

School of Computer Science

Scientific Research & Literature

in Fulfilment of

SPEC9997

Maksymilian Drzezdzon

C15311966

Degree: TU060/2

Module Supervisor: Giancarlo Salton

Declaration of Ownership: I declare that the attached work is entirely my own and that all sources have been acknowledged: 🗹  
**Date: 2021/03/24**

Investigating differences in performance between SVM and DWD in Schizophrenia classification



***Maksymilian Drzezdzon***

*C15311966*

A dissertation submitted in partial fulfilment of requirements of Technological University Dublin for the degree of

M.Sc. in Computer Science

May 2022

# DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computer Science, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute’s guidelines for ethics in research.

**Signed:** Maksymilian Drzezdzon

**Date:** 17/03/2022

# ABSTRACT

**Introduction**

There is an ever-growing emphasis on mental health with greater demands for already sparse mental health professionals. Applying machine learning as another criteria or test to satisfy a diagnosis for a mental illness such as schizophrenia would not only save valuable resources such as mental health practitioners time but also get patients the care they desperately need faster.

Young girls tend to be underdiagnosed with attention deficit hyperactivity disorder (ADHD) because how unalike its manifestation is when compared to boys, this levels out during adulthood, however having better tools for diagnosing disorders especially ones as serious as schizophrenia will greatly improve people’s quality of life and allow for medical intervention.

Schizophrenia is a disabling mental illness with huge time requirements to attain a diagnosis, a patient could avoid years of misdiagnosis/waiting if less severe or noticeable symptoms could be diagnosed/detected sooner through better diagnostic tools.

Obsessive compulsive disorder (OCD) and psychosis tend to be misdiagnosed as one another because the DSM-5 (diagnostic and statistical manual of mental disorders 5th edition) categorizes certain behavioral traits belonging to specific illnesses which in practice requires a lot more time and caution for evaluation, medications used to treat a patient experiencing psychosis which they may not have would actually exacerbate their OCD, leaving space for doubling down on/reinforcing a misdiagnosis that is then further mis/interpreted as psychosis after the fact.

There have been a few short comings in diagnosing of serious mental health disorders, there is no process to date that properly diagnoses dissociative identity disorder, despite it being acknowledged as a mental illness in the 1950s, research between then and now has been conducted but later found fraudulent or difficult to reproduce.

**Objective**

The goal of this study is to compare misclassification rates between state-of-the-art implementations of support vector machine (SVM) and distance weighed discrimination (DWD) applied to schizophrenia detection/classification through prototyping of classification models based on state-of-the-art research using preprocessed MRI and fMRI image modalities. Followed up by identifying areas for future work.

Give some indication of possible directions this work could go in the future.

Describe the sections of articles to come

**Results**

Body text - Body text Body text Body text Body text Body text Body text Body text

Keywords: Diagnosis Prediction, Schizophrenia Classification, Precision Psychiatry, Support Vector Machine, Distance Weight Discrimination

Word Count: **xxxxxx**

# ACKNOWLEDGEMENTS

Dataset used for this thesis is from a

Collection of this dataset was made at the [Mind Research Network](http://www.mrn.org/) under an NIH NIGMS Centers of Biomedical Research Excellence (COBRE) grant 5P20RR021938/P20GM103472 to Vince Calhoun (PI).

Notes for later:

Thank supervisor etc

REMOVE THIS LATER THIS WAS CUT FROM ABSTRACT – not really needed, add to intro instead

One study found and attained an accuracy of 70% with a 75% baseline when classifying schizophrenia on synthetic data with a DFNN model. In another study an accuracy of 84.4% was attained when using a SVM model on FNC between independent components extracted by ICA. Finally, one study achieved an accuracy of 87% on network maps extracted by ICA with an SVM model.

, provide scope of schizophrenia and MRI image formats and the digital imaging and communications (DICOM) standard which provides a central medium for imaging modalities.

Table of Contents

[DECLARATION i](#_Toc105600552)

[ABSTRACT ii](#_Toc105600553)

[ACKNOWLEDGEMENTS iii](#_Toc105600554)

[Chapter 1 – INTRODUCTION 1](#_Toc105600555)

[Background 1](#_Toc105600556)

[Research Problem 2](#_Toc105600557)

[Research Question 3](#_Toc105600558)

[Research Objectives 4](#_Toc105600559)

[Aim: 4](#_Toc105600560)

[Objectives: 4](#_Toc105600561)

[Research Methodologies 5](#_Toc105600562)

[Scope and Limitations 5](#_Toc105600563)

[Document Outline – TODO update at end 6](#_Toc105600564)

[Chapter 2 – LITERATURE REVIEW 7](#_Toc105600565)

[Introduction 7](#_Toc105600566)

[Approaches to solve the problem 8](#_Toc105600567)

[Gaps in Research 9](#_Toc105600568)

[Machine Learning in Mental Health 10](#_Toc105600569)

[Overview of Schizophrenia 10](#_Toc105600570)

[HDLSS - TODO read paper 10](#_Toc105600571)

[Imaging Techniques – TODO read that other paper 10](#_Toc105600572)

[Functional Network Connectivity – FNC 11](#_Toc105600573)

[Sourced-Based Morphometry – SBM 11](#_Toc105600574)

[Evaluation 11](#_Toc105600575)

[Chapter 3 – Design and Methodology 12](#_Toc105600576)

[Hypothesis 13](#_Toc105600577)

[Hypothesis testing 13](#_Toc105600578)

[Dataset 14](#_Toc105600579)

[Acknowledgement 14](#_Toc105600580)

[Dataset 14](#_Toc105600581)

[Data preparation 15](#_Toc105600582)

[Model training and tuning - TODO needs work 15](#_Toc105600583)

[Feature selection 16](#_Toc105600584)

[RFE - Recursive Feature Elimination 16](#_Toc105600585)

[RRF - Regularized Random Forest 17](#_Toc105600586)

[LASSO - Last Absolute Shrinkage and Selection Operator 18](#_Toc105600587)

[Rpart - Decision Tree Rpart 20](#_Toc105600588)

[BORUTA 21](#_Toc105600589)

[QVT – (Data) Quality and Variable Training (Samples) 22](#_Toc105600590)

[Software/Programing Environment used 22](#_Toc105600591)

[Project Approach 22](#_Toc105600592)

[Evaluation 23](#_Toc105600593)

[Chapter 4 – Results and Evaluation 25](#_Toc105600594)

[Results 25](#_Toc105600595)

[Explain results 27](#_Toc105600596)

[Conclusion 27](#_Toc105600597)

[Chapter 5 – Conclusion and Future Work 28](#_Toc105600598)

[Research Overview 28](#_Toc105600599)

[Research findings 28](#_Toc105600600)

[Future work and recommendations 28](#_Toc105600601)

[References 29](#_Toc105600602)

# Chapter 1 – INTRODUCTION

## Background

Machine learning is gaining popularity in many industries, one such industry is healthcare which requires explainable results, some of the most powerful machine learning techniques available based on deep learning don’t produce explainable models.

A diagnosis for dissociative disorders or schizophrenia among other disorders rely on the interpretation of an assessment completed by a clinician, (HSE, 2021) this process is time consuming and prone to error especially when a patient does not exhibit severe symptoms such as delusions or the flattening of emotions, which is a condition where a person is unable to express emotions the same way other people might (Timothy J. Legg, 2017) Similar overlapping symptoms can cause a misdiagnosis and leave a patient astray for years before finally being diagnosed for instance with schizophrenia or a disassociation disorder.

Currently there is no non-invasive methods for diagnosing schizophrenia and no established biomarker for diagnosis besides using the process of elimination (Cynthia S. Weickert, 2013) (Heline Mirzakhanian, 2014).

One way to reduce misdiagnosis is with the use of machine learning classification in conjunction with MRI and fMRI images. Once a biomarker for other illnesses such as depression or anxiety can be identified they can then be acknowledged and ‘omitted’ when searching for definitive biomarkers that help hone in on schizophrenia or dissociative disorders.

Following the traditional approach of elimination with a little more precision that could cater to different demographics as expressed in precision medicine, however that is outside the scope of this project.

An obstacle that occurs when collecting data, apart from privacy concerns and difficulty obtaining such data due to regulations, schizophrenia only afflicts ~1% of the population making it very scarce. (mentalhelp.net, 2021) When working with images one can rotate them to synthesize more data for model training, however this may not be ideal.

Machine learning algorithms and traditional statistical techniques such as support vector machine, linear discriminant analysis, multivariate analysis, neural networks, regression, k-nearest neighbor, k-means clustering and random forest have been found to be most popular and effective when attempting to detect/classify schizophrenia. (Adrian B. R. Shatte, 2019)

When talking about data, high dimensional low sample size (HDLSS) is typically not the first thing that comes to mind. However, in computer vision, complex mental illness classification such as schizophrenia via magnetic resonance imaging (MRIs), gene analysis, chemo-metrics handling it is one of many obstacles experts still wrestle with today (J. S. Marron, 2007) (Oh, 2020) (Chang Su, 2020) (Wanwan Zheng, 2022) (J.Rodríguezb, 2018).

(HDLSS) data is data that has more features than samples. When working with HDLSS data variations of classification methods such as support vector machine (SVM) and distance weighted discrimination (DWD) are used, (Delaram Sadeghi, 2021) each implementation tries to address shortcomings of its predeceasing method such as, neighbor-less nature (Cheema, 2015) or data piling in SVM (Jeongyoun Ahn, 2015) which is what prompted the creation of DWD. (J. S. Marron, 2007) What makes these challenges unique is that traditional classification approaches don't work and require tailored solutions. Most of these implementations don't have framework implementations yet which make experimentation difficult. (J. S. Marron, 2007) (Lock, 2020) Although deep learning models command impressive results (0.96 on an AUC) they tend to over fit and miss-classify younger cohorts with less severe/profound symptoms (Oh, 2020) (Cortes, 2021) (Marjane Khodatars, 2021).

## Research Problem

This project investigates the classification effectiveness of SVM and DWD on modalities extracted from MRIs.

“If this can be accomplished through machine learning classification, what is the accuracy and error rate of each model, how do they compare?”

Support vector machine (SVM) is the state-of-the-art when working with HDLSS data. (Tianmeng Lyu, 2017) (Delaram Sadeghi, 2021) It’s a supervised learning algorithm (Boser, 1992) that utilizes a hyper-plane to separate two classes and assign them to either 'group'.

However, there is an issue if you don’t have enough data which would be the norm in a HDLSS setting, points mapped onto vectors end up or are close to 0 if not actually 0 making them identical, resulting in new samples not being properly grouped using a vanilla SVM classifier.

There is a hybrid approach called DWSVM which incorporates DWD which finds the classification direction and then determines the intercept using SVM which should deal with data pilling, however there is no publicly available implementation of this algorithm to compare. (Xingye Qiao, 2015)

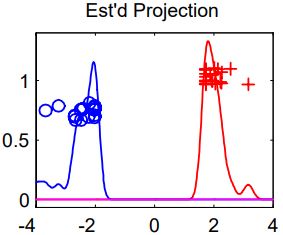


Figure 1: Example SVM data piling - Linear SVM, C = 1000, dimension = 39 (J. S. Marron, 2007)

This phenomenon is referred to as “*data-piling*”, which causes SVM to over-fit because it can’t use the data vector to differentiate points between the two classes resulting in tightly centered groups (J. S. Marron, 2007) (Jeongyoun Ahn, 2015) (AHN, 2010). This is where DWD and SVMs many implementations come into play. The solution to SVMs data piling is DWD.

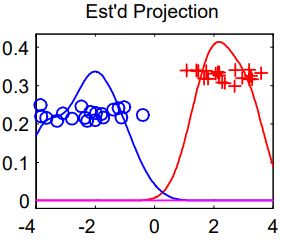


Figure 2: Example DWD no data piling - Distance Weighted Disc., dimension = 39 (J. S. Marron, 2007)

Looping back to the previous section regarding tailored solutions, there is a plethora of DWD implementations of which most don't have public code implementations appended or published making it difficult to compare approaches for experimentation.

## Research Question

"What are the differences in performance between implementations of SVM and DWD when classifying Schizophrenia using HDLSS data through fMRI/FNC features and sMRI/SBM loadings?"

## Research Objectives

This project is based on secondary research, using the Mind Research Network data-set supported by a systematic review of existing literature on SVM and DWD implementations for schizophrenia classification.

This study aims to investigate the difference in effectiveness between DWD and SVM through classification accuracy metrics such as F1 score, Log Loss and AUC to gauge its performance when applied to HDLSS data for schizophrenia classification using the Mind Research Network data-set. Leading to the acceptance or rejection of the null hypothesis based on results gathered from model evaluation.

### Aim:

To quantify the differences in misclassification rate through mean squared error between state-of-the-art implementations of SVM and DWD for HDLSS data in mental illness detection/classification

### Objectives:

* Implement the following implementations of SVM: Polynomial, Radial, Linear and Sigmoid
* Implement the following implementations of DWD: Polynomial, Radial, Linear, Sparse
* Setup feature selection methods for each algorithm implementation
* Gather metrics from each model on unfiltered data vs subsets gathered from feature selection
* Repeat feature selection and training if time allows with QVT implementation (Wanwan Zheng, 2022)
* Format results into table with regards to which algorithm was used, feature selection, improvement criteria & QVT impact if implemented

## Research Methodologies

In this experiment the performance between SVM and DWD implementations will be examined using HDLSS data with an objective of using quantitative research methods via the development of classification models evaluated by F1 score, Log Loss, and AUC that will lead to the acceptance or rejection of the null hypothesis based on sample results gathered from model evaluation

A review of previous literature was carried out, scoping the area of machine learning applications to mental health diagnostics and classification. A dataset has been found and cited for this undertaking. Each algorithm will have several feature selection methods used in conjunction with QVT if time allows for it is a significant feature selection augmentation for HDLSS datasets. (Wanwan Zheng, 2022)

## Scope and Limitations

The scope of this project is to build classification models comparing their performance on HDLSS data.

* This project will attempt to provide metrics on algorithm implementations as a means of comparing them for future work, given the timeframe and number of implementations used some may not be finetuned perfectly which would should be taken into account.
* This project is limited to the MLSP dataset.
* Preprocessing MRIs from schizconnect is time consuming and out of scope for the time allocated for this project, which would have otherwise yielded more test and train data.
* Severity of schizophrenia could be a factor in model sensitivity, this detail wasn’t recorded into the dataset, meaning that if this model was used in a production environment it may struggle on a broader patient cohort.

## Document Outline – TODO update at end

Chapter 2 - Literature review

Reviewing existing literature in mental illness classification using (update this with whats in chp 2). Surface level overview of schizophrenia and motivations of this research along with imaging techniques. Review methods used in acquiring data, FNC and SBM.

Chapter 3 - Design and methodology

This chapter focusses on how the project is conducted, the dataset used, its features, data preparation and the proposed solution. Model evaluation is discussed along with a description of how the project will be conducted providing methods employed to test the hypothesis.

The hypothesis is stated followed by how it will be tested, later a description of the dataset used is provided, how the data was initially acquired by the original research team, how it was prepared and explored for this project. A brief list of software, languages and tools used.

Chapter 4 - Results and evaluation

This chapter focusses on summarizing results of the experiment and evaluate the proposed method. The chapter concludes with a discussion on strengths and limitations of the proposed solution highlighting potential improvements and areas for future research.

Chapter 5 - Conclusion

A summary of key findings, conclusions and areas for future research.

# Chapter 2 – LITERATURE REVIEW

## Introduction

This chapter covers the approach to solve the problem, reviews gaps in research, provides an overview of machine learning in mental health, an anecdote of schizophrenias impact on the healthcare system followed by an overview of FNC and SBM methods used to extract the data.

The state-of-the-art research focuses on two categories of methods, machine learning predominately via SVM (Olivier Chapelle, 2002) and deep learning techniques (Delaram Sadeghi, 2021). The downfall of deep learning (DL) especially in the HDLSS space is due to it requiring a lot of data, this of course is difficult to facilitate when you don’t have access to more data. (Oh, 2020).

The main differences between a deep learning model compared to traditional machine learning from a model building perspective is feature acquisition, as illustrated in figure 3 (Delaram Sadeghi, 2021) (A. Jović, 2015) (Mwangi, 2014). Hence the focus on feature selection augmentation such as QVT and linear dependency filtering. (J.Rodríguezb, 2018) (Wanwan Zheng, 2022)

LINEAR DEPENDENCY PAPERS

Traditional machine learning models require more manual data curation than deep learning models that use auto feature selection. This is where the majority of the work is when working with HDLSS datasets as there is a need to be precise with how you extract useful features and discard others.

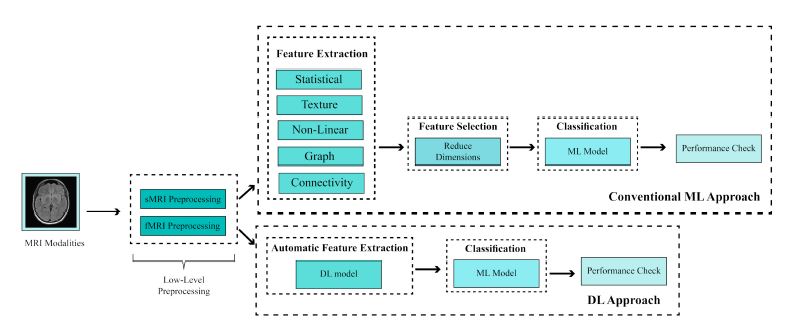


Figure 3: Difference in machine learning vs deep learning model diagnosing (Delaram Sadeghi, 2021)

Most used feature extraction techniques can be grouped into categories which are (Delaram Sadeghi, 2021) (Federico Calesella, 2021) statistical that encompass the mean, variance and standard deviation. Textural features, a more specialized method in medical imaging utilized in deep learning (Georgiadis, 2008), non-linear feature extraction uses neuro-imaging modalities with a publicly available framework hosted on GitHub called NeAT (Adrià Casamitjana, 2020) (Pinotsis, 2014) and graph models (Yizhen Xiang, 2020) (Sugai Liang, 2018) are used in a HDLSS setting to combat how computationally expensive feature selection is with larger HDLSS data-sets. (Zhihong Zhang, 2011)

Feature reduction narrows down to principal component analysis as a common approach when leveraging traditional machine learning methods. (Delaram Sadeghi, 2021) (Svante Wold, 1987) The large majority of these studies focus around SVM or deep learning methods utilizing methods such as convolutional neural networks, recurrent neural networks, generative adversarial networks and the list goes on. Other more sophisticated implementations such as CADS are used, however those require huge compute power along with profound expertise in the topic to tune these models (Delaram Sadeghi, 2021).

These can be validated using methods such as T-tests and analysis of variance (ANOVA) for feature selection in neuro-medical images. In all studies reviewed accuracy, sensitivity and recall were used to compare performance (Delaram Sadeghi, 2021) (Mwangi, 2014).

### Approaches to solve the problem

Current approaches to tackle HDLSS data are different implementations that build upon SVM, DWD or deep learning methods. Current state-of-the-art methods seem to use deep learning (Delaram Sadeghi, 2021) (Qingbo Yin, 2020) (Huang, 2012) (Artemiou, 2021). As mentioned previously DWD has seen little adoption over SVM or deep learning with research beginning to pile up over the last few years, over a decade after its initial proposal in 2004/2007. (J. S. Marron, 2007)

### Gaps in Research

J. S. Marron authored DWD (J. S. Marron, 2007) to tackle SVMs data piling problem. DWD is sensitive to the sample size ratio between classes denoted by the intercept term β, this is a problem because when taking into consideration the differences between cohorts age, stage of schizophrenia, type of schizophrenia among other intricacies that make it hard to diagnose and distinguish between, once and if accounted for, this can cause/causes a class imbalance, this limitation is blunted with refined implementations of DWD such as wDWD and DWSVM but at a significantly higher computational cost (Qingbo Yin, 2020). With so many different SVM and DWD implementations that lack code samples its very difficult to parse and know which method preforms best in which context.

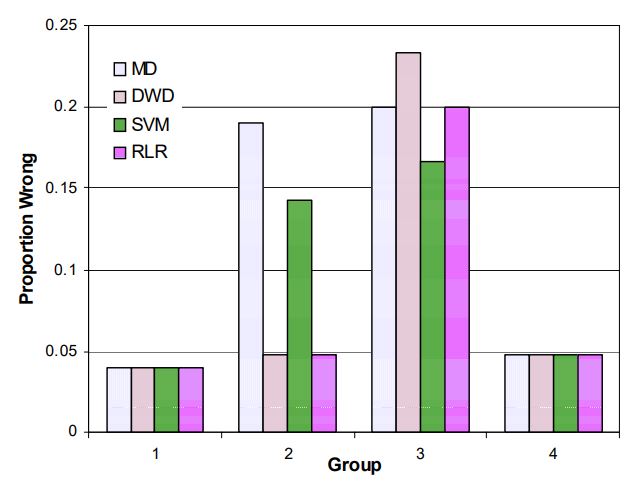


Figure 3: Error rates for gene expression data (J. S. Marron, 2007)

Despite how well DWD and its various implementations cater to HDLSS data (Huang, 2012) (Artemiou, 2021) it’s not widely used in most state of the art research (Delaram Sadeghi, 2021) even with its competitive performance seen in figure 4 above (J. S. Marron, 2007) , this seems like an application/implementation/ methodological gap where DWD and its following implementations are only being picked up on more recently. The goal is to compare different SVM and DWD implementations with available implementations online and compile a set of results in the 20 weeks allocated for this project.

An alternative study could be conducted in the future to implement other variations of SVM and DWD in python or R libraries as most don't have ready to use implementations or have not been maintained. A machine learning R library called caret has built in support for model kernels, implementing these methods could submitted to the author for review.

## Machine Learning in Mental Health

Common applications of ML in healthcare encompass detection, aid in determining diagnosis, prognosis, treatment and support, public health applications, research and clinical administration.

These can be further segregated into pre-diagnosis screening tools and risk models that identify individual’s predisposition or susceptibility to develop mental health conditions. (Adrian B. R. Shatte, 2019)

In this project the disorder being examined is schizophrenia with mention of disassociation as they have overlapping symptomology.

Other avenues worth exploring are the use of NLP to detect onset of schizophrenia, similar was accomplished with Alzheimer’s disease. (Elif Eyigoz, 2020) Potentially analyzing social media activity could also be a viable path for future work depending on data availability as eccentric behavior is linked to schizophrenia, given the correlation between social isolation as a result of serious mental illnesses. (Sheila J. Linz PMHNP-BC, 2013)

## Overview of Schizophrenia

Schizophrenia is a serious mental illness for the afflicted but also costly to the healthcare system. With limited healthcare resources such as psychiatrists time for diagnosis and continued care managing said condition. Finding better ways to help with diagnostics would not only help the individual get treated for the correct condition but also alleviate stress put on clinicians.

Schizophrenia was first identified by Emil Kraepelin in 1896 under the name dementia praecox. (R M Ion, 2002) It’s very difficult to diagnose schizophrenia due to the fact that it overlaps with many other mental illnesses or conditions such as disassociation, ADHD, psychosis etc. (Selwyn B. Renard, 2017)

### HDLSS - TODO read paper

High dimensionality low sample size data (Dan Shen, 2016)

### Imaging Techniques – TODO read that other paper

#### DICOM

Digital imaging and communications in medicine standard is a data interchange protocol for biomedical image format’s structure. (W. Dean Bidgood, 1997)

#### MRI

Magnetic resonance imaging is used in radiology to take none invasive images of brain and brain stem structure. (Michael Harkin, 20217)

#### fMRI

Functional magnetic resonance images are similar to MRIs but depict the changes in blood oxygen levels. It’s been used in conjunction with statistical methods for classification for concluding inferences about brain states. (Glover, 2012)

#### PET

A positron emission tomography scan uses a chemical/dye containing tracers which can be viewed by a PET scanner which then detects which cells that absorb more of this chemical. In brain disorder classification this can be used to detect levels of glucose similar to SBM weights, PET scans can be used to inspect regions of the brain that use more or less glucose. Currently its used for Alzheimer’s disease and depression. (Brian Krans, 2018) depression is sometimes diagnosed in tandem with schizophrenia.

Maybe write here about the use of AI/ML with Imaging techniques ?? otherwise this is redundant. Do they have an impact of the model?

## Functional Network Connectivity – FNC

FNC are correlation values that summarize the overall connection between independent brain maps over time through correlation in statistical analysis. FNC describe patterns of brain function. (Elena A. Allen, 2014) It’s important to note that this data refers to the state at a given point in time, meaning patients must be in the same state when this kind of data is being gathered. This is done through MRI, fMRI, EEG or MEG. In this project these values were acquired from fMRI from schizophrenic patients and healthy controls at rest (rs-MRI) with group independent component analysis. (Elena A Allen 1, 2011)

## Sourced-Based Morphometry – SBM

SBM loadings are weights of brain maps gathered from gray matter concentration maps using independence component analysis. These values are also derived from MRI scans and represent a patient’s brain structure. (Judith M Segall 1, 2012) The goal behind such metrics is that they provide cognitive capability for each region of the brain through statistical analysis.

# Chapter 3 – Design and Methodology

This chapter will focus on the experiment used to evaluate models and accept or reject the null hypothesis. Several sets of models are implemented for both SVM and DWD. All models will be tested on the Mind Research Network MLSP 2014 dataset. Data preparation, feature selection is described along with the train/test process.

The F1 score, Cohens Kappa, McNamar’s Test P-value, PPV (positive predicted value), NPV (negative predicted value) and AUC will be collected for each model’s assessment along with the public and private Kaggle scores.

Finally, a state-of-the-art implementation of DW-SVM would be implemented and compared to the rest of the models.

## Hypothesis

**Null Hypothesis:**

If SVM implementations are used to classify schizophrenic patient’s vs healthy controls using fMRI/FNC features (correlation values that summarize connection between brain maps over time) and sMRI/SBM loadings (weights of brain maps derived from gray matter concentration of all subjects) then the mean difference between the F1 score, Log Loss and AUC of the top 5 SVM models and top 5 DWD models will not be statistically significant.

**Alternate Hypothesis:**

If SVM implementations are used to classify schizophrenic patient’s vs healthy controls using fMRI/FNC features (correlation values that summarize connection between brain maps over time) and sMRI/SBM loadings (weights of brain maps derived from gray matter concentration of all subjects) then the mean difference between the F1 score, Log Loss and AUC of the top 5 SVM models and top 5 DWD models will be statistically significant.

## Hypothesis testing

The type of statistical test is dependent on the sampling technique used, for each model trained, in this scenario, using k-fold validation will mean that estimated metrics are dependent, it further limits the possibility for truly independent samples (Brownlee, Statistical significance tests for comparing machine learning algorithms, 2019) (Brownlee, A gentle introduction to estimation statistics for Machine Learning, 2019).

Using a Non-parametric tests such as a paired t-test Wilcoxon singed rank test can be used to validate the hypothesis, albeit it holds less statistical power. Alternatively, estimation statistics can be used such as effect size, interval estimation, confidence intervals or meta-analysis that can be used to accept or reject the null hypothesis.

There are a few ways to do this, such as the use of PCA to lower the number of features by their highest eigen-values for dimension reduction or factor analysis followed by a MANOVA for a global hypothesis test.

This is preferred if data is of Gaussian distribution validated or otherwise in objective 8, this rule has been slightly relaxed over time \cite{muller\_stewart\_2006}. Another option is PCA or factor analysis, for PCA to be successful it requires a “simple co-variance structure, at least asymptotically” (Yueh-Yun Chi, 2013).

Factor analysis is sensitive to an unequal ratio of observations to variables which also holds true for PCA. There are other specialized alternatives, however more time needs to be allocated to identify better suited method/s.

Reject the null hypothesis (Ho) if p < 0.05 from the above hypothesis test using a MANOVA with the use of PCA/factor analysis is found.

## Dataset

Methods used to acquire these coordinates were FNC & SBM, group independent component analysis and independent component analysis.

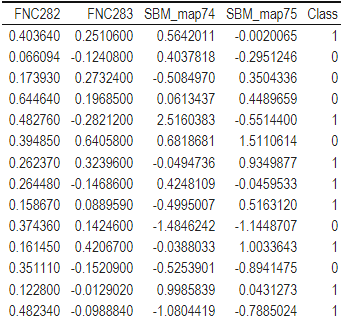
### Acknowledgement

Collection of this dataset was made at the [Mind Research Network](http://www.mrn.org/) under an NIH NIGMS Centers of Biomedical Research Excellence (COBRE) grant 5P20RR021938/P20GM103472 to Vince Calhoun (PI)

### Dataset

The MLSP dataset consists of 86 observations with 411 features in which 46 patients are labeled as schizophrenic and 40 labeled as healthy controls. The response variable for this dataset is the class column which represents a patient’s condition coded using binary values of 0 and 1, 0 indicating a healthy control and 1 a schizophrenic patient. The MLSP datasets features are divided between two/three datasets. The FNC portion holds 378 features whereas the SBM dataset has 32, with the ‘third’ dataset storing the labels adding up to 411 features, if the ID column is counted as a feature in the above description, it would bump the total feature count by 3 as its present across the three datasets.

Table 1: MLSP subset data sample



### Data preparation

The FNC and SBM train sets are loaded in and merged by the ID column alongside the labels dataset that contains the binary value for whether a patient has schizophrenia. Each row is a patient with 410 features spread between SBM and FNC. The experiment was run against several subsets of the original dataset that were filtered by select feature selection algorithms along with removing linear dependencies on remaining features before training each model. Each subset was saved to a csv for ease of use as experiments were carried out. The goal of this approach was to remove noise and reduce

### Model training and tuning

Each model was first trained using a train control parameter using repeated cross validation 10 times with 10 folds, allowing parallel processing if a model supports it to save time before saving the final predictions. This was done to get a sense of initial parameters that would be selected for a model. Afterword a grid would be constructed using values that encompass the initial values selected to try find better settings. Typically, this approach increased the models AUC by 0.03-0.14 and NPV/PPV by 0.10. Models such as sparse DWD need more manual tunning as lambda values would go below 0.

Use bullet points?

Make a diagram?

Talk about different kernels used?

## Feature selection

Feature selection is a process in which an algorithm is used to filter/rank features in order to measure how much they contribute to predicting the target variable and which features lower the accuracy of a model, the latter is very important for linear models. (Why is it important for linear models)

SVM and DWD are linear classifiers, this is why data was filtered for linear dependencies to further reduce noise **(citation).** The following feature selection algorithms and plots were used to asses which and how many features to keep.

Write more about what dwd and svm are sensitive to?

### RFE - Recursive Feature Elimination

RFE is a wrapper feature selection algorithm that uses another algorithm to establish a baseline. (X. Zeng, 2009) In this project random forest was passed to RFE as a baseline algorithm. Random forest was picked as it seems to be the most agnostic feature selection algorithm because it uses decision trees to gauge a features contribution to the model, cross validation was also used as a means of gathering features, resampling 10 times. In figure 4 you can see that the large majority of features in the MLSP dataset don’t seem to contribute to predicting the target class but rather reduce performance. In figure 5 features that held a variable importance score of ~1.8 were kept.

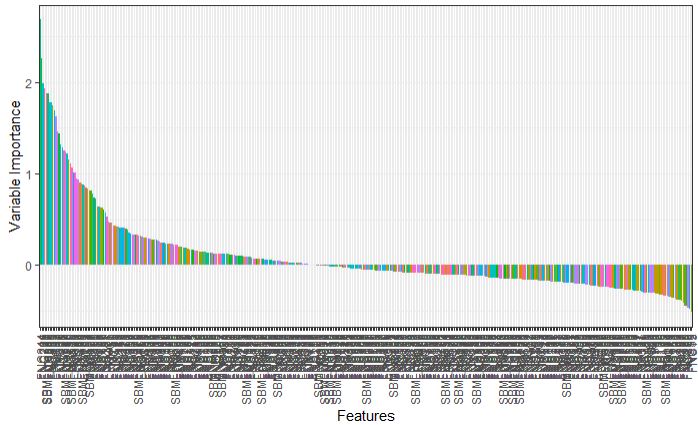


Figure 4: All features ranked by importance by RFE using the full MLSP dataset

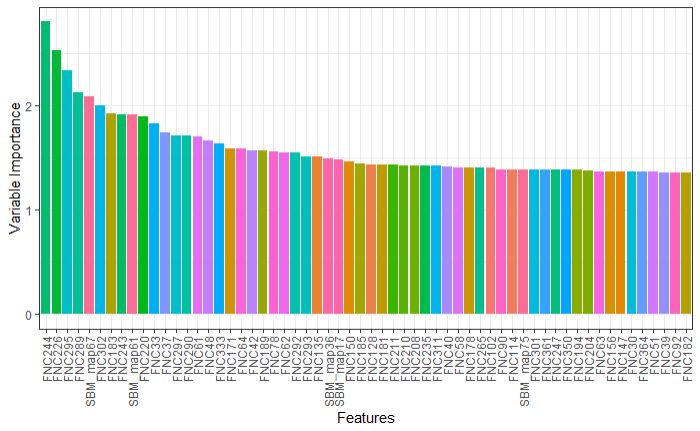


Figure 5: RFE selected features from full MLSP dataset

### RRF - Regularized Random Forest

The version used from the R package caret computes one prototype per class. A prototype is a group of data points based on their similarity. For each observation in the input dataset a nearest neighbor is selected. Later a prototype that has most neighbors of a class is identified. This prototype is then the medoid of these neighbors for its class. (Deng, 2022)

This algorithm was used to provide a baseline when compared to the wrapped alternative constructed using RFE.

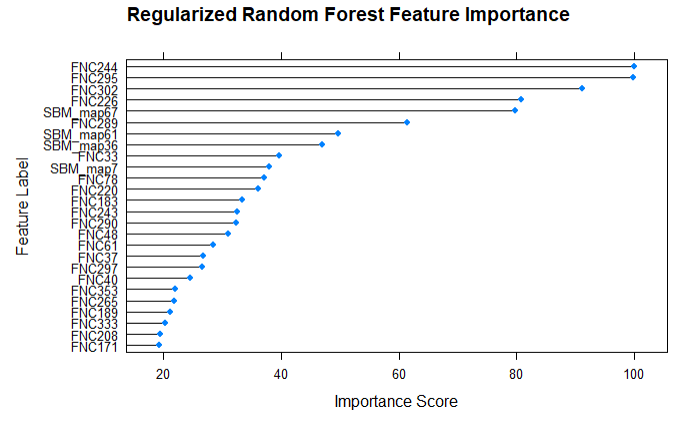


Figure 6: Feature importance on select features from the full MLSP dataset

### LASSO - Last Absolute Shrinkage and Selection Operator

The LASSO algorithm fits a least squares criterion:

Where all predictors are centered and scaled to put on the same scale, is chosen using k-fold cross validation which was used for all feature selection methods. This results in the ’s being shrunk to 0 given that lambda is large enough, this is accomplished by getting the absolute value of which can shrink it to 0. What this does is it allows for filtering out features that provide little to no use to the model.

The left dashed line is the min value whereas the right dashed line is 1se. min is the value for which the model has the lowest cross validation MSE, as seen in figure 7 said value would encompass 19 features and 1se would include 12. 1se is the MSE with a larger value of within 1 standard error of the lowest MSE. (Sunghoon Kwon, 2013)

In this project 20 features were kept despite 1se being the typical choice, this decision was made because features would then be checked for linear dependencies delegating some of that responsibility to another function instead of solely depending on LASSO.

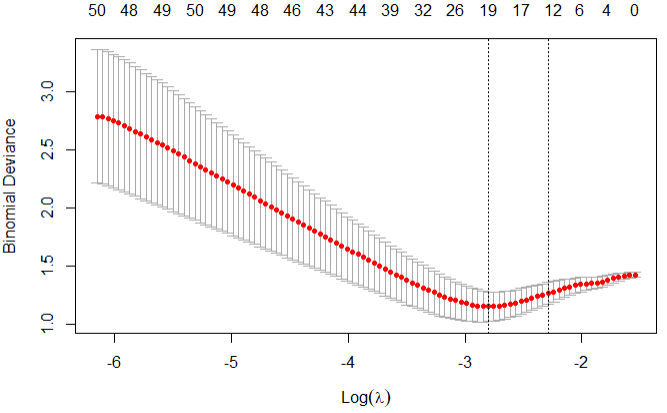


Figure 7: LASSO trained using the full MLSP dataset

### Rpart - Decision Tree Rpart

The Rpart algorithm recursively splits data which is dictated by the independent variable that would yield the largest reduction in variability of the predictor. These kinds of models are very transparent in how they work and can be easily interpreted.

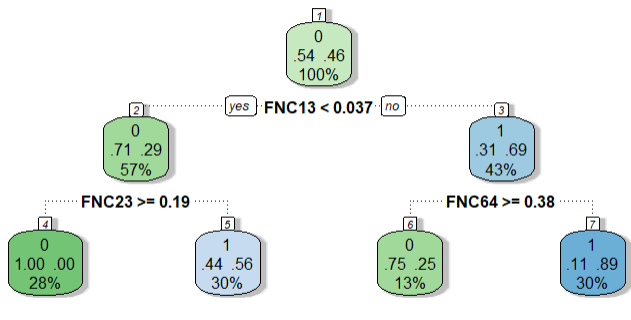


Figure 8: Rpart tree for full MLSP dataset

As seen in figure 8 above, 46% (0.46) of all patients without a schizophrenia diagnosis indicated by 0 within the top most green tag (the root of the tree) had a FNC13 value of less than 0.037, furthermore, 69 % (0.69) of people in that group who did have schizophrenia indicated by 1 had a value of FNC64 greater or equal to 0.38 and finally of that 69%, 89% had diagnosed schizophrenia.

Alternatively, 54% (0.54) of all non-schizophrenic patients that did have a FNC13 value greater than 0.037 within their group 71% (0.71) had a FNC23 value of 0.19 or greater and were all not previously diagnosed with schizophrenia making up 28% of all healthy controls. This could indicate that the FNC23 feature is a predictor of the absence of schizophrenia and could/should be studied in isolation by medical scientists.

This is a small example tree plot, it could be further expanded visualizing more of the “if-then” rulesets found.

### 

### BORUTA

Boruta is another wrapper algorithm, however, its specifically built around the random forest classification algorithm, unlike RFE you don’t have a choice of which algorithm you use as a core. Feature importance is calculated using z-scores via shuffling values from each column and internally labeling them as shadow features of the original feature and training a random forest model using this setup. Following this step, the algorithm checks original features if they have a higher z-score than the maximum z-score among shadowed features, before removing all shadowed features from the final list of retained features. (Kursa, 2013)

Shuffling is used to compare the importance of a feature versus random shuffles of shadow features, similarly to generating a dummy feature within the range of a real feature, anything below that dummy feature would be discarded as this dummy feature is obsolete because its randomly generated given a start and stop range.

In figure 9, the blue boxplots are the min, average and max z-scores. Red represents rejected features, yellow are tentative features, albeit in this plot there is only 1 and green resembling retained features. (Kursa, 2013)

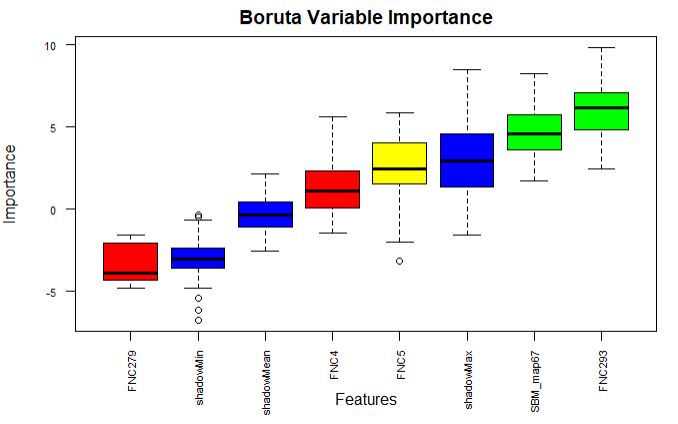


Figure 9: Boruta variable importance boxplots for full MLSP dataset

### QVT – (Data) Quality and Variable Training (Samples)

Internally QVT uses mutual information as a metric of gauging the benefit of using one variable over another when initially ranking them with the best preforming samples being put on top.

It uses a kindred forward selection approach supplemented with the feature selection algorithm passed in, meaning that from the ranked list synthesized above it will use a subset ***s*** to create a new collection of samples by selecting the top K samples from each class stored in the list above ***F1*** followed by preforming feature selection using the passed in algorithm saving in a new list **F2**. Where size is the number of samples in ***s***. ***F1*** is a first pass of the data.

Using the original ranked feature list ***F1*** and the new one ***F2*** created in the step above it will calculate the average importance score for overlapping features from the two lists and record feature and its average score in a new list ***Fc***. Then for each feature in ***F1*** and ***F2*** store its importance score, append to ***Fc*** and output the final list. (Wanwan Zheng, 2022)

## Software/Programing Environment used

The R programming language was used with R studio, every library used in the [r notebook](https://github.com/Maks-Drzezdzon/Masters-Classes-L-O/blob/master/Thesis/models_notebook.Rmd) is available via the c-ran package manager.

## Project Approach

The scope is limited to examining classification techniques such as SVM, DWD and their implementations over a period of ~13 weeks applied to FNC/SBM correlation values gathered from HDLSS data from the Mind Research Network’s Schizophrenia Dataset consisting of 35,432 observations gathered from 162 patients and 169 healthy controls. The Aim of this study is to derive differences between implementations and their classification accuracy via the F1 score

## Evaluation

Models are tested using 30% of each input dataset split before training. The original MLSP dataset contains 410 features with the 411th being the label. The main metric used is AUC with F1 score, Cohens Kappa, McNemar’s Test P-value, positive predicated value and negative predicted value further supplemented by Kaggle’s private and public scores gathered from the original competition created in 2014.

* AUC was used to gauge how well a model is able to distinguish between classes in a classification model. (Whitehill, 2019)

Metric interpretation: Measured on a scale of 0-1, the higher the AUC the better the model can discern between predicted classes.

* F1 score was used to get an idea of how the model’s precision and recall preform. The reason why the F2 score wasn’t used is because it prioritizes recall, in a case like this you don’t want to have false negatives as well as false positives hence F1 score was the better choice to measure all misclassifications.

Metric interpretation: Measured on a scale of 0-1, the higher the F1 score the better the average recall and precision, a downfall of one will have a knock-on effect on the overall F1 score.

* Cohens Kappa compares how good a model’s predictions are compared to random chance (McHugh, 2012) – this was used as a more robust means of validating the model and to a lesser extent the public Kaggle score.

A downfall of this metric in this project is that it depends on representation of the thing being tested. Meaning when there is an underrepresentation of something, such as in a HDLSS setting a low Kappa value may not actually be representative of how many of a model’s predications can be explained by chance. What it also does is it checks for potential bias by taking correct classification by chance into account. When a dataset is balanced it automatically pushes the value up (Viera AJ, 2005).

Metric interpretation: Measured on a scale of 0-1, the higher the score the better, which means that x% of the models’ predictions are not attributed to chance.

* McNemar’s Test P-value is a distribution free hypothesis test typically used in biological research to discriminate intergroup differences which translate to a probability that this difference could result in different responses, in medicine it measures the effect of a treatment vs its control (Pembury Smith, 2020).

When McNemar’s Test P-value is applied in a machine learning context such as these binary classification models this metric is comparing whether one predicted class is more error prone than another when a model is trained on a single dataset split into train and test. It can also be used to compare two algorithms which wasn’t done in this project due to time constraints, in that scenario it would compare whether one algorithm is more accurate or error prone than its comparate (Dietterich, 1998).

Metric interpretation: Measured on a scale of 0-1, when the p-value is greater than 0.05 it means that the model misclassifies one class more than the other.

When comparing 2 algorithms if the p-value is less than 0.05 then it means that these algorithms have different error rates and one would be better than the other, the latter would need another metric to confirm which and by how much, the test only checks for if there is a difference between the two.

* Positive predicted value is a measure of true positives

Metric interpretation: Measured on a scale of 0-1, what % of true positives were correct.

* Negative predicted value is a measure of true negatives

Metric interpretation: Measured on a scale of 0-1, what % of true negatives were correct.

* Private score - Kaggle uses 99% of data available to validate results
* Public score - Kaggle uses 1% of the data available for the private score

The private score would be much more important as a good or bad public score could have been attained by chance. An unlabeled dataset is used to test each model, predicted values are submitted for evaluation at kaggle.com/c/mlsp-2014-mri, the higher the score the better the model preforms.

# Chapter 4 – Results and Evaluation

## Results

Table 2: SVM model results table

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SVM Kernel | Model Metrics | | | | | | Kaggle Evaluation Scores | | Feature Selection |
| Metric used | F1 | Cohens Kappa | Mcnemar’s  Test P-value | PPV | NPV | AUC | Private | Public | Method used |
| Polynomial | 0.6364 | 0.3548 | 0.72367 | 0.7 | 0.6667 | 0.68 | 0.79 | 0.630 | N/A |
|  | 0.6957 | 0.4373 | 1 | 0.7273 | 0.7413 | 0.72 | 0.679 | 0.616 | RFE |
|  | 0.72 | 0.441 | 1 | 0.6923 | 0.75 | 0.72 | 0.779 | 0.777 | RRF |
|  | 0.555 | 0.3464 | 0.077 | 0.83 | 0.631 | 0.67 | 0.63 | 0.58 | LASSO |
|  | 0.5833 | 0.1987 | 1 | 0.5833 | 0.615 | 0.6 | 0.679 | 0.54 | Rpart |
|  | 0.8696 | 0.7588 | 0.7588 | 0.91 | 0.857 | 0.88 | 0.741 | 0.71 | BORUTA |
| Radial | 0.5263 | 0.2671 | 0.1824 | 0.7143 | 0.6111 | 0.63 | 0.708 | 0.643 | N/A |
|  | 0.6957 | 0.4373 | 1 | 0.7273 | 0.7143 | 0.72 | 0.737 | 0.713 | RFE |
|  | 0.74 | 0.44 | 0.44969 | 0.667 | 0.8 | 0.72 | 0.751 | 0.741 | RRF |
|  | 0.6 | 0.3506 | 0.2888 | 0.75 | 0.647 | 0.71 | 0.741 | 0.571 | LASSO |
|  | 0.6364 | 0.3548 | 0.72367 | 0.7 | 0.6667 | 0.68 | 0.669 | 0.58 | Rpart |
|  | 0.833 | 0.6795 | 1 | 0.833 | 0.8462 | 0.84 | 0.751 | 0.705 | BORUTA |
| Linear | 0.5714 | 0.2718 | 0.505 | 0.667 | 0.625 | 0.63 | 0.684 | 0.54 | N/A |
|  | 0.7273 | 0.5161 | 0.68309 | 0.8 | 0.733 | 0.76 | 0.674 | 0.616 | RFE |
|  | 0.6667 | 0.5098 | 0.041 | 1 | 0.733 | 0.75 | 0.779 | 0.714 | RRF |
|  | 0.88 | 0.7604 | 1 | 0.8462 | 0.9167 | 0.88 | 0.54 | 0.59 | LASSO |
|  | 0.5 | 0.0385 | 1 | 0.5 | 0.539 | 0.52 | 0.679 | 0.607 | Rpart |
|  | 0.8462 | 0.682 | 0.617 | 0.7857 | 0.91 | 0.84 | 0.718 | 0.763 | BORUTA |
| Sigmoid | 0.5714 | 0.2718 | 0.505 | 0.667 | 0.625 | 0.63 | 0.79 | 0.696 | N/A |
|  | 0.6957 | 0.4373 | 1 | 0.7273 | 0.7143 | 0.72 | 0.774 | 0.652 | RFE |
|  | 0.72 | 0.4409 | 1 | 0.6923 | 0.75 | 0.72 | 0.746 | 0.808 | RRF |
|  | 0.75 | 0.5192 | 1 | 0.75 | 0.7692 | 0.76 | 0.713 | 0.643 | LASSO |
|  | 0.72 | 0.4409 | 1 | 0.6923 | 0.75 | 0.72 | 0.723 | 0.634 | Rpart |
|  | 0.818 | 0.678 | 0.617 | 0.9 | 0.8 | 0.84 | 0.674 | 0.741 | BORUTA |
|  | ***Note****: RRF – Regularized Random Forest, RFE – Recursive Feature Elimination, LASSO – Last Absolute Shrinkage and Selection Operator, Rpart – Decision Tree Rpart, PPV – Positive Predicted Value, NPV – Negative Predicted Value* | | | | | | | | |

Table 3: DWD model results table

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DWD Kernel | Model Metrics | | | | | | Kaggle Evaluation Scores | | Feature Selection |
| Metric used | F1 | Cohens Kappa | Mcnemar’s  Test P-value | PPV | NPV | AUC | Private | Public | Method used |
| Linear | 0.7 | 0.513 | 0.22 | 0.875 | 0.7059 | 0.75 | 0.644 | 0.584 | N/A |
|  | 0.8462 | 0.6815 | 0.617 | 0.785 | 0.909 | 0.84 | 0.794 | 0.61 | RFE |
|  | 0.8182 | 0.677 | 0.617 | 0.9 | 0.8 | 0.84 | 0.702 | 0.714 | RRF |
|  | 0.761 | 0.595 | 0.371 | 0.888 | 0.75 | 0.8 | 0.651 | 0.629 | LASSO |
|  | 0.63 | 0.35 | 0.723 | 0.7 | 0.667 | 0.68 | 0.756 | 0.562 | Rpart |
|  | 0.4 | 0.257 | 0.0076 | 1 | 0.59 | 0.62 | 0.653 | 0.53 | BORUTA |
| Radial | 0.7143 | 0.367 | 0.2888 | 0.625 | 0.777 | 0.69 | 0.641 | 0.642 | N/A |
|  | 0.7692 | 0.522 | 0.683 | 0.7143 | 0.8182 | 0.76 | 0.817 | 0.741 | RFE |
|  | 0.75 | 0.5192 | 1 | 0.75 | 0.769 | 0.76 | 0.717 | 0.607 | RRF |
|  | 0.7273 | 0.51 | 0.683 | 0.8 | 0.733 | 0.76 | 0.741 | 0.571 | LASSO |
|  | 0.6 | 0.3506 | 0.288 | 0.75 | 0.647 | 0.67 | 0.75 | 0.53 | Rpart |
|  | 0.833 | 0.6795 | 1 | 0.833 | 0.846 | 0.84 | 0.642 | 0.642 | BORUTA |
| Poly | 0.6667 | 0.359 | 1 | 0.6667 | 0.6923 | 0.68 | 0.684 | 0.633 | N/A |
|  | 0.88 | 0.76 | 1 | 0.8462 | 0.9167 | 0.88 | 0.679 | 0.669 | RFE |
|  | 0.9565 | 0.9196 | 1 | 1 | 0.928 | 0.96 | 0.641 | 0.642 | RRF |
|  | 0.667 | 0.5098 | 0.04 | 1 | 0.6842 | 0.75 | 0.6358 | 0.571 | LASSO |
|  | 0.6957 | 0.434 | 1 | 0.727 | 0.714 | 0.72 | 0.717 | 0.531 | Rpart |
|  | 0.8182 | 0.6774 | 0.617 | 0.9 | 0.8 | 0.84 | 0.669 | 0.718 | BORUTA |
| Sparse | 0.69 | 0.36 | 0.72 | 0.642 | 0.727 | 0.68 | 0.607 | 0.58 | N/A |
|  | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | RFE |
|  | 0.72 | 0.4409 | 1 | 0.6923 | 0.75 | 0.72 | 0.679 | 0.736 | RRF |
|  | 0.7 | 0.513 | 0.22 | 0.875 | 0.7 | 0.75 | 0.70 | 0.48 | LASSO |
|  | 0.4 | 0.257 | 0.0076 | 1 | 0.59 | 0.62 | 0.717 | 0.642 | Rpart |
|  | 0.47 | 0.26 | 0.0455 | 0.8 | 0.6 | 0.63 | 0.664 | 0.59 | BORUTA |
|  | ***Note****: RRF – Regularized Random Forest, RFE – Recursive Feature Elimination, LASSO – Last Absolute Shrinkage and Selection Operator, Rpart – Decision Tree Rpart, PPV – Positive Predicted Value, NPV – Negative Predicted Value* | | | | | | | | |

Currently sparse DWD wasn’t working with RFE and produced wrong values or N/A.

## Explain results

Review metrics here and comment on them, is all.

## Conclusion

Draw conclusion from results

# Chapter 5 – Conclusion and Future Work

## Research Overview

xxx

## Research findings

xxxx

## Future work and recommendations

There is a large body of potential work that could be done, a lot of which was cut from this project. Schizconnect has over 500 gigabytes of MRI data spread across several database. These data need to be processed with SPM in order to feed to a machine learning algorithm, alternately a library that can read the unprocessed files would suffice. Once that’s done a larger sample can be used for the above experiment, most importantly allowing for a larger test set.

A lot of DWD implementations were dropped as any code implementations found were not compatible with test frameworks before or after implementation, didn’t have documentation on hyperparameters available, weren’t maintained and some were removed from cran-r and were unavailable for download likewise less common SVM implementations were just as difficult to interact with and test.

Re-implementing these into python and/or R would be great for more through comparisons between algorithms in HDLSS settings, potentially submitting the final project to the maintainer of caret an R machine learning library as additional kernel functions would be great.

Another portion that was dropped is the implementation of the QVT algorithm. This is a specialized method used for improving HDLSS feature selection. This alone could be a project in itself when used to compare benefits on specialized HDLSS algorithms. (citation)

Other forms of feature selection augmentation can be used such as survival analysis () citation

Look up more FS methods for HDLSS data

# References

A. Jović, K. B. (2015). A review of feature selection methods with applications. *Electronics and Microelectronics (MIPRO)*, pp. 1200-1205, doi: 10.1109/MIPRO.2015.7160458.

Adrià Casamitjana, V. V. (2020). NeAT: a Nonlinear Analysis Toolbox for Neuroimaging. *neuroinformatics* , 517–530 (2020). https://doi.org/10.1007/s12021-020-09456-w.

Adrian B. R. Shatte, D. M. (2019). Machine learning in mental health: A systematic scoping review of methods and applications. *Psychological medicine*, 49(9), 1426–1448. https://doi.org/10.1017/S0033291719000151.

AHN, J. &. (2010). The maximal data piling direction for discrimination. *Biometrika*, 97(1), 254–259. http://www.jstor.org/stable/27798914.

Artemiou, A. R. (2021). Sufficient dimension reduction based on distance-weighted discrimination. *Scandinavian Journal of Statistics*, https://doi.org/10.1111/sjos.12484.

Boser, G. V. (1992). *Support Vector Machines.* Boston, MA: Springer.

Brian Krans, C. S. (2018, September 17). *What Is a PET Scan?* Retrieved from healthline: https://www.healthline.com/health/pet-scan#purpose

Brownlee, J. (2019, August 8). *A gentle introduction to estimation statistics for Machine Learning*. Retrieved from machinelearningmastery.com: https://machinelearningmastery.com/estimation-statistics-for-machine-learning/?fbclid=IwAR0mdkFN\_hFzvAtIinRey2LueMdUR8oAeQ3hX3pqz\_UHHJBuc9-iwKEn\_ug

Brownlee, J. (2019, August 8). *Statistical significance tests for comparing machine learning algorithms*. Retrieved from machinelearningmastery.com: https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms

Chang Su, Z. X. (2020). Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, 10(1), 116. https://doi.org/10.1038/s41398-020-0780-3.

Cheema, M. S. (2015). High dimensional low sample size activity recognition using geometric classifiers. *Digit. Signal Process*, 61–69. https://doi.org/10.1016/j.dsp.2015.03.019.

CONNOLLY, M. (2020, Jan 17). *adhdireland*. Retrieved from Easy-to-Miss ADHD Symptoms in Girls: https://adhdireland.ie/what-adhd-looks-like-in-girls/

Cortes, J. &.-R. (2021). Going deep into schizophrenia with artificial intelligence. *Schizophrenia Research*, 10.1016/j.schres.2021.05.018.

Cynthia S. Weickert, T. W. (2013). Biomarkers in Schizophrenia: A Brief Conceptual Consideration. *Disease markers*, 10.1155/2013/510402.

Dan Shen, H. S. (2016). The Statistics and Mathematics of High Dimension Low Sample Size Asymptotics. *Statistica Sinica*, 26(4), 1747–1770. 10.5705/ss.202015.0088.

Delaram Sadeghi, A. S.-D. (2021). An Overview on Artificial Intelligence Techniques for Diagnosis of Schizophrenia Based on Magnetic Resonance Imaging Modalities: Methods, Challenges, and Future Works. *Advanced Researches In Biomedical Engineering Lab.*, https://doi.org/10.48550/arXiv.2103.03081.

Deng, H. (2022, 05 30). *Package ‘RRF’*. Retrieved from cran.r-project.org: https://cran.r-project.org/web/packages/RRF/RRF.pdf

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms . *Neural Comput*, https://doi.org/10.1162/089976698300017197.

Elena A Allen 1, E. B. (2011). A baseline for the multivariate comparison of resting-state networks. *Frontiers in systems neuroscience*, 5, 2. https://doi.org/10.3389/fnsys.2011.00002.

Elena A. Allen, E. D. (2014). Tracking Whole-Brain Connectivity Dynamics in the Resting State . *Cerebral Cortex*, 663–676, https://doi.org/10.1093/cercor/bhs352.

Elif Eyigoz, S. M. (2020). Linguistic markers predict onset of Alzheimer's disease. *The Lancet*, https://doi.org/10.1016/j.eclinm.2020.100583.

Federico Calesella, A. T. (2021). A comparison of feature extraction methods for prediction of neuropsychological scores from functional connectivity data of stroke patients. *Brain Informatics* , https://doi.org/10.1186/s40708-021-00129-1.

Georgiadis, P. &. (2008). Improving brain tumor characterization on MRI by probabilistic neural networks on non-linear transformation of textural features. *Computer methods and programs in biomedicine*, 89. 24-32. 10.1016/j.cmpb.2007.10.007 .

Glover, G. H. (2012). Overview of Functional Magnetic Resonance Imaging. *Neurosurgery clinics of North America*, 10.1016/j.nec.2010.11.001.

Heline Mirzakhanian, F. S. (2014). Biomarkers in psychosis: an approach to early identification and individualized treatment. *Biomarkers in medicine*, 10.2217/bmm.13.134.

HSE. (2021, April 4). *Diagnosis Schizophrenia*. Retrieved from HSE.ie: https://www2.hse.ie/conditions/mental-health/schizophrenia/schizophrenia-diagnosis.html#:~:text=There's%20no%20single%20test%20for,the%20result%20of%20other%20causes.

Huang, H. &. (2012). R/DWD: Distance Weighted Discrimination for Classification, Visualization and Batch Adjustment. *Bioinformatics* , 28. 1182-3. 10.1093/bioinformatics/bts096.

J. S. Marron, M. J. (2007). Distance-Weighted Discrimination. *Journal of the American Statistical Association*, 102(480), 1267–1271. http://www.jstor.org/stable/27639976.

J.Rodríguezb, L. I. (2018). On feature selection protocols for very low-sample-size data. *Elsevier*, 10.1016/j.patcog.2018.03.012.

Jeongyoun Ahn, Y. J. (2015). Sparse HDLSS discrimination with constrained data piling. *Computational Statistics & Data Analysis*, 74-83, https://doi.org/10.1016/j.csda.2015.04.006.

Judith M Segall 1, E. A. (2012). Correspondence between structure and function in the human brain at rest. *Frontiers in neuroinformatics*, 6, 10. https://doi.org/10.3389/fninf.2012.00010.

Kursa, M. B. (2013). Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(11), 1–13. https://doi.org/10.18637/jss.v036.i11.

Lock, E. F. (2020). Bayesian Distance Weighted Discrimination. *arXiv*, 10.48550/ARXIV.2010.03111.

Marjane Khodatars, A. S. (2021). Deep learning for neuroimaging-based diagnosis and rehabilitation of Autism Spectrum Disorder: A review. *Computers in Biology and Medicine*, 104949. https://doi.org/10.1016/j.compbiomed.2021.104949.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 276–282, ncbi.nlm.nih.gov/pmc/articles/PMC3900052/?report=classic.

mentalhelp.net. (2021, April 4). *Schizophrenia Symptoms, Patterns and Statistics and Patterns*. Retrieved from mentalhelp.net: https://www.mentalhelp.net/schizophrenia/statistics/#:~:text=Worldwide%20about%201%20percent%20of,this%20year%20around%20the%20world.

Michael Harkin, S. H. (20217, March 29). *What is a head MRI?* Retrieved from healthline: https://www.healthline.com/health/head-mri

Mwangi, B. T. (2014). A Review of Feature Reduction Techniques in Neuroimaging. *Neuroinformatics*, 229–244 (2014). https://doi.org/10.1007/s12021-013-9204-3.

Oh, J. a.-L.-U.-H. (2020). Identifying Schizophrenia Using Structural MRI With a Deep Learning Algorithm. *Frontiers in Psychiatry*, 10.3389/fpsyt.2020.00016.

Olivier Chapelle, V. V. (2002). Choosing Multiple Parameters for Support Vector Machines. *Springer Machine Learning*, 131–159. https://doi.org/10.1023/A:1012450327387.

Pembury Smith, M. R. (2020). Effective use of the McNemar test. *Behav Ecol Sociobiol*, https://doi.org/10.1007/s00265-020-02916-y.

Pinotsis, D. &. (2014). Extracting novel information from neuroimaging data using neural fields. *Nonlinear Biomedical Physics*, 2. 5. 10.1186/1471-2202-15-S1-O4. .

Qingbo Yin, E. A. (2020). Population-guided large margin classifier for high-dimension low-sample-size problems. *Elsevier*, https://doi.org/10.1016/j.patcog.2019.107030.

R M Ion, M. D. (2002). The British reaction to dementia praecox 1893-1913. Part 1. *Hist Psychiatry*, 10.1177/0957154X0201305103.

Reinders, A. M. (2019). Aiding the diagnosis of dissociative identity disorder: pattern recognition study of brain biomarkers. *The British journal of psychiatry* , 215(3), 536–544. https://doi.org/10.1192/bjp.2018.255.

Selwyn B. Renard, R. J. (2017). Unique and Overlapping Symptoms in Schizophrenia Spectrum and Dissociative Disorders in Relation to Models of Psychopathology: A Systematic Review. *Schizophrenia bulletin*, 10.1093/schbul/sbw063.

Sheila J. Linz PMHNP-BC, M. R. (2013). The Phenomenon of Social Isolation in the Severely Mentally ill. *Perspect Psychiatr Care*, 49: 243-254, https://doi.org/10.1111/ppc.12010.

Sugai Liang, R. V. (2018). Neurocognitive Graphs of First-Episode Schizophrenia and Major Depression Based on Cognitive Features. *Neuroscience bulletin*, 10.1007/s12264-017-0190-6.

Sunghoon Kwon, S. H. (2013). A small review and further studies on the LASSO. *Journal of the Korean Data and Information Science Society*, 10.7465/jkdi.2013.24.5.1077.

Svante Wold, K. E. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 37-52, doi.org/10.1016/0169-7439(87)80084-9.

Tianmeng Lyu, E. F. (2017). Discriminating sample groups with multi-way data. *Biostatistics*, Pages 434–450, 10.1093/biostatistics/kxw057.

Timothy J. Legg, P. (2017, August 4). *What is Flat Affect?* Retrieved from healthline: https://www.healthline.com/health/flat-affect#:~:text=Flat%20affect%20is%20a%20condition,affect%20shows%20no%20facial%20expressions.

Viera AJ, G. J. (2005). Understanding interobserver agreement: the kappa statistic. *Family medicine*, 360–363.

W. Dean Bidgood, J. M. (1997). Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging. *Journal of the American Medical Informatics Association*, 10.1136/jamia.1997.0040199.

Wang, B. Z. (2018). Another look at distance-weighted discrimination. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1369-7412. https://doi.org/10.1111/rssb.12244.

Wanwan Zheng, M. J. (2022). Improving the Performance of Feature Selection Methods with Low-Sample-Size Data . *The Computer Journal*, https://doi.org/10.1093/comjnl/bxac033.

Whitehill, J. (2019). How Does Knowledge of the AUC Constrain the Set of Possible Ground-Truth Labelings? *Worcester Polytechnic Institute*, https://doi.org/10.1609/aaai.v33i01.33015425.

X. Zeng, Y. -W. (2009). Feature Selection Using Recursive Feature Elimination for Handwritten Digit Recognition. *Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 1205-1208, doi: 10.1109/IIH-MSP.2009.145.

Xingye Qiao, L. Z. (2015). Distance-weighted Support Vector Machine. *International Press*, https://dx.doi.org/10.4310/SII.2015.v8.n3.a7.

Yizhen Xiang, J. W.-X. (2020). Schizophrenia Identification Using Multi-View Graph Measures of Functional Brain Networks. *Frontiers in Bioengineering and Biotechnology*, 10.3389/fbioe.2019.00479 .

Yueh-Yun Chi, K. E. (2013). Two-Step Hypothesis Testing When the Number of Variables Exceeds the Sample Size. *Communications in Statistics - Simulation and Computation*, 1113-1125. 10.1080/03610918.2012.659819.

Zhihong Zhang, E. R. (2011). *A Graph-Based Approach to Feature Selection.* Berlin: Springer.