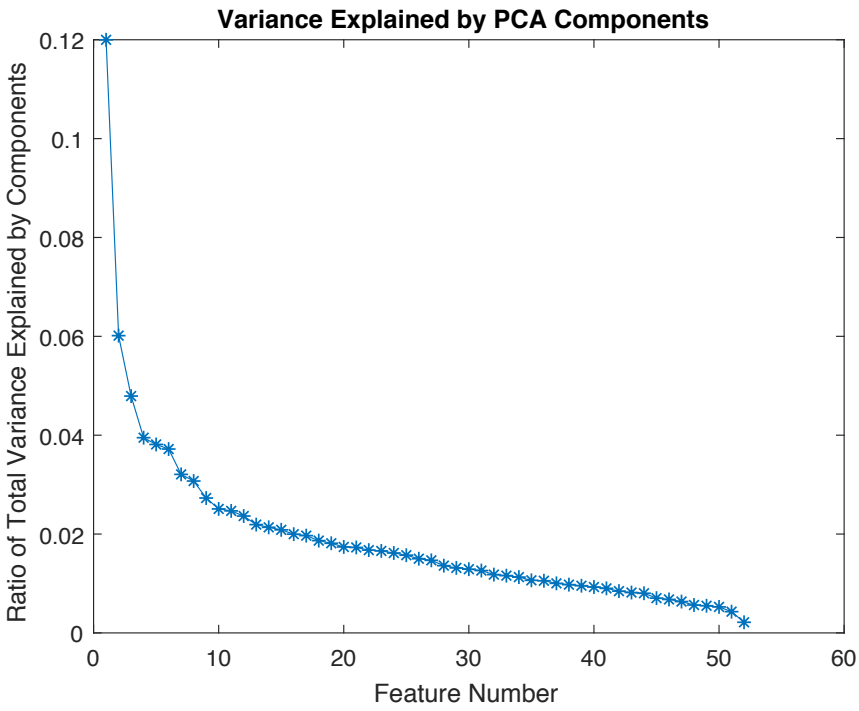


Figure 3.2 PCA components using rTMS dataset

3.5.2 Data Over-Sampling

Imbalance in a dataset refers to the case where class distribution is not uniform, meaning we have a majority class with more instances than those of a minority class. Class imbalance in a dataset can hinder a classifier's performance in terms of its sensitivity towards detecting the minority class. In our dataset, the ratio of response to non-response is 38:62, and remission to non-remission is 25:75. Imbalances in the dataset can be overcome in a number of ways. When the dataset is large in size, under-sampling is done where samples of the majority class is removed using either ensemble learning or cluster methods, keeping only the most informative samples until almost 1:1 class ratio is achieved. However when the dataset size is small, as in our situation, under-sampling can significantly reduce the ability of a dataset to construct an accurate classifier, and hence, an

over-sampling technique is favored. Over-sampling has been previously done by simply duplicating instances of the minority class with replacement, however studies [59] have shown that the prediction accuracy of minority class instances did not significantly increase using this technique. A new method known as the Synthetic Minority Over-sampling Technique (SMOTE) [60] oversamples the minority class by creating new synthetic instances using k-nearest neighbors algorithm. The designers of this over-sampling method showed that it results in classifiers with higher generalization of the minority class due to bigger and less specific decision regions [60]. Figure 3.3 shows their comparison of SMOTE with standard oversampling with replacement, proving how accuracy of the minority class detection improves with different degrees of oversampling.

Using SMOTE, we calculate the degree of over-sampling needed based on our class imbalance ratio which is 63% for the response dataset and 200% for the remission dataset. Next, a subset of the minority class is chosen at random, and for each sample within the subset, a variable vector is found. For variables with continuous values, this vector is calculated using the Euclidean distance between the minority sample under consideration and its nearest neighbor, which is then scaled using a randomized factor between 0 and 1 to create a new sample lying between the original sample and its nearest neighbor. In terms of variables consisting of a number of categories, a majority vote among the k-nearest neighbors found for the considered observation will determine the new variable category. Binary variables should remain unchanged. Over-sampling lead to a larger dataset $\{X_t, y\}_o$ where $t = 15$ variables/features and $n = 474$ patient samples for response dataset and $n = 584$ patient samples for remission dataset. To prevent any information leakage between the testing and training sets, over-sampling is done independently for each of the sets.

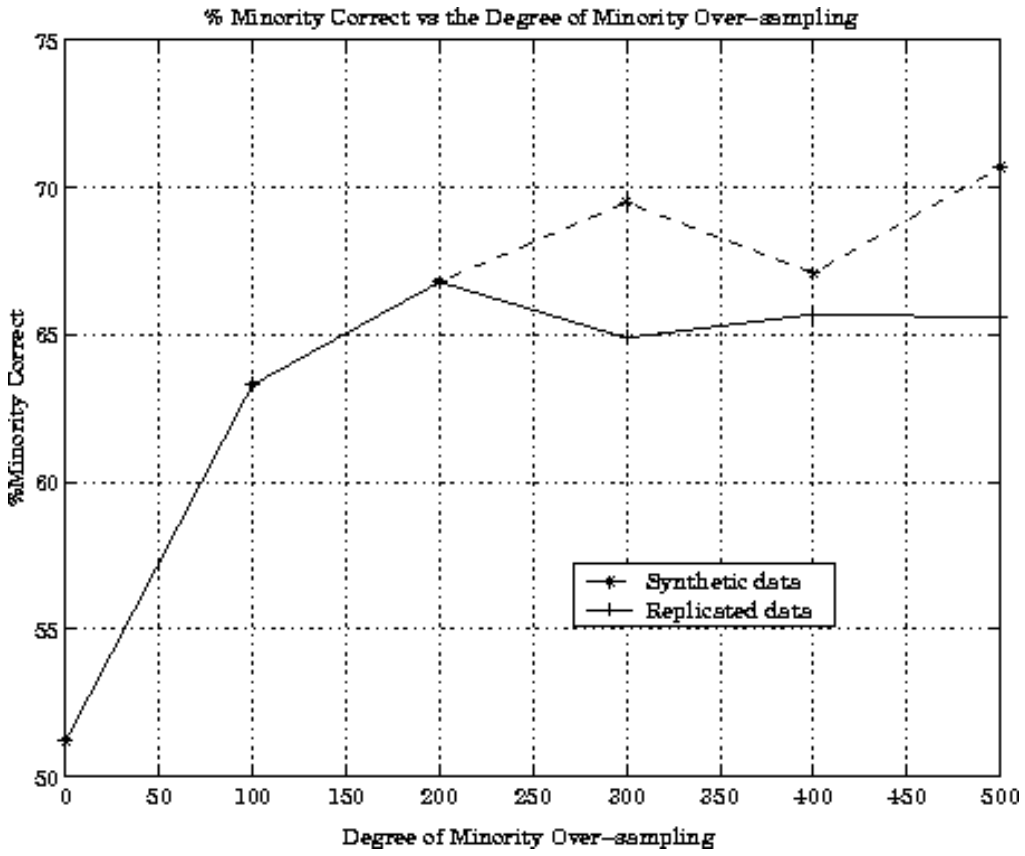


Figure 3.3 SMOTE performance compared with standard over-sampling [60]

3.5.3 Classification Models

Selecting the right classification algorithm is a crucial step in designing a good ML model. There does not exist a single ML model that works well for all kinds of classification tasks, as model performance is highly dependent on the nature of the data and the complexity of the task. An analysis of the data size, variable types and interrelationships can give us a better idea about a more narrow group of ML models to target. Our data set has a mixture of categorical and continuous features, and for that reason, we focus on evaluating three different ML algorithms known for their ability to handle mixed features well; ANN, RF and LR. ANN has the capability

of approximating complex functions or mappings between input variables and output response to a high degree of accuracy. RF is an ensemble classifier known for its high robustness and interpretable structure that resembles the process of medical reasoning and decision making. Furthermore, LR is one of the most widely used algorithms in the medical field, mainly due to its easy interpretability and mathematical mapping between the independent variables and the output response. As discussed, each of these ML algorithms have their own perks, and performance evaluations will help us decide which of them to deploy.

Machine Learning Classifier	Preliminary Results
Support Vector Machine	Accuracy: 63.7% Sensitivity: 30.3% Specificity: 85.2%
Linear Discriminant Analysis	Accuracy: 65.4% Sensitivity: 68.4% Specificity 62.5%
K Nearest Neighbor	Accuracy: 50.6% Sensitivity: 42.1% Specificity 56.1%
Gaussian Naïve Bayes	Accuracy: 65.6% Sensitivity: 70.5% Specificity 60.7%
Decision Tree	Accuracy: 59.5% Sensitivity: 61.6% Specificity 59.5%

Table 3.4 Preliminary study results

Medical data is highly complex and usually consists of several data types, so determining the best classifiers can be a challenging task. Although most ML classifiers have certain underlying assumptions, these assumptions may not be necessarily met when dealing with real data. Therefore, to test the assumptions we made regarding our choice of classifiers and better understand our specific data, we carried out a preliminarily analysis of some of the most commonly deployed ML algorithms. Our wide selection of classifiers included Support Vector Machine (SVM) [61], Linear Discriminant Analysis (LDA) [62], K Nearest Neighbor (KNN) [63], Gaussian

Naïve Bayes (GNB) [64] and Decision Tree (DT) [65]. The results of this preliminary analysis is shown in Table 3.4 above. All the results were obtained using 4-fold cross validation. As can be seen in the results, all the classifiers have low performance, measured using the accuracy metric (accuracy < 66%). As for SVM, there is a huge bias towards predicting the response instances (class label 1), as reflected by the very high specificity measure (85.2%) and very low sensitivity measure (30.3%). Although SVM is an excellent classifier for high sparsity problems, it does not perform well in smaller data sets with mixed data types. The above results confirms our initial choices of selecting LR, RF and ANN to be used as classifiers in this problem. This will be further discussed and proved in Chapter 4.

We developed an ANN with an input layer, 2 hidden layers and an output layer for each treatment outcome. The input layer consisted of 15 nodes, one node per variable/feature, and the output layer consisted of a single node that can either output a 1 or a 0. The dimensionality and number of hidden layers were found through a grid search, along with other parameters which included 1) The activation function used in the hidden layer, 2) The loss function solver for weight optimization, which represents the L2 penalty (regularization term) parameter and 3) The number of maximum iterations. Selected parameters can be seen in Table 3.5 below. This neural network has a feed-forward architecture which means that the information moves along the network layers from the input layer, across the hidden layer and into the output layer, without going through any loops. The function which was used to activate the nodes in the hidden layer is the tanh function which is represented as:

$$tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

Where z is the weighted sum of the inputs and biases of all inputs to the node. Tanh is generally more suitable than the sigmoid function as it has larger derivatives which allows for a faster minimization of the loss function. In addition, the tanh function has a range of [-1,1] which maps the outputs to center around 0 rather than 0.5 as in the case of sigmoid, which allows for easier learning when the information is passed to the next layer. Backpropagation is used to adjust the weights in the network layers through a minimization of the cost function. The cost function used in our model is the logarithmic loss, also known as the cross-entropy cost function [66]. Cross entropy loss for binary classification can be shown as:

$$Cross\ Entropy\ Loss = -(ylog(p) + (1-y)log(1-p))$$

Where y is the expected class output and p is the predicted probability. The Adam optimizer [67] is an extension of the stochastic gradient descent and is used in the optimization of the cross entropy loss as it updates the network weights during the training phase. Two main measures were taken to avoid overfitting. Firstly, the dataset was validated using 4-fold cross-validation where the samples were split to 75% training and 25% testing, iterated over 4 times. Additionally, an early stopping mechanism was in place such that 10% of the training samples were used as validation, and a validation score was estimated once no improvement was seen within a tolerance level of $1x10^{-6}$ for 10 iterations, causing the training to immediately stop. Once training was complete, the optimal neural network was evaluated using a testing set to analyze its ability to accurately predict the outcome to rTMS treatment.

Parameter	Response Neural Network	Remission Neural Network
Number of hidden layers	2	2
Nodes per hidden layer (i,j)	(25, 12)	(26, 14)
Activation function	tanh	tanh
Loss function solver	Adam	Adam
Alpha α	0.1	0.1
Maximum iteration	3000	5000
Learning rate	0.015	0.015

Table 3.5 Hyperparameter values for the deep neural networks

The LR model was built based on the below equation where $x1, x2, \dots, x15 \in R_n$ represent the top 15 variables, $y_i = \{0,1\}$ represent the output of the model and w_n represent the coefficients/weights of the logistic function

$$y = f(x) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_{15}x_{15}$$

$$where\ f(x) = log\left(\frac{p}{1-p}\right) = e^{w_0 + w_1x_1 + w_2x_2 + \cdots + w_{15}x_{15}}$$

L1 regularization using the liblinear optimizer [68] with penalty parameter $C = 0.8$ was used for our LR models, as determined using a grid search with 4-fold cross validation. L1 regularization can be described as minimizing the below function with weight vector $w \in R_n$:

$$minimize \quad \|w\|_1 + C \sum_{i=1}^{l} (\log (1 + e^{-y_i w^T x_i})$$

$$where \quad \|w\|_1 = \sum_{i=1}^{n} |w_i|$$

Finally, the RF classifiers consisted of 50 decision trees, each having a maximum depth of 5, as found by a parameter grid search. Bootstrapping was enabled, which means that each tree was built using a sub-sample of the original data samples, drawn with replacement. In our RF classifier, the quality of the binary node splits were evaluated based on the Gini index given by:

$$Gini\ Index = 1 - (P_0^2 + P_1^2)$$

Where P_0 and P_1 represent the proportion of samples having class labels 0 and 1 respectively [69]. The nodal impurity for a specific split is minimized (Gini index = 0) when a pure state is achieved, having all samples at the split belonging to a single class:

$$Gini\ Index\ Min = 1 - (P_0^2 + P_1^2) = 1 - (1^2 + 0^2) = 1 - (0^2 + 1^2) = 0$$

Meanwhile, the nodal impurity for a specific split is unfavorably maximized (Gini index = 0.5) when a mixed state is achieved, where the samples at the split are equally distributed:

$$Gini\ Index\ Max = 1 - (P_0^2 + P_1^2) = 1 - (0.5^2 + 0.5^2) = 0.5$$

The motivation behind assessing nodal impurity is to achieve child nodes that have high nodal purity in terms of the output variable, by minimizing the Gini index. The tree splits are determined at each node using a greedy search that tests all possible splits to select the one resulting in the lowest impurity (Gini Index).

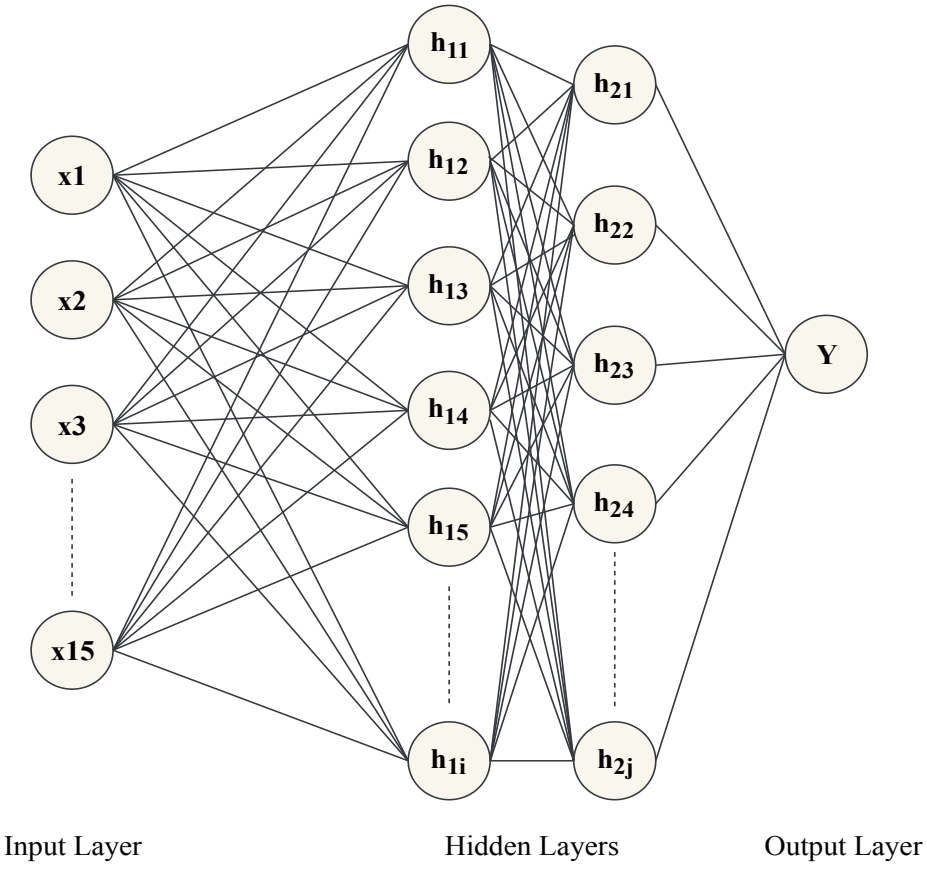


Figure 3.4 Architecture of DANN for treatment outcome prediction

3.6 Model Evaluation

We compared the performance of the deep neural network model with that of logistic regression and of random forests using the top 15 selected variables/features. Each model was evaluated using several metrics; accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) [70]. Sensitivity measures the fraction of all the positive-labeled instances that are correctly identified as positive. Similarly, specificity measures the fraction of all the negative-labeled instances that are correctly identified as negative. PPV is the probability that an instance identified as positive, truly is positive and similarly NPV is the probability that an instance identified as negative, truly is negative. All these metric values were obtained as an average of all validation folds using 4-fold cross-validation. Receiver Operating Characteristic (ROC) curve is an important tool used to evaluate binary classifications. It provides a visualization of every possible sensitivity (also called true positive rate) and 1-sensitivity (also called false positive rate) pair for different thresholds [71]. Additionally, the Area Under the ROC Curve (AUC) reflects how well a classifier is able to distinguish between two different classes [72]. An ROC analysis was conducted for each of DANN, LR and RF classifiers, with two variants per classifier to predict each of remission and response. Finally, all models were tested for accuracy significance using p-values calculated with permutation tests and 4-fold cross validation.

Chapter 4: Results & Discussion

4.1 Results

4.1.1 Selected Variables

All 52 variables combining the IDS-30 scale descriptors, BSI-A scale descriptors and other clinical and sociodemographic data were used to construct two RF variable selectors, for response as well as remission outcomes. The top predictors of response outcome, represented as a fraction of the total Gini importance, can be seen in Figure 4.1 below. The most important predictor of treatment response is the IDS-30 descriptor for future outlook with importance weight 0.0571 (STD 0.0166), followed by the IDS-30 descriptor of feeling sad 0.0436 (0.0143), BSI-A descriptor of feeling fear 0.0423 (0.0174) and the sociodemographic variables indicating age 0.0368 (0.0091) and financial status 0.0343 (0.0086). Financial status is a categorical variable consisting of 7 subcategories. Using one-hot encoding, we further expanded the financial status variable to analyze the individual categorical importance and found that employed status 0.0393 (0.01918) had the highest contribution to prediction accuracy followed by family support 0.0075 (0.0089), receiving insurance disability 0.0073 (0.0051), unemployed 0.0044 (0.0033), supported by the Ontario disability support program 0.0030 (0.0025), supported by spouse 0.0022 (0.0021) or supported by the Ontario works program 0.0013 (0.0020). Similarly for remission outcome, Figure 4.2 shows the top outcome predictors. IDS-30 descriptors for suicidality and future outlook had the highest importance scores with 0.0892 (0.0142) and 0.0873 (0.0268) respectively. IDS-30 descriptor of anxiety 0.0659 (0.0153) and sadness 0.0480 (0.0222) ranked 3rd and 4th most important predictors of remission followed by the sociodemographic feature representing financial status 0.0353 (0.0075).

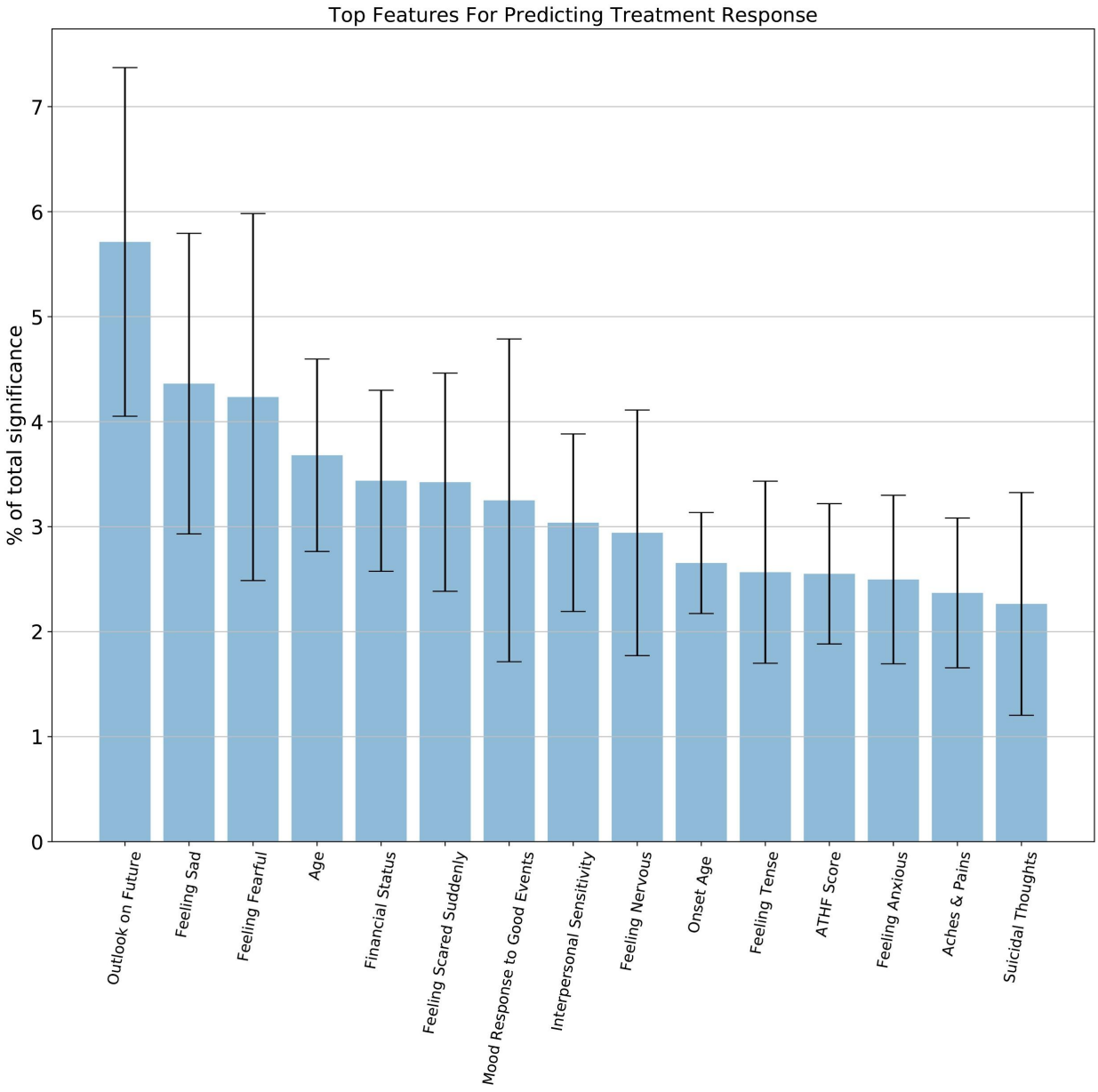


Figure 4.1 Top 15 variables/features associated with response prediction

The most insignificant variables were common for both response and remission models within the pool of 52 variables. Binary variables indicating different types of pharmacotherapy a patient might have undergone during treatment had almost no importance on predicting response or remission with feature scores ranging between (0.0003-0.0056) for response and (0.0002-0.0041)

for remission. Additionally, the undergone treatment variant (rTMS or iTBS) showed low importance for both response 0.0080 (0.0049) and remission 0.0056 (0.0035).

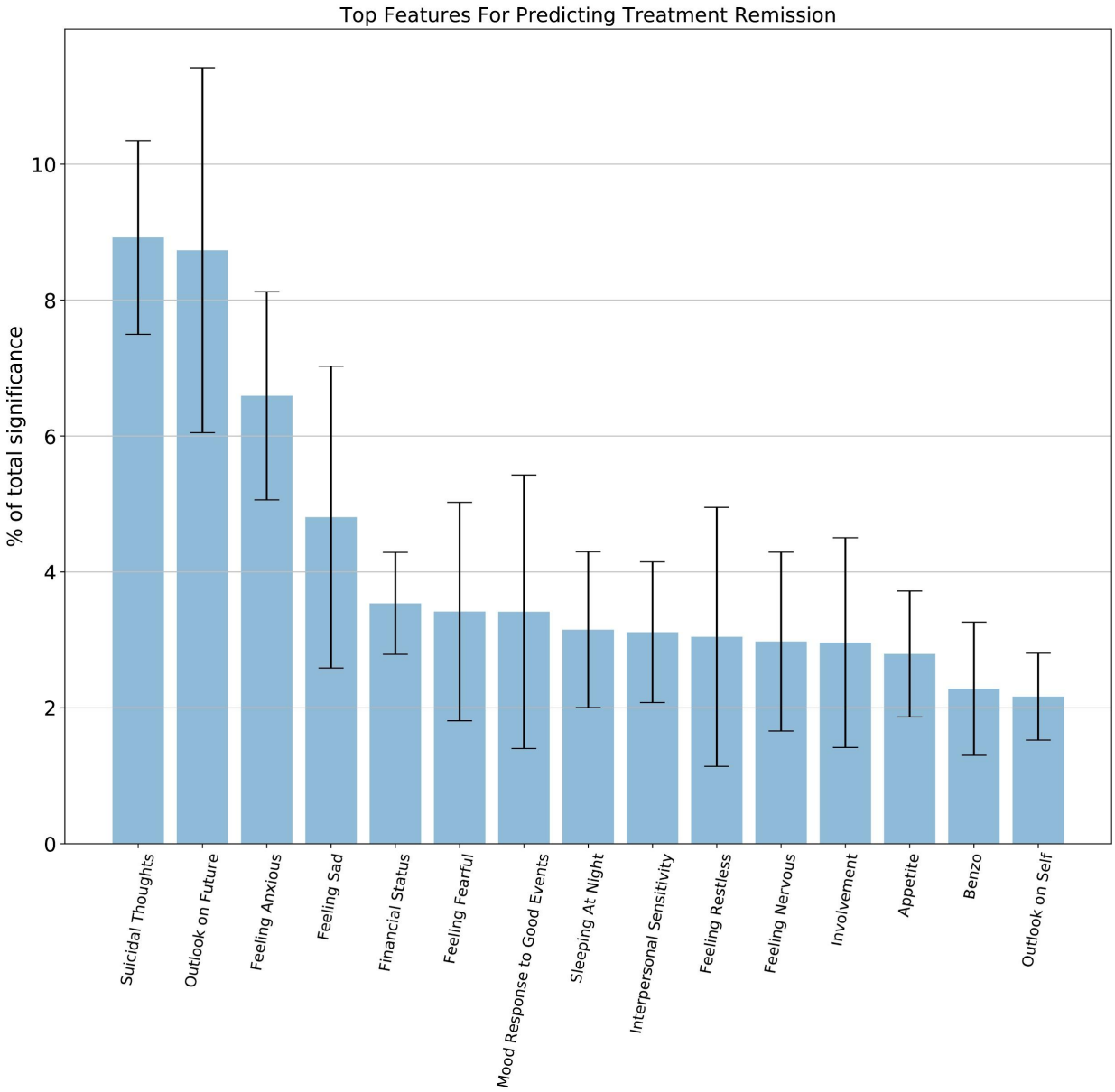


Figure 4.2 Top 15 variables/features for remission prediction

Furthermore, the RF variable selection models showed good performance when compared with chance performance. The response prediction model had an accuracy of 72.8 % (STD 2.5), sensitivity of 67.1% (5.1), specificity of 73.4 % (3.6), PPV of 73.4 % (3.6) and NPV of 69.3% (3.4). Remission prediction yielded better results with an accuracy of 86.1 % (STD 1.0), sensitivity of 84.9% (4.5), specificity of 83.9 % (3.0), PPV of 85.0 % (2.0) and NPV of 84.6% (3.3). Results of the ROC analysis can be seen in Figure 4.3 below. The AUC values for response and remission prediction were calculated as 0.76 and 0.92 respectively, where an ideal predictor would have an AUC of 1.0 and a random predictor would have an AUC of 0.5.

4.1.2 Performance Metrics

The set of the 15 top variables were extracted from both response and remission RF feature selection models and each inserted into a deep neural network with 2 hidden layers. For comparison, LR and RF models were also developed using each of the top variable/feature sets. Table 4.1 provides a comparability analysis of all three model performances for each of response and remission outcome predictions. All evaluation metrics were obtained using an average of 4-fold cross validation results. All models were tested for accuracy significance and had $p < 0.01$. Deep neural networks yielded better accuracy in both response and remission prediction with 73.8% and 84.4% respectively. RF models came in next with accuracies of 71.5% and 82.7% for response and remission prediction respectively. LR, RF and DANN models all yielded significantly superior results when compared with random chance, standardly taken at an accuracy of 50%. It is also evident that remission prediction models had overall better performance than that of response. As demonstrated in Table 4.1, remission had better class separability when compared with response. Ideally, the probability curve of the positive class should peak around a probability

score of 1.0, and similarly 0.0 for the negative class. However, in the case of response prediction, a majority of the instances belonging to the responders class are misclassified as non-responders, explaining the lower performance results in comparison with that of the remission model.

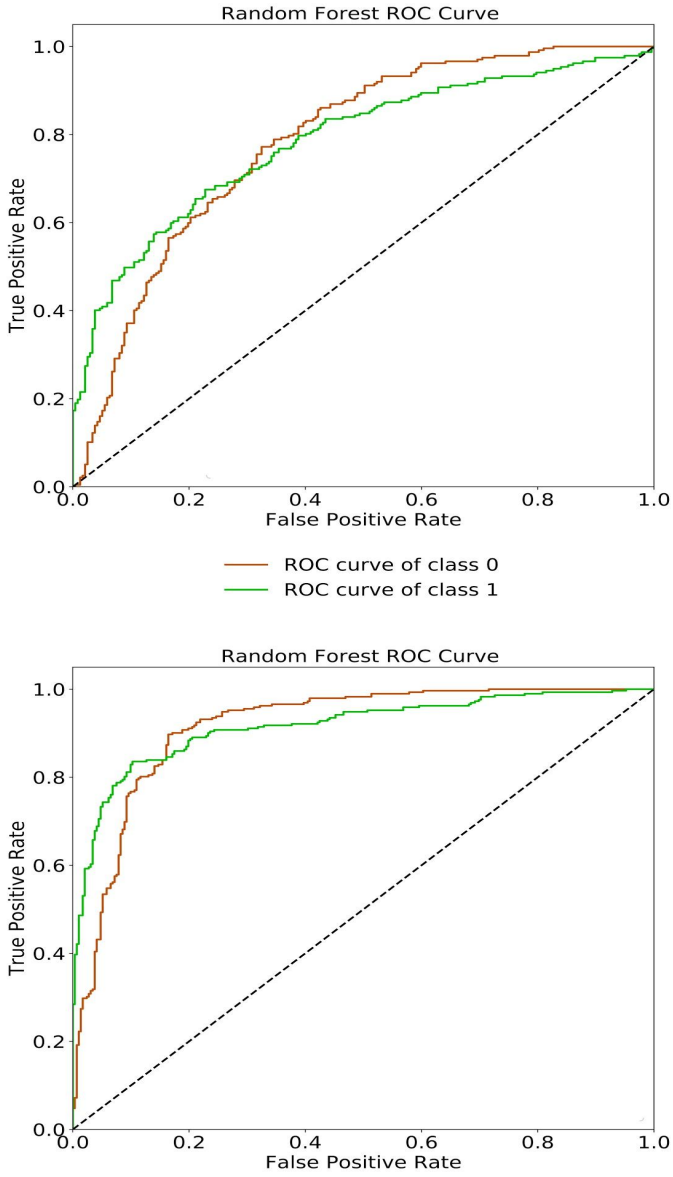


Figure 4.3 ROC curve for RF feature selection models

(upper curve - response model, lower curve - remission model)

4-fold Cross Validation						
	Deep Neural Network		**Logistic regression**		**Random forest**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Response Models						
Accuracy	73.8%	4.5	68.1%	2.0	71.5%	2.6
Sensitivity	70.0%	4.0	67.9%	6.1	69.2%	4.0
Specificity	77.6%	7.9	68.4%	3.8	72.6%	3.7
PPV	76.2%	6.4	68.3%	1.6	73.2%	1.7
NPV	72.1%	3.3	68.3%	3.0	73.4%	5.0
Remission Models						
Accuracy	84.4%	2.4	74.8%	1.6	82.7%	1.9
Sensitivity	78.8%	2.5	79.1%	1.1	82.4%	3.7
Specificity	90.1%	5.0	70.5%	2.5	86.6%	1.5
PPV	89.1%	4.8	72.9%	1.8	81.5%	3.9
NPV	80.9%	1.3	77.1%	1.4	88.0%	2.4

Table 4.1 Performance measures in predicting treatment outcome

4.1.3 ROC Analysis

ROC curves represents a visual trade-off between the benefits (true positives) and costs (false positives) of a classifier. Figure 4.4, Figure 4.6 and Figure 4.8 shows ROC curves for the DANN, RF and LR models respectively, where the top curve represents response prediction and the lower one represents remission prediction. AUC of an ROC curve is another way to measure a classifier's accuracy, with a value of 1 representing a perfect classifier and a value of 0.5 representing a random classifier. The average AUC for the DANN model was 0.78 for response and 0.91 for remission, for the RF model it was 0.76 for response and 0.90 for remission and for the LR model it was 0.72 for response and 0.83 for remission. To visualize how good each of the models were at discriminating between response/non-response and remission/non-remission, we plotted probability distribution curves for each class in Figure 4.5, Figure 4.7 and Figure 4.9 for each of DANN, RF and LR models respectively. Ideally in these class distribution plots, the positive class plot should be on the right side, peaking around a score of 1, while the negative class should be all the way towards the left, peaking around a score of 0. The overlap between the positive and negative class distributions should be minimal in the case of a good classifier. This is evident when comparing class distribution plots for the DANN models with that of the LR models. The LR models had the lowest accuracy when compared with RF and DANN, and looking at Figure 4.9, we can see this through the large overlap region between the two class distributions plots.

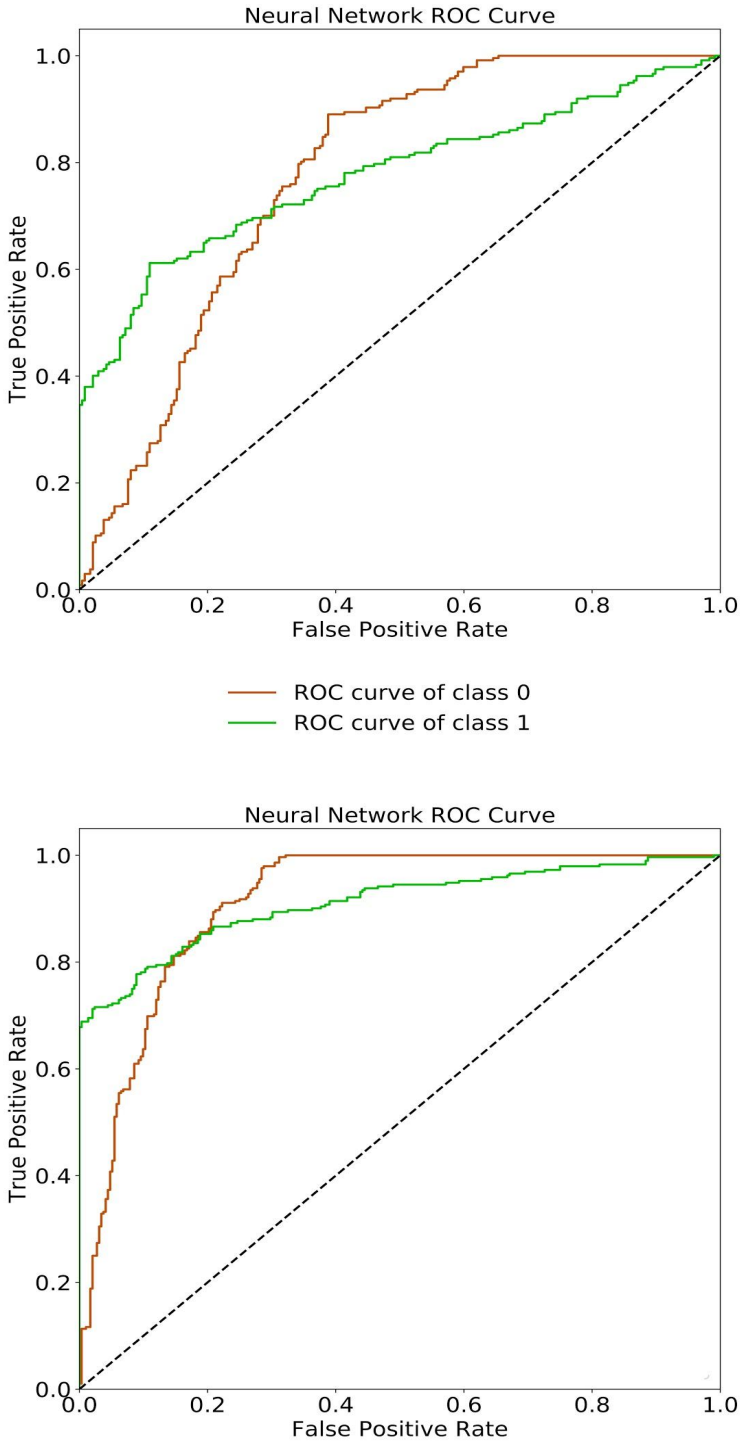
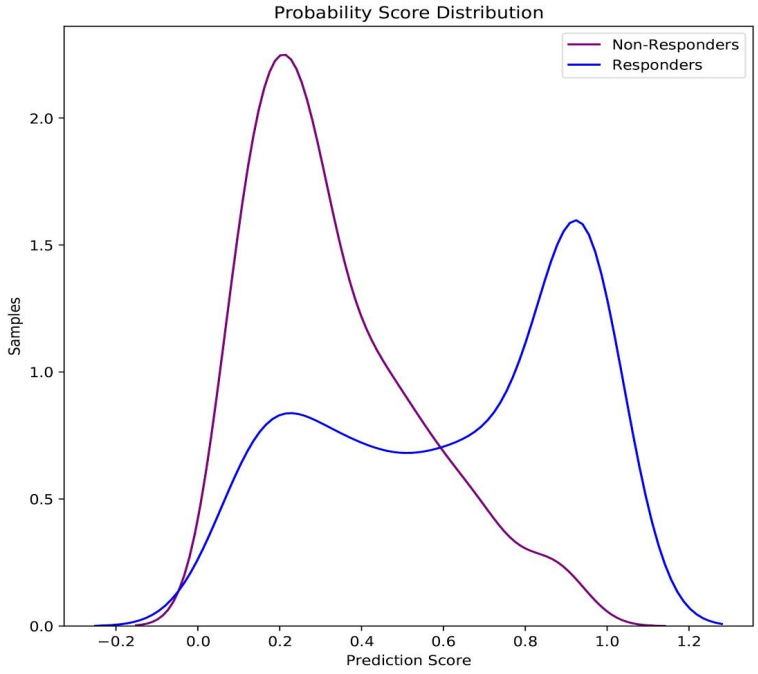


Figure 4.4 ROC curve for DANN models

(upper curve - response model, lower curve - remission model)



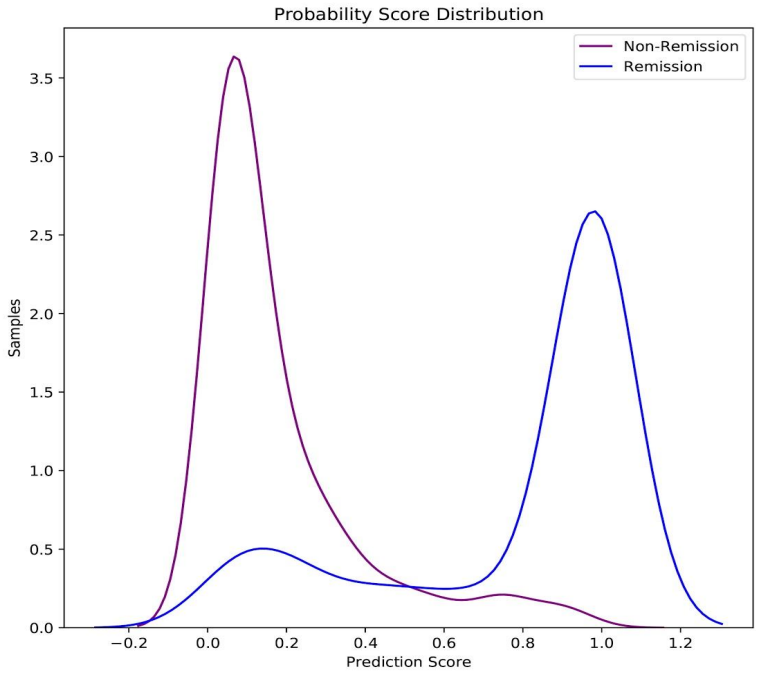
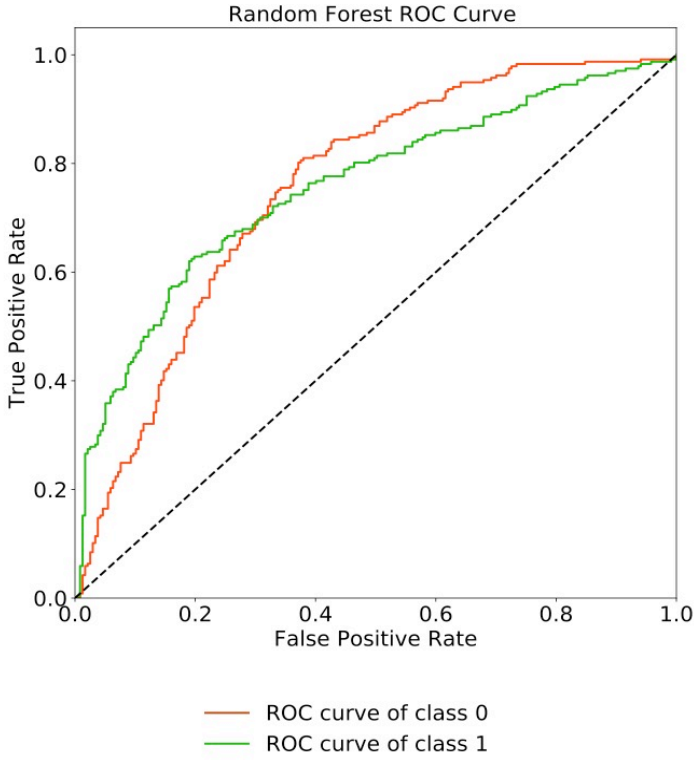


Figure 4.5 Probability score distributions for DANN models

(upper curve - response model, lower curve - remission model)



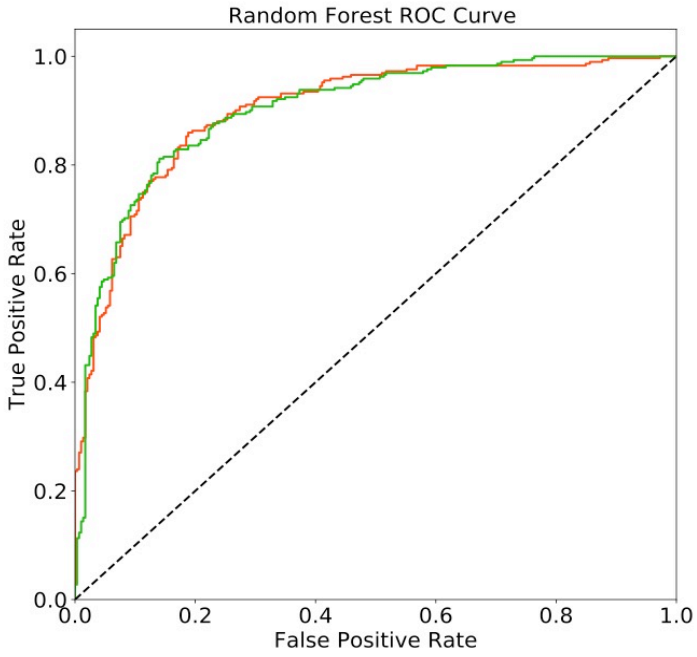
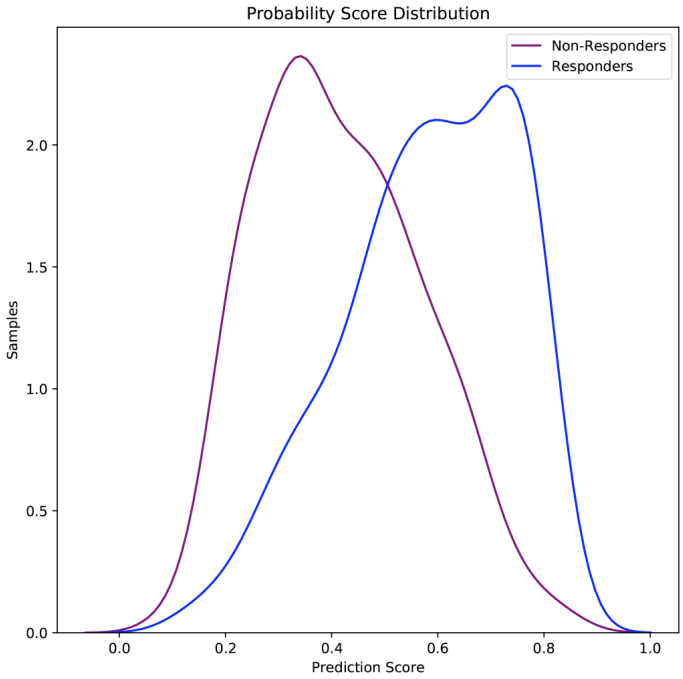


Figure 4.6 ROC curves for RF models

(upper curve - response model, lower curve - remission model)



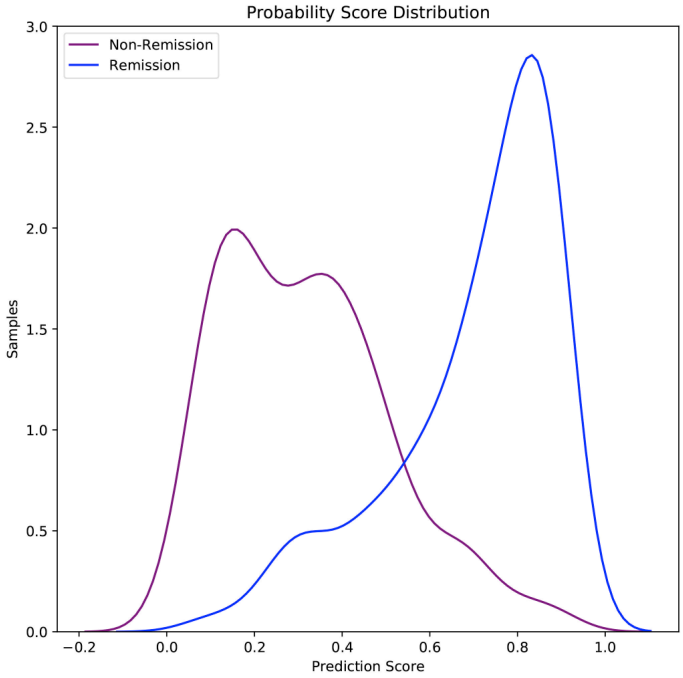


Figure 4.7 Probability score distributions for RF models

(upper curve - response model, lower curve - remission model)

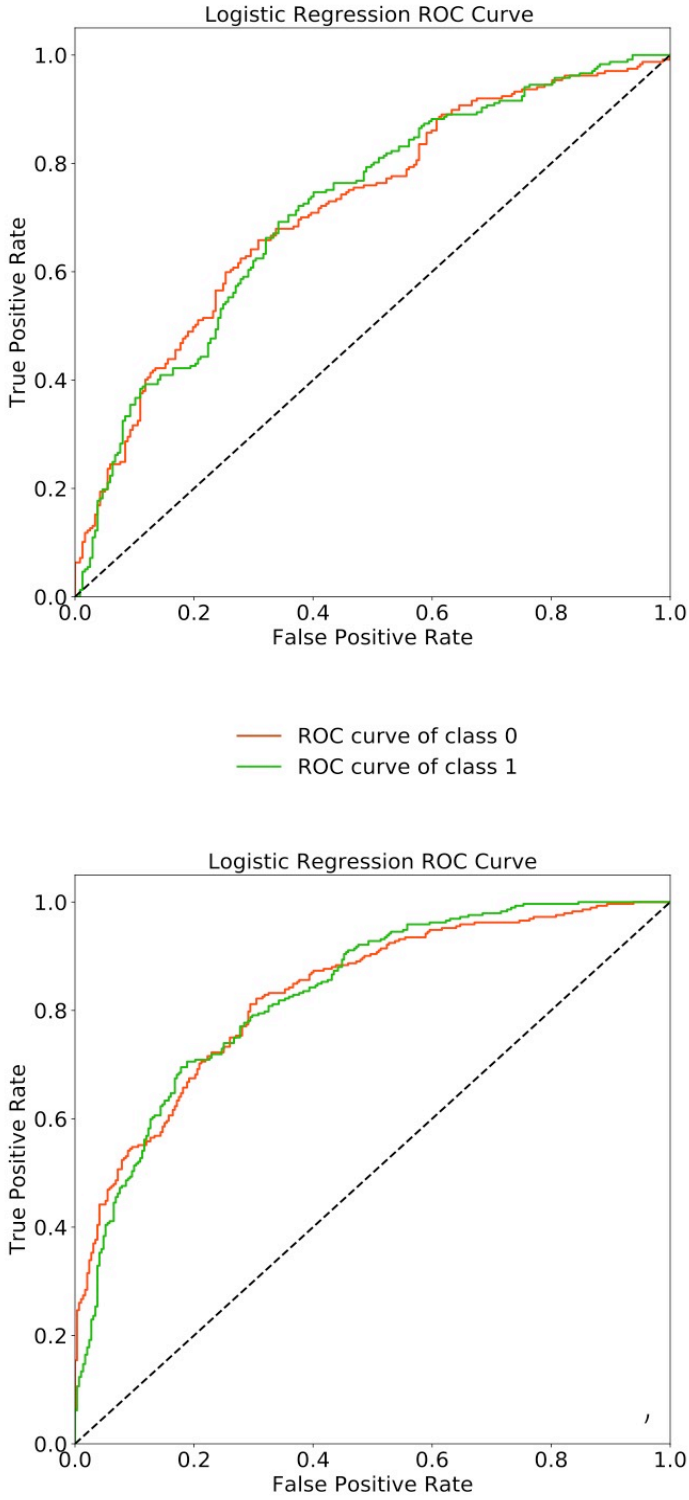
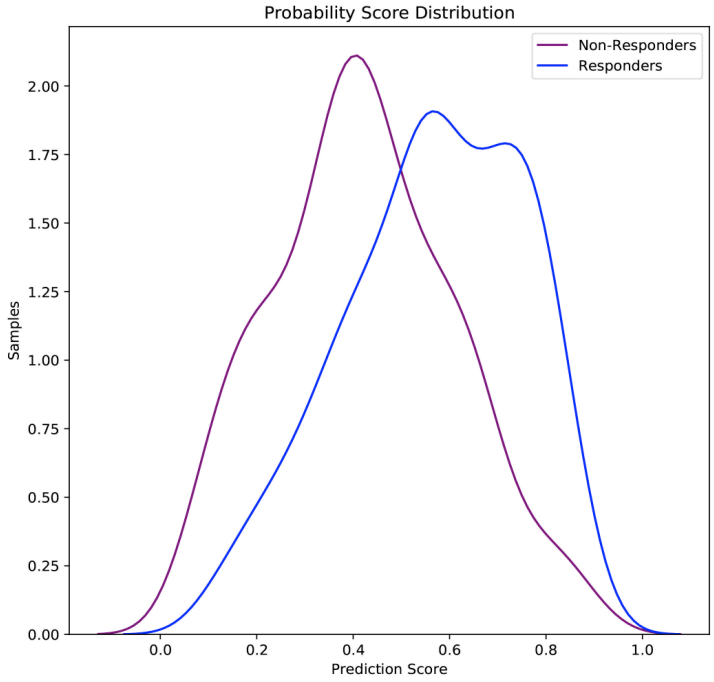


Figure 4.8 ROC curves for LR models

(upper curve - response model, lower curve - remission model)



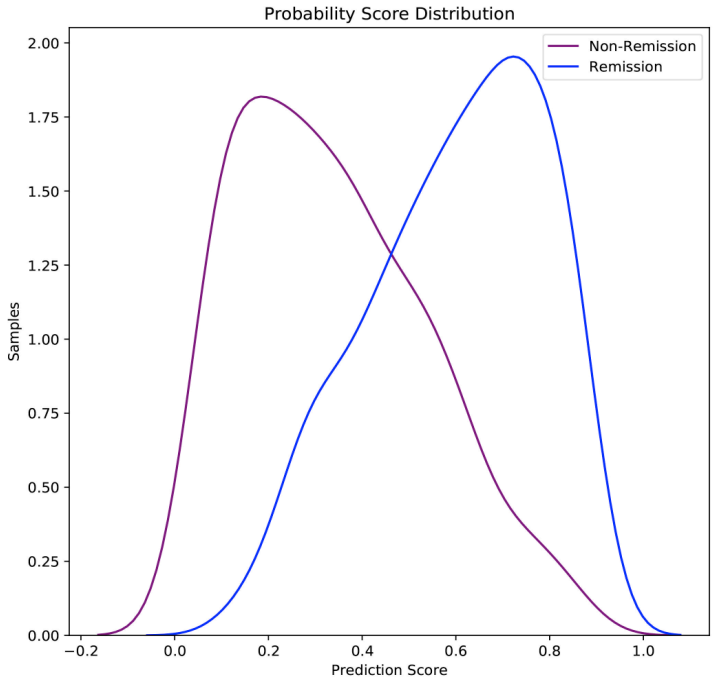


Figure 4.9 Probability score distributions for LR models

(upper curve - response model, lower curve - remission model)

4.2 Discussion

The result shown in this thesis highlight the possible advantage that clinical assistive tools could provide to aid physicians in deciding whether a MDD patient is likely to respond to rTMS treatment. Our system is based on a set of 15 clinical and sociodemographic variables/features at baseline, which can be collected from a patient within a couple of mins. The results demonstrated in this thesis are the first of its kind among brain stimulation studies, which is based on machine learning algorithms utilizing only clinical and sociodemographic data to predict the rTMS treatment outcome in MDD patients. Using a DANN, the patient's showing responsive symptoms can be accurately distinguished with an accuracy of 73.8% from those not likely to respond to treatment. Remission prediction had superior results with an accuracy of 84.4% using the DANN model. This difference in performance is notably backed by the nature of patient score distribution. Figure 4.10 shows how the % change in IDS-30 score between baseline and post-treatment phases varies between patients. Three main patient groups are visible in the distribution plot, mainly the highly responders on the far right, the highly non-responders on the far left and those wavering around the 50% score change in the region encompassed by the two green lines. The response is defined as a 50% or more drop in the IDS-30 score, and therefore a threshold right at the 50% mark separates responders from non-responders. This causes a lot of difficulty in identifying patients hovering right close to the 50% mark. In the case of remission prediction however, the threshold is shifted towards the right, close to 79.2% which is the average % score change calculated for patients with remission. With this new threshold, the first 2 groups representing non-responders and those having response without remission is combined into a single group, reducing the complexity of the prediction task.

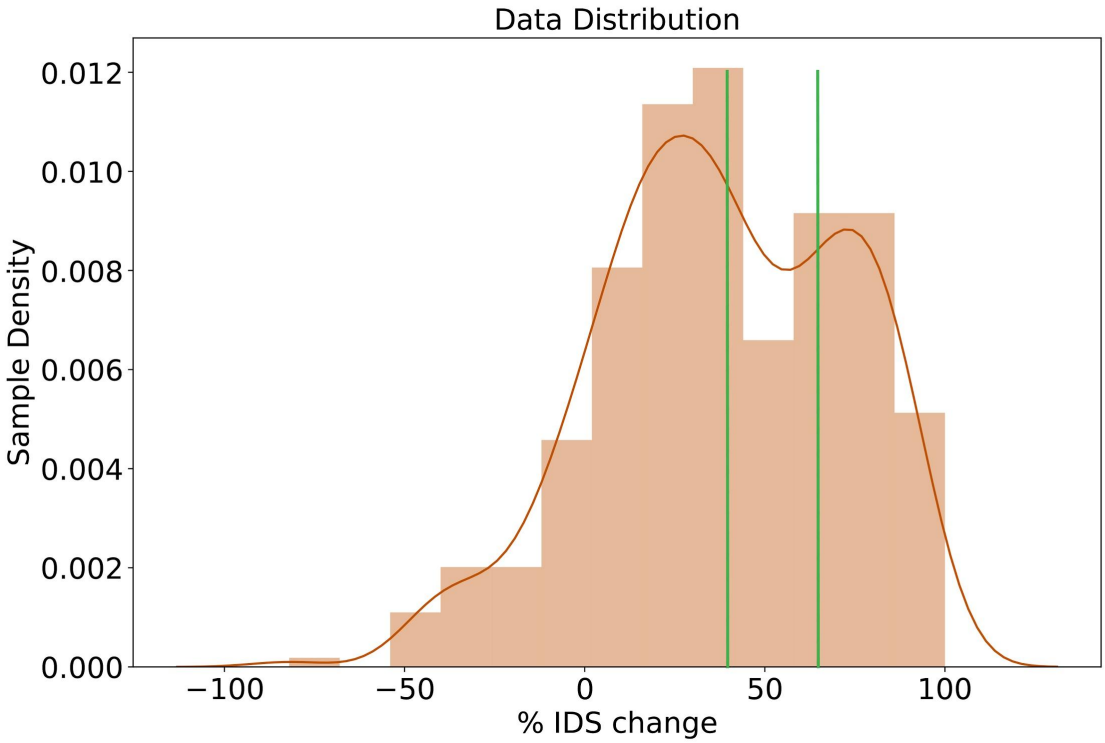


Figure 4.10 Sample distribution curve of our dataset

The DANN demonstrates superior performance for both prediction tasks in comparison with RF and LR. This can be attributed to DANN's ability to infer complex relationships between the predictors and the output variables, that may not be feasible using LR. Additionally, DANN provides high flexibility with a large set of hyperparameters, different activation functions and optimizers that can be modeled depending on the size of the dataset and the nature of the task.

The main downside of using a DANN is its concept of a 'black box' with vague understanding of the underlying variable relationships as well as its high computational cost. RF is known for its robustness to overfitting, application ease with only a small number of parameters and its embedded feature selection ability, but lacks feature interpretability. Therefore, given only

the small difference in the performances of RF and DANN, one should identify whether interpretability is an important factor, in which RF would be preferred over DANN.

One of the main debates surrounding the applicability of machine learning based models in the medical field is their generalization capacity. Usually smaller datasets have limitations when complex models are applied to them. The bias-variance tradeoff means that the higher accuracy achieved using a given dataset may lead to poor generalization when given unseen data. Dealing with model overfitting is a common issue in machine learning models. We took several measures to reduce model overfitting which included using cross-validation technique to train and test our model. This is sufficient for internal validation, however a more rigorous approach is necessary for clinical approval. External validation requires the use of an independent dataset of similar structure but collected from different sources and tested with the model under investigation, showing adequate results. This wasn't possible in this work due to the unavailability of publicly available datasets that have resembling characteristic to our model data, which is mainly due to the restrictions and privacy concerns put within the medical domain.

Analysis of the most important predictors of treatment outcome revealed interesting findings. There is a large overlap between the top predictors for response and for those of remission. IDS-30 descriptor for future outlook was the single best predictor of treatment response and the second best predictor of remission. Optimism about the future was found to have a strong correlation with depression [73]. Suicidality, which is also portrayed using the IDS-30 scale was found to be the top predictor of treatment remission among MDD patients. This concurs with results found by the group for the study of resistant depression [74], which shows suicidality being ranked the third descriptor to be affiliated with treatment resistant depression, preceded by anxiety

comorbidity which was the third most important predictor of treatment remission in our results. Other common predictors found in both response and remission models are sadness, mood response to good events and interpersonal sensitivity descriptors from IDS-30, the fear descriptor from BSI-A and financial status. A study analyzing outcome predictors of citalopram antidepressant showed strong correlation between higher income and higher remission rate [75], which overlaps to a degree with our results showing employment as a strong predictor of both response and remission. A study on the same data used in this thesis showed the non-inferiority of iTBS to standard 10 Hz rTMS treatment [31], which is affirmed with our results having low importance score for the predictor which represents treatment condition.

Chapter 5: Conclusion & Future Work

5.1 Summary

MDD is a widely prevalent disorder that affects thousands of people every year. Current treatments in place for depression vary based on patient severity and treatment history. Deep brain stimulation therapy like rTMS has been reserved for patients who are deemed unresponsive to adequate dosage of an antidepressant trial. Due to its lengthy procedures and large costs, there is a great benefit to identifying patients who are unlikely to show responsiveness to rTMS treatment. In this thesis, we introduce a novel framework that allows us to first detect the most important variables/features related to treatment outcome and then train and deploy a machine learning based model on the extracted features for two different classification tasks. An in-depth analysis of the most important variables extracted using the RF algorithm was followed by a discussion providing supportive arguments from other related studies. Classification tasks consisted of identifying patients likely to be responders to rTMS treatment and identifying patients likely to achieve remission from MDD. To approach this, we designed and trained a DANN with backpropagation and optimized it using a grid search of the hyperparameters to maximize performance. Evaluation consisted of cross-validation metrics, ROC analysis and comparison with the RF and LR models. Our deep neural network model achieved an accuracy of 73.8% in distinguishing responders form non-responders and 84.4% in identifying remission candidates. Meanwhile, the RF and LR classifiers had lower performances with accuracies of 71.5% (response), 82.7% (remission) and 68.1% (response), 74.8% (remission) respectively. We show how a selective model with only a small number of variables and limited preprocessing can have high performance with accurate predictions. We find that our model has significantly higher performance compared with pure chance, which is standardly the basis of how clinicians propose rTMS treatments to patients.

Our findings support the introduction of new clinical assistive tools to aid physicians with decision making in terms of treatment plans. It is fairly difficult for a physician to make an informed judgment about a patient by just looking at a small number of clinical and sociodemographic characteristics, which is where machine learning can be used to instantly provide case-by-case predictions. There is a lot of exciting opportunities for clinical adoption, however this is pending external validation.

5.2 Future Work

This thesis can be extended in the future to accommodate several possibilities:

5.2.1 Response & Remission as Continuous Scales

This work can be further improved by altering the task to perform regression in place of the existing binary classification. Regression analysis would allow us to categorize patients according to continuous response and remission scales. This additional information will remove the ambiguity of a standard binary classification, where a patient likely to achieve a 49% score drop would be classified as a non-responder, and discouraged from undergoing treatment, which would otherwise provide significant help. Similarly, this can also be achieved through a multi-classification adjustment, where instead of 2 binary classes, we relabel the data with 10 classes for example, based on scores representing least responsive to most responsive patients. Physicians will hence better explain the outcome expectations and allow patients to make an informed decision.

5.2.2 Time-Series Classification

Data collection at several time points between treatment initiation and termination would allow for a time series classification using Long Short Term Memory (LSTM) [76] units in a Recurrent Neural Network (RNN) [77] architecture. RNN is a variant of an ANN that is capable of processing and predicting sequenced patterns using LSTM memory units to capture temporal characteristics in time-series data. Using a RNN model would potentially produce more accurate predictions by considering how patient characteristics change within the first few treatment sessions and using that additional information in the learning process. The more time-series data points available per patient, the more refined a prediction would be. Instead of a class label, it would also be possible to predict a time-series sequence showing how response and remission rates change post-treatment. This would give physicians a better tool for analyzing the underlying long term effect of rTMS treatment.

Finally, we hope that our work can further motivate other work and trials to investigate similar approaches in data mining and machine learning for the advancement of mental health research.

Bibliography

[1] Anonymous "Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016," Lancet, vol. 390, pp. 1211-1259, Sep 16,. 2017.

[2] A. Cipriani, T.A. Furukawa, G. Salanti, J.R. Geddes, J.P. Higgins, R. Churchill, N. Watanabe, A. Nakagawa, I.M. Omori, H. McGuire, M. Tansella and C. Barbui, "Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis," Lancet, vol. 373, pp. 746-758, Feb 28,. 2009.

[3] A. Cipriani, P. Brambilla, T. Furukawa, J. Geddes, M. Gregis, M. Hotopf, L. Malvini and C. Barbui, "Fluoxetine versus other types of pharmacotherapy for depression," Cochrane Database Syst Rev, pp. CD004185, Oct 19,. 2005.

[4] A.J. Rush, S.R. Wisniewski, D. Warden, J.F. Luther, L.L. Davis, M. Fava, A.A. Nierenberg and M.H. Trivedi, "Selecting among second-step antidepressant medication monotherapies: predictive value of clinical, demographic, or first-step treatment features," Arch. Gen. Psychiatry, vol. 65, pp. 870-880, Aug. 2008.

[5] K. Rost, P. Nutting, J.L. Smith, C.E. Elliott and M. Dickinson, "Managing depression as a chronic disease: a randomised trial of ongoing treatment in primary care," Bmj, vol. 325, pp. 934, -10-26. 2002.

[6] K.R. Connolly, A. Helmer, M.A. Cristancho, P. Cristancho and J.P. O'Reardon, "Effectiveness of transcranial magnetic stimulation in clinical practice post-FDA approval in the United States: results observed with the first 100 consecutive cases of depression at an academic medical center," J Clin Psychiatry, vol. 73, pp. 567, Apr. 2012.

[7] A.J. Rush, C.M. Gullion, M.R. Basco, R.B. Jarrett and M.H. Trivedi, "The Inventory of Depressive Symptomatology (IDS): psychometric properties," Psychological Medicine, vol. 26, pp. 477-486, /05. 1996.

[8] A.J. Rush, M.H. Trivedi, H.M. Ibrahim, T.J. Carmody, B. Arnow, D.N. Klein, J.C. Markowitz, P.T. Ninan, S. Kornstein, R. Manber, M.E. Thase, J.H. Kocsis and M.B. Keller, "The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression," Biological Psychiatry, vol. 54, pp. 573-583, 2003.

[9] M. Hamilton, "Development of a rating scale for primary depressive illness," Br J Soc Clin Psychol, vol. 6, pp. 278-296, Dec. 1967.

[10] L.L. Carpenter, P.G. Janicak, S.T. Aaronson, T. Boyadjis, D.G. Brock, I.A. Cook, D.L. Dunner, K. Lanocha, H.B. Solvason and M.A. Demitrack, "Transcranial Magnetic Stimulation (TMS) for major depression: A multisite, naturalistic, observational study of acute treatment outcomes in clinical practice", Depression and Anxiety, vol. 29, pp. 587-596, Jul. 2012.

[11] C. Ciobanu, M. Girard, B. Marin, A. Labrunie and D. Malauzat, "rTMS for pharmacoresistant major depression in the clinical setting of a psychiatric hospital: Effectiveness and effects of age," Journal of Affective Disorders, vol. 150, pp. 677-681, 2013.

[12] K.R. Connolly, A. Helmer, M.A. Cristancho, P. Cristancho and J.P. O'Reardon, "Effectiveness of transcranial magnetic stimulation in clinical practice post-FDA approval in the United States: results observed with the first 100 consecutive cases of depression at an academic medical center," J Clin Psychiatry, vol. 73, pp. 567, Apr. 2012.

[13] W.K. Silverstein, Y. Noda, M.S. Barr, F. Vila-Rodriguez, T.K. Rajji, P.B. Fitzgerald, J. Downar, B.H. Mulsant, S. Vigod, Z.J. Daskalakis and D.M. Blumberger, "Neurobiological

Predictors of Response to Dorsolateral Prefrontal Cortex Repetitive Transcranial Magnetic Stimulation in Depression: A Systematic Review," Depression and Anxiety, vol. 32, pp. 871-891, -12-01. 2015.

[14] D.E. Goldberg and J.H. Holland, "Genetic algorithms and machine learning," Mach.Learning, vol. 3, pp. 95-99, 1988.

[15] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436, 2015.

[16] Y. Lee, R. Ragguett, R.B. Mansur, J.J. Boutilier, J.D. Rosenblat, A. Trevizol, E. Brietzke, K. Lin, Z. Pan, M. Subramaniapillai, T.C.Y. Chan, D. Fus, C. Park, N. Musial, H. Zuckerman, V.C. Chen, R. Ho, C. Rong and R.S. McIntyre, "Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review," J Affect Disord, vol. 241, pp. 519-532, Dec 01,. 2018.

[17] T.T. Erguzel, S. Ozekes, S. Gultekin, N. Tarhan, G. Hizli Sayar and A. Bayram, "Neural Network Based Response Prediction of rTMS in Major Depressive Disorder Using QEEG Cordance," Psychiatry Investig, vol. 12, pp. 61-65, -1. 2015.

[18] F. De Jonghe, S. Kool, G. Van Aalst, J. Dekker and J. Peen, "Combining psychotherapy and antidepressants in the treatment of depression," J.Affect.Disord., vol. 64, pp. 217-229, 2001.

[19] P.G. Mottram, K. Wilson and J.J. Strobl, "Antidepressants for depressed elderly," Cochrane

[20] S. Rossi, M. Hallett, P.M. Rossini, A. Pascual-Leone and Safety of TMS Consensus Group, "Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research," Clinical Neurophysiology, vol. 120, pp. 2008-2039, 2009.

[21] Anonymous "Transcranial magnetic stimulation," 2015.

[22] D. Kolar, "Current status of electroconvulsive therapy for mood disorders: a clinical review," Evidence-Based Mental Health, vol. 20, pp. 12, 2017.

[23] R.W. Lam, P. Chan, M. Wilkins-Ho and L.N. Yatham, "Repetitive transcranial magnetic stimulation for treatment-resistant depression: a systematic review and metaanalysis," The Canadian Journal of Psychiatry, vol. 53, pp. 621-631, 2008.

[24] E.M. Wassermann and T. Zimmermann, "Transcranial magnetic brain stimulation: Therapeutic promises and scientific gaps," Pharmacology & Therapeutics, vol. 133, pp. 98-107, 2012.

[25] M.S. George, S.H. Lisanby, D. Avery, W.M. McDonald, V. Durkalski, M. Pavlicova, B. Anderson, Z. Nahas, P. Bulow, P. Zarkowski, P.E. Holtzheimer, T. Schwartz and H.A. Sackeim, "Daily left prefrontal transcranial magnetic stimulation therapy for major depressive disorder: a sham-controlled randomized trial," Arch. Gen. Psychiatry, vol. 67, pp. 507-516, May. 2010.

[26] K. Nguyen and L.G. Gordon, "Cost-Effectiveness of Repetitive Transcranial Magnetic Stimulation versus Antidepressant Therapy for Treatment-Resistant Depression," Value Health, vol. 18, pp. 597-604, Jul. 2015.

[27] N. Bakker, S. Shahab, P. Giacobbe, D.M. Blumberger, Z.J. Daskalakis, S.H. Kennedy and J. Downar, "rTMS of the Dorsomedial Prefrontal Cortex for Major Depression: Safety, Tolerability, Effectiveness, and Outcome Predictors for 10 Hz Versus Intermittent Theta-burst Stimulation," Brain Stimulation, vol. 8, pp. 208-215, 2015.

[28] S.M. McClintock, I.M. Reti, L.L. Carpenter, W.M. McDonald, M. Dubin, S.F. Taylor, I.A. Cook, J. O'Reardon, M.M. Husain, C. Wall, A.D. Krystal, S.M. Sampson, O. Morales, B.G.

Nelson, V. Latoussakis, M.S. George and S.H. Lisanby, "Consensus Recommendations for the Clinical Application of Repetitive Transcranial Magnetic Stimulation (rTMS) in the Treatment of Depression," J Clin Psychiatry, vol. 79, Jan/Feb. 2018.

[29] Y. Huang, M.J. Edwards, E. Rounis, K.P. Bhatia and J.C. Rothwell, "Theta burst stimulation of the human motor cortex," Neuron, vol. 45, pp. 201-206, Jan 20,. 2005.

[30] N. Grossheinrich, A. Rau, O. Pogarell, K. Hennig-Fast, M. Reinl, S. Karch, A. Dieler, G. Leicht, C. Mulert and A. Sterr, "Theta burst stimulation of the prefrontal cortex: safety and impact on cognition, mood, and resting electroencephalogram," Biol.Psychiatry, vol. 65, pp. 778-784, 2009.

[31] D.M. Blumberger, F. Vila-Rodriguez, K.E. Thorpe, K. Feffer, Y. Noda, P. Giacobbe, Y. Knyahnytska, S.H. Kennedy, R.W. Lam, Z.J. Daskalakis and J. Downar, "Effectiveness of theta burst versus high-frequency repetitive transcranial magnetic stimulation in patients with depression (THREE-D): a randomised non-inferiority trial," Lancet, vol. 391, pp. 1683-1692, 04 28,. 2018.

[32] S.B. Kotsiantis, I. Zaharakis and P. Pintelas, "Supervised machine learning: A review of classification techniques," Emerging Artificial Intelligence Applications in Computer Engineering, vol. 160, pp. 3-24, 2007.

[33] M.A.T. Figueiredo and A.K. Jain, "Unsupervised learning of finite mixture models," IEEE Trans.Pattern Anal.Mach.Intell., vol. 24, pp. 381-396, 2002.

[34] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural

language processing-Volume 10, pp. 79-86, 2002.

[35] F. Mosteller and J.W. Tukey, "Data analysis and regression: a second course in statistics." Addison-Wesley Series in Behavioral Science: Quantitative Methods, 1977.

[36] A.K. Jain, M.N. Murty and P.J. Flynn, "Data clustering: a review," ACM Computing Surveys (CSUR), vol. 31, pp. 264-323, 1999.

[37] A. Liaw and M. Wiener, "Classification and regression by randomForest," R News, vol. 2, pp. 18-22, 2002.

[38] L. Breiman, "Random forests," Mach.Learning, vol. 45, pp. 5-32, 2001.

[39] H. Kavade, "A logistic regression model to predict incident severity using the human factors analysis and classification system," 2009.

[40] R.C. Kessler, H.M. van Loo, K.J. Wardenaar, R.M. Bossarte, L.A. Brenner, T. Cai, D.D. Ebert, I. Hwang, J. Li, P. de Jonge, A.A. Nierenberg, M.V. Petukhova, A.J. Rosellini, N.A. Sampson, R.A. Schoevers, M.A. Wilcox and A.M. Zaslavsky, "Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports," Mol. Psychiatry, vol. 21, pp. 1366-1371, 10. 2016.

[41] J.L. Krupnick, S.M. Sotsky, I. Elkin, S. Simmens, J. Moyer, J. Watkins and P.A. Pilkonis, "The role of the therapeutic alliance in psychotherapy and pharmacotherapy outcome: Findings in the National Institute of Mental Health Treatment of Depression Collaborative Research Program," Focus, vol. 64, pp. 532-277, 2006.

[42] B. Yegnanarayana, Artificial neural networks, PHI Learning Pvt. Ltd., 2009, .

[43] A.M. Chekroud, R.J. Zotti, Z. Shehzad, R. Gueorguieva, M.K. Johnson, M.H. Trivedi, T.D. Cannon, J.H. Krystal and P.R. Corlett, "Cross-trial prediction of treatment outcome in depression: a machine learning approach," The Lancet Psychiatry, vol. 3, pp. 243-250, 2016.

[44] A. Kautzky, M. Dold, L. Bartova, M. Spies, T. Vanicek, D. Souery, S. Montgomery, J. Mendlewicz, J. Zohar, C. Fabbri, A. Serretti, R. Lanzenberger and S. Kasper, "Refining Prediction in Treatment-Resistant Depression: Results of Machine Learning Analyses in the TRD III Sample," J Clin Psychiatry, vol. 79, Jan/Feb. 2018.

[45] R. Iniesta, K. Malki, W. Maier, M. Rietschel, O. Mors, J. Hauser, N. Henigsberg, M.Z. Dernovsek, D. Souery, D. Stahl, R. Dobson, K.J. Aitchison, A. Farmer, C.M. Lewis, P. McGuffin and R. Uher, "Combining clinical variables to optimize prediction of antidepressant treatment outcomes," J Psychiatr Res, vol. 78, pp. 94-102, 07. 2016.

[46] R.H. Perlis, "A clinical risk stratification tool for predicting treatment resistance in major depressive disorder," Biol. Psychiatry, vol. 74, pp. 7-14, Jul 01,. 2013.

[47] T.T. Erguzel and N. Tarhan, "Machine Learning Approaches to Predict Repetitive Transcranial Magnetic Stimulation Treatment Response in Major Depressive Disorder," pp. 391-401, 2016/9/21.

[48] A. Khodayari-Rostamabad, J.P. Reilly, G.M. Hasey, H. deBruin and D. MacCrimmon, "Using pre-treatment electroencephalography data to predict response to transcranial magnetic stimulation therapy for major depression," pp. 6418-6421, 2011.

[49] D.V. Sheehan, Y. Lecrubier, K.H. Sheehan, P. Amorim, J. Janavs, E. Weiller, T. Hergueta, R. Baker and G.C. Dunbar, "The Mini-International Neuropsychiatric Interview (M.I.N.I.): the

development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10," J Clin Psychiatry, vol. 59 Suppl 20, pp. 57, 1998.

[50] L.R. Derogatis and N. Melisaratos, "The Brief Symptom Inventory: an introductory report," Psychological Medicine, vol. 13, pp. 595-605, /08. 1983.

[51] J.P. O'Reardon, H.B. Solvason, P.G. Janicak, S. Sampson, K.E. Isenberg, Z. Nahas, W.M. McDonald, D. Avery, P.B. Fitzgerald, C. Loo, M.A. Demitrack, M.S. George and H.A. Sackeim, "Efficacy and safety of transcranial magnetic stimulation in the acute treatment of major depression: a multisite randomized controlled trial," Biol. Psychiatry, vol. 62, pp. 1208-1216, Dec 01,. 2007.

[52] M.S. George, S.H. Lisanby, D. Avery, W.M. McDonald, V. Durkalski, M. Pavlicova, B. Anderson, Z. Nahas, P. Bulow, P. Zarkowski, P.E. Holtzheimer, T. Schwartz and H.A. Sackeim, "Daily left prefrontal transcranial magnetic stimulation therapy for major depressive disorder: a sham-controlled randomized trial," Arch. Gen. Psychiatry, vol. 67, pp. 507-516, May. 2010.

[53] C. Cusin, H. Yang, A. Yeung and M. Fava, "Rating Scales for Depression," in Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health, Humana Press, Totowa, NJ, 2009, pp. 7-35.

[54] I. Jolliffe, "Principal component analysis," in International encyclopedia of statistical science, Springer, 2011, pp. 1094-1096.

[55] Y. Saeys, T. Abeel and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 313-325, 2008.

[56] K.J. Archer and R.V. Kimes, "Empirical characterization of random forest variable importance measures," Comput.Stat.Data Anal., vol. 52, pp. 2249-2260, 2008.

[57] J. Snoek, H. Larochelle and R.P. Adams, "Practical bayesian optimization of machine learning algorithms," in Advances in neural information processing systems, pp. 2951-2959, 2012.

[58] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Ijcai, pp. 1137-1145, 1995.

[59] N. Japkowicz, "The class imbalance problem: Significance and strategies," in Proc. of the Int'l Conf. on Artificial Intelligence, 2000.

[60] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[61] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," IEEE Intelligent Systems and their Applications, vol. 13, pp. 18-28, 1998.

[62] S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K. Mullers, "Fisher discriminant analysis with kernels," in Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop. pp. 41-48, 1999.

[63] J.M. Keller, M.R. Gray and J.A. Givens, "A fuzzy k-nearest neighbor algorithm," IEEE Trans.Syst.Man Cybern., pp. 580-585, 1985.

[64] G.H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pp. 338-345, 1995.

[65] S.R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE Trans.Syst.Man Cybern., vol. 21, pp. 660-674, 1991.

[66] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," IEEE Trans.Inf.Theory, vol. 26, pp. 26-37, 1980.

[67] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv Preprint arXiv:1412.6980, 2014.

[68] R. Fan, K. Chang, C. Hsieh, X. Wang and C. Lin, "LIBLINEAR: A library for large linear classification," Journal of Machine Learning Research, vol. 9, pp. 1871-1874, 2008.

[69] L. Breiman, "Some properties of splitting criteria," Mach.Learning, vol. 24, pp. 41-47, 1996.

[70] A. Baratloo, M. Hosseini, A. Negida and G. El Ashal, "Part 1: simple definition and calculation of accuracy, sensitivity and specificity," 2015.

[71] J.A. Hanley and B.J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve." Radiology, vol. 143, pp. 29-36, 1982.

[72] A.P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognit, vol. 30, pp. 1145-1159, 1997.

[73] H. Tindle, B.H. Belnap, P.R. Houck, S. Mazumdar, M.F. Scheier, K.A. Matthews, F. He and B.L. Rollman, "Optimism, response to treatment of depression, and rehospitalization after coronary artery bypass graft surgery," Psychosomatic Medicine, vol. 74, pp. 200-207, Feb. 2012.

[74] D. Souery, P. Oswald, I. Massat, U. Bailer, J. Bollen, K. Demyttenaere, S. Kasper, Y. Lecrubier, S. Montgomery, A. Serretti, J. Zohar and J. Mendlewicz, "Clinical factors associated

with treatment resistance in major depressive disorder: results from a European multicenter study," J Clin Psychiatry, vol. 68, pp. 1062-1070, Jul. 2007.

[75] Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice

[76] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, pp. 1735-1780, 1997.

[77] T. Mikolov, M. Karafiát, L. Burget, J. Černocký and S. Khudanpur, "Recurrent neural network based language model," in Eleventh Annual Conference of the International Speech Communication Association, 2010.