National College of Ireland

Classification of Seizure Disorders using Machine Learning

MSc Reseach Project
Data Analytics

Dibyajyoti Bose
x15010732

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Dibyajyoti Bose
Student ID:	x15010732
Programme:	Data Analytics
Year:	2016
Module:	MSc Reseach Project
Lecturer:	Vikas Sahni
Submission Due Date:	22/08/2016
Project Title:	Classification of Seizure Disorders using Machine Learning
Word Count:	4984

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	12th September 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Abstract

The unpredictable nature of seizures poses a risk to patients with epilepsy.The aim of this study is to predict epileptic-seizures from EEG/MRI signals and patient-medical history by using machine-learning classifiers. The dataset is collected from NRS Medical Hospital, India, and comprises patients EEG/MRI findings, details of convulsion and type of seizure disorder. Although many studies have been performed to classify seizures based on para-clinical evidence obtained from conventional MRI and EEG, very little work had been hitherto done to characterize various types of seizures by taking into consideration associated clinical factors such as concomitant diseases, impaired-ADL, etiology of seizures, drug/alcohol abuse and family-history. This paper presents a supervised machine-learning approach that classifies seizure types (partial and generalized) using a dataset containing 150 records. In this paper, multiple machine learning techniques are applied and their performance is examined. The impact of feature engineering and parameter tweaking are explored with the objective of achieving superior predictive performance. Out of the various machine learning algorithms used, namely ANN,kNN and Random Forest, the the kNN classifier showed the best results (90% accuracy and 87% sensitivity). The model is evaluated on the testing data by using k-fold cross validation. The study will help clinicians to make diagnosis of epilepsy and initiate timely treatment; also, it will help a primary physician to decide the next step without the intervention of a trained neurologist.

Keywords- Seizure; machine learning; cross-validation; performance metrics

1 Introduction

The aim of the research is to classify Seizure disorder patients as high or low risk and partial or generalized using MRI/EEG scans and patient medical history from a high dimensional dataset using machine learning approach to aid early treatment, thus reducing morbidity in remote places where neurologists are not easily accessible. The risk of developing epilepsy is more common in elderly than the young, causing premature mortality. There are no anti-epileptic drugs that can cure seizure disorder. Seizure Analyses are primarily performed on EEG and MRI. An epileptic seizure can be classified into partial and generalized seizure. Partial seizure starts in one area of the brain while generalized seizure occurs in both hemispheres of the brain.

The paper presents and evaluates machine learning approach for constructing patient-specific classifiers that detect the onset of a seizure disorder through analysis of the EEG and MRI scan as well as previously unexplored but highly relevant clinical parameters. The challenge is to identify features to separate seizure disorder from other types of brain activity and implement an appropriate machine learning framework as proposed by Ghiassian et al [4]. Diagnosing seizure disorder in the early curable stages is very important and can save the life of a patient. Machine learning can help physicians in the time-consuming diagnosis tasks. In the Third World or remote places, such technology could help in early diagnosis especially where doctors are often not available or overworked. Machine learning paradigm will help physicians to make near-perfect diagnoses as well as help in cost reduction and optimization on a societal level, as it pertains to lack of preventative care and resource utilization. The machine learning algorithms were cross-validated to estimate how accurately a predictive model performs in practice on an

independent unknown data set to limit the problems of overfitting. This exciting development has huge potential and can be considered as a valuable tool to predict results before performing real laboratory experiments, thus saving labour, time and cost. This paper explores whether the combination of several classifiers can improve the results on classification of seizure disorder. The objective of this study is to identify the factors, which facilitate the diagnosis of epilepsy and classify the seizure types.
The number of clinical studies showing evidence of symptoms before seizures is very limited and thus there is little existing literature that systematically illustrates all the factors that can be taken into consideration for building predictive models in machine learning for seizure classification. Also there is limited training data for epileptic seizures which can be a major challenge.
We analyzed data over 6 months for 150 patients with partial and generalized seizures from NRS Medical Hospital, India with baseline clinical and demographic data. This study uses Random Forest (RF), K-Nearest Neighbors (kNN), and Artificial Neural Network (ANN) to predict the seizure types. The study aims to evaluate different algorithms used and compares the performance of each model. A comprehensive classication of seizures has been described based on feature extraction of high dimensional EEG signals. However, to the best of our knowledge, a classication of seizures by type (i.e. Generalized vs. Partial) using diagnostic scheme and the clinical characteristics have not been performed.
For this study, Cross Industry Standard Process for Data Mining. (CRISP-DM) methodology is used. This paper reports all details of conducted study in the following sections. Section I gives the introduction of the whole work. Section II explores the existing related works in the area, especially that which has been used in various contexts of this research. Section III describes the methodology used. Section IV provides the details of implementation of the modeling work. Section V includes the evaluation and Section VI gives the conclusion and future work.

2 Related Work

The study of electrical activity in the brain through the EEG is one of the most important tools that helps the doctors in making decisions about treatment and diagnosis of neurological disorders. The physician has to evaluate a number of factors from the current test results to previous decisions made on other patients to diagnose the epileptic seizure of a patient. In this crucial step, a physician may need an accurate tool that can help him to list the previous decisions made on the patient having the same factors. In recent years, a number of studies have been carried out to determine the risk factors associated with seizure recurrence. In a research article Phabphal et al[14] studied 278 patients older than 65 years with first seizure; they found that etiology and abnormal EEG features were the significant factors for seizure recurrence whereby mood disorder, sleep deprivation and stroke were the common co-morbidities. The multivariate regression analysis carried out for this study showed that age, sex and antiepileptic drugs were not significant factors for predicting seizure recurrence.

Andrews et al as well as Kotsopoulos et al [2, 9], conducted other studies. The one by Andrews et al was for a sample of 83 patients with uncontrolled seizures using discriminant analysis. It showed earlier onset age and higher seizure frequency were the two factors which predicted difficulty in controlling seizures. A similar research by Lacombe

et al [10] was conducted to characterize the distribution of seizure types on 104 horses presented for seizure disorders and to characterize the various types of seizures by identifying associated clinical factors. Seizures were then classified based on seizure types and accepted definitions in both human and animal epileptology. The univariate and multivariate logistic regression analyses were done and significant associations were found between seizure type and gender, frequency of seizures and presence of seizures during hospitalization. However, the findings suggested that seizure type was not significantly associated with etiology and that the clinical presentation was independent of the underlying disease. In a similar study conducted[12] on 792 patients (humans) at the time of their first diagnosis of epileptic seizures was undertaken to identify the factors presenting prognosis of epilepsy. The study revealed that the number of seizures in the early phase of epilepsy is the single most important predictive factor for both early and long-term remission of seizures.

The paper by Graves et al[5] discusses the risk factors associated with febrile seizures. This risk is increased in patients younger than 18 months and those with a lower fever, short duration of fever before seizure onset, or a family history of febrile seizures. Gotman proposed a system for automatic recognition of inter-ictal epileptic activity in prolonged EEG recordings using a spike and sharp wave recognition method. Extensions to this work are presented in Kofer and Gotman [8], Qu and Gotman [15], while recent works have focused on the use of functional magnetic resonance imaging (fMRI)and the correlation between cerebral hemodynamic changes and epileptic seizure events visible in EEG [11].

The results from the study by Shoeb at al [17] used a SVM classifier on EEG recordings from 24 subjects to distinguish between seizure and non-seizure with a classification accuracy of 96% for sensitivity with a false positive rate of 0.08 per hour using the CHB-MIT database. In a similar study five records were evaluated from CHB-MIT dataset containing 65 seizures by using a linear discriminant classifier[7]. The overall accuracy was 91.8%, the sensitivity was 83.6%, and specificity was 100%. [20] This paper used the Fuzzy Sugeno classifier which achieved 98.1% for overall accuracy. The FRE dataset research by Acharya et al [1] proposed a seizure detection system to train a neural network where 21 seizure records were used to train the classifier and 65 were used for testing with 94.9% accuracy The EEG can provide information to nd the patterns to the inter-ictal epileptiform discharges about the brain location where the abnormality is created and can be used to identify type of seizure disorder syndrome. The paper by Doescher et al [3] discusses the relationships between MRI and EEG findings.As part of an ongoing prospective study with 181 children (90 girls and 91 boys) with new-onset seizures, the association between EEG and MRI abnormalities are being explored. Of the 50 children with a normal EEG, however, 21 (42%) were found to have an abnormal MRI.The limitations of the study was the relatively small sample size but nevertheless the findings indicate that a normal EEG does not reliably predict a normal MRI with first seizures.

The findings indicate that EEG results are not the only good indicators of seizure disorder, and MRI results should be taken into consideration for further clinical findings.

Machine Learning provides algorithms and techniques that can help solving diagnostic and prediction of disease progression in a variety of medical domains. There is a lot of scope and ongoing work for applying machine learning in medical diagnosis in specialised diagnostic problem. Medical diagnostic reasoning is a very important application area of computer-based systems. This paper provides an overview of the reliability of the

decisions of classifiers for seizure disorder in the development of intelligent data analysis in medicine, from a machine learning perspective. The paper also discusses the approach to which machine learning algorithm can be used to classify seizure disorder; this can be especially difficult to diagnose accurately in clinical practice scenarios. This field of medical science is of utmost importance to researchers and data scientist to verify some unexplained phenomena in neurological disorders which is not yet approved by orthodox medical community, but could play an important role in the future in overall medical diagnosis and treatment.

Another study of inter-cranial EEG by Nigam et al [13], demonstrated that windowing-technique positively impacts classification performances.

A study of related risk factors focusing on early and late onset of seizures particularly in the aftermath of a stroke by Wang et al [18] demonstrated that late seizures are associated with cortical involvement, thus making them a high-risk factor in a post stroke seizure. However, the caveat is that this was a retrospective study and therefore, it's utility in predictive classification remains limited.

Senanayake et al [16] study considers etiological factors like meningitis, febrile seizures, genetic factors, head-injuries as a pre-disposing trigger for seizures, some of which have also been considered in the present work.

3 Methodology

This project is structured around the Cross Industry Standard Process for Data Mining (CRISP-DM) framework. CRISP-DM is a comprehensive data mining and business process model that breaks down the life-cycle of a data mining project into six phases thereby providing a complete blueprint for conducting a data mining project.

A. Development of the project

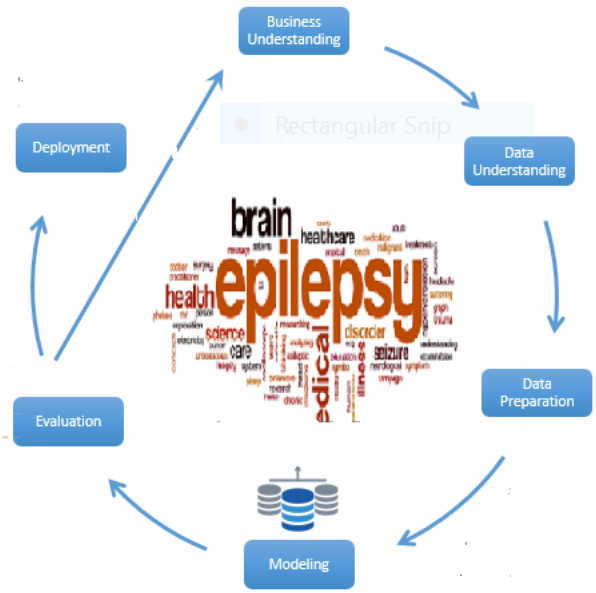


Figure 1: CRISP-DM Methodology

As shown in Figure 1, the development will follow the six main stages below, in accordance with the CRISP-DM regime we have presented. Epilepsy is a common neurological disorder and occurs when many nerve cells fire simultaneously in the brain - leading to seizures. It affects around 50 million people in the world. So application of machine learning to seizure prediction could help doctors save lives of the patients. By drawing insights at the similarities and differences between a large sample of patients clinical cases help doctors in their diagnosis of the seizure types which depends on several factors including the frequency and severity of the seizures and the person's age, overall health, and medical history. An accurate diagnosis of the type of epilepsy is also critical to choosing the best treatment. This will be beneficial for both medical practitioners and patients won't have to undergo unnecessary treatments that do not lead to a cure.
We formulated the prediction task by translating the business questions to data mining goals and specified the data mining problem type (classification, prediction, clustering, etc.). Since the project deals with the objective of predicting if the seizure is partial or generalized based on 16 extracted features, including Etiology of seizure, Concomitant diseases, EEG, MRI, patient medical history etc., the problem is defined as a classification problem. Next an initial assessment of tools and techniques to be used is evaluated. The data is acquired from the NRS Medical Hospital repository and initial data preparation tasks were carried out. This task included the distribution of key attributes (target attribute for prediction task) relationships and statistical analysis. Since the success of machine learning depends on how the data is presented the process of feature engineering is applied based on domain knowledge and dataset attributes to transform the raw data into features that can better represent the underlying problem to the predictive models, resulting in improved model performance on unseen data. The details of feature engineering process are discussed in the implementation part.
Next, appropriate bar charts, histogram, and correlation were explored to indicate the data characteristics that suggest interesting insights from the data subsets. The data quality was addressed, like imputing missing values and rectifying errors in dataset to further proceed with the next step for exploration. The dataset from the data preparation phase was used for modelling the major analysis work of the project. This task also includes transformation of values for existing attributes. We first attempted PCA for dimensionality reduction but it did not provide any improvements for our model. The correlation between the attributes are limited and the original feature set is well-designed. As the first step in modelling a specific modelling technique was studied since we have already selected a tool (R) during the business understanding phase. We implemented three different machine learning algorithms on the dataset, ranging from Random Forest, K-Nearest Neighbors and Artificial Neural Network. To overcome the problem of overfitting we have performed 5-fold cross validation and build the models on the training data and tested on the test data as suggested by Wolpert [19]. We plotted the variable importance from RF and concluded which variables have a significant importance in the model. The model evaluation was done by various performance measures that will make predictions by the trained model on the test dataset as proposed in Hothorn's work [6].
For the purpose of our model evaluation, metrics of prediction performance are used including Accuracy, Recall (Sensitivity), and Specificity. In the end of the project, the final report is prepared with a comprehensive presentation of our data mining results and any future scope of improvement. The performances of all the models are evaluated and compared, with the aim of understanding which method provides the best performance. It should be brought into knowledge that different methods may be favorable over

one another in different business applications, depending on which performance metrics should be focused in each business scenario. But for the scope of this project we have focused more on accuracy of the model as discussed with the domain experts.

4 Implementation

A. Preparation and splitting of the data

The original data set has 150 observations with 16 fields including patient name and age. A common challenge with nominal categorical variables is that it may decrease the model performance. So we have used dummy encoding to represent one level of categorical variable. Presence of a level is represented by 1 and absence is represented by 0. For every level present, one dummy variable is created. The original raw data that was collected from the hospital was in below excel format:

Age	Gender	Concomitant diseases	Urban/Rural	Onsetage	Impaired ADL	Frequency (/month)	Duration (minute)	last attack days back	Family history of Seizure	Febrile Seizure	Sleep deprivation	Drug/ alcohol abuse	EEG	CT Brain and MRI	Etiology of Sz	Type of Sz	Abnormal discharge
16	M	Mood Dis	Urban	14yrs 3 months	Y	0.75	0.75	25	N	N	Yes	N	focal Dysrhythmia	Rt focal isch change	HIE	Gen SZ	Slow Wave-40 Millisecond s
14	F	Headache	Urban	12yrs 5 months	N	1-2	15	170	N	N	No	N	Poly spike	Normal	Primary	Gen SZ	Multiple Spike- 35 to 45 MS
4	F	Nil	Urban	3yrs 2months	N	00-Jan	0.5	5	N	y	No	N	Gen Slow wave	Normal	Primary	Gen SZ	45 to 55 MS

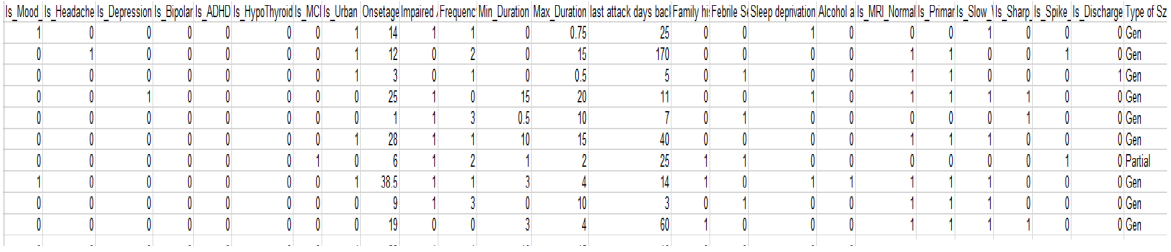
Figure 2: Seizure Disorder NRS Hospital Data

Attributes	Description
Name	Name of the patient
Age	Age of the patient
Gender	Male or Female
Urban/Rural	Urban or Rural
Concomitant Disease	Associated Disease along with the Present Disease
Onsetage	Age at which seizure happened
Frequency (/month)	Number of times seizure occurs n months
Impaired ADL	Impairment of daily activity of any Individual
Duration (/minute)	Duration of seizure
Last attack days back	Last attack of seizure
Family History of Seizure	History of Seizure
Febrile Seizure	Febrile convulsion, associated with a high body temperature
Sleep Deprivation	Sleep
Drug/Alcohol Abuse	Chronic alcohol abuse in patients
EEG	Slow, Sharp and Spike wave
CT Brain and MRI	Identification and localization of brain lesion
Etiology of Seizure	Primary/Secondary
Abnormal Discharge	Electrical discharge of brain cells all at once
Type of Seizures	Partial/Generalized

Figure 3: Seizure attributes Description

The raw data was collected and domain knowledge was gathered after discussing with the neurologists to identify which features might be relevant. Accordingly feature engineering was applied to reduce the complexity of the data and to yield high performance. Dummy variables, also called Indicator Variables are useful to take categorical variable as a predictor in statistical models. Categorical variable can take values 0 and 1.So we converted all the categorical variables into numerical variables. We classified the EEG signals wavelength as Slow, Sharp and Spike. The sharp wave is the transient, clearly distinguishable from background activity with pointed peak at conventional paper speeds and a duration of 70 milliseconds to 200 milliseconds (MS). The spike wave is same as sharp wave but with duration of 20 MS to less than 70 MS. Poly Spike is the multiple Spike discharge. The slow wave is classically same as spike but of higher amplitude than the spike discharge. The CT Brain and MRI was coded as normal or not and concomitant diseases were encoded with 7 binary dummy variables to indicate presence or not. Similarly, the etiology of seizure was coded with 2 binary variables as primary or secondary to indicate 0 or 1. Since the duration of the attack was coded as range of values so we created 2 separate variables denoting the minimum and maximum duration which the seizure lasted. Thus all the variables were encoded in binary and numerical values to

boost the performance when applying machine learning models. The patient demographics like Name, Age, Gender was removed from the dataset and feature engineering was done as illustrated below: -



Is_Mood	Is_Headache	Is_Depression	Is_Bipolar	Is_ADHD	Is_HypoThyroid	Is_MCI	Is_Urban	Onsetage	Impaired	Frequenc	Min_Duration	Max_Duration	last attack days bac	Family hi	Febrile Si	Sleep deprivation	Alcohol a	Is_MRI_Normal	Is_Primar	Is_Slow_I	Is_Sharp_	Is_Spike_	Is_Discharge	Type of Sz
1	0	0	0	0	0	0	1	14	1	1	0	0.75	25	0	0	1	0	0	0	1	0	0	0	Gen
0	1	0	0	0	0	0	1	12	0	2	0	15	170	0	0	0	0	1	1	0	0	1	0	Gen
0	0	0	0	0	0	0	1	3	0	1	0	0.5	5	0	1	0	0	1	1	0	0	0	1	Gen
0	0	1	0	0	0	0	0	25	1	0	15	20	11	0	0	1	0	1	1	1	1	0	0	Gen
0	0	0	0	0	0	0	0	1	1	3	0.5	10	7	0	1	0	0	0	0	0	1	0	0	Gen
0	0	0	0	0	0	0	1	28	1	1	10	15	40	0	0	0	0	1	1	1	0	0	0	Gen
0	0	0	0	0	0	1	0	6	1	2	1	2	25	1	1	0	0	0	0	0	0	1	0	Partial
1	0	0	0	0	0	0	1	38.5	1	1	3	4	14	1	0	1	1	1	1	1	0	0	0	Gen
0	0	0	0	0	0	0	0	9	1	3	0	10	3	0	1	0	0	1	1	1	0	0	0	Gen
0	0	0	0	0	0	0	0	19	0	0	3	4	60	1	0	0	0	1	1	1	1	0	0	Gen

Figure 4: Feature Engineering of Attributes

For the ease of simplicity, the seizures were classified into Partial and generalized though the original excel received from the hospital categorized the partial seizure types into simple and complex partial seizure. But for the purpose of this study we have kept it simple in terms of 2 categories as partial and generalized seizures. These two are our target variables on which we applied the classifiers to predict the seizure types.
This study uses a K-fold cross validation and the data is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times. Since the dataset is small and computation time was less K-fold cross validation is used to overcome the problem of overfitting.
For the purpose of this study we set the value of K as 5 and estimated the confusion matrix for each of the 5 folds. The rationale of using K-fold cross validation is since the test set is very small, so there will be a lot of variation in the performance estimate for different samples of data, or for different partitions of the data to form training and test sets when using random train/test spilt. K-fold validation reduces this variance by averaging over k different partitions, so the performance estimate is less sensitive to the partitioning of the data. By rotating through the partitions of data for training v/s CV, it helps ensure that the resulting model end up selecting is well rounded and not particularly biased towards the particulars of how the training data was split into training and CV datasets.

A. Development of the project

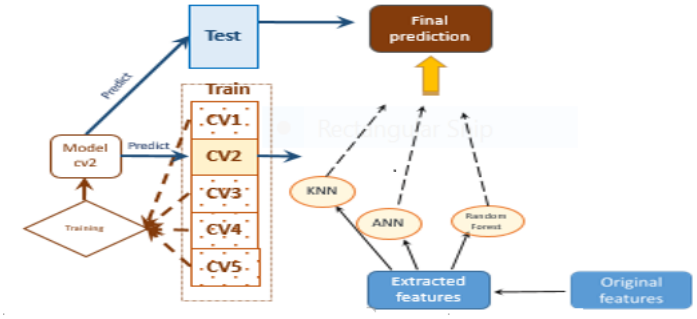


Figure 5: The work flow of modelling

B.Implementation of machine learning models

1) K-nearest neighbor (KNN) method: the KNN model is performed in R with the class library. KNN requires the input data be normalized in range of values. The values of the 24 extracted features from the original data are scaled into the range between 0 and 1. The scaled data are then used for KNN modelling. Preliminary tests are performed to tune the model and find the optimal value of k=9 for this study, and the same k value is used for every run of KNN modelling for all the 5-fold cross validation.

2) Random forest (RF): RF is performed using the random forest package in R, with mtry=10 and ntrees=200 selected from preliminary tests. The featured engineer attributes are used as the input for this model. The variable importance from RF is plotted.

3) Artificial neural network (ANN): ANN is performed with the neural net package in R. It is good practice to normalize the data before training a neural network because depending on dataset some times the algorithm will not converge before the number of maximum iterations allowed. With the activation function as sigmoid, and one hidden layer with 5 neurons, learning rate 0.1, the ANN is trained on the normalized data. By setting a unique seed value before splitting the train set into five folds cross-validation (CV) sets, or say five folds, the same row indexing for the five CV sets are used for all the different models. The accuracy, sensitivity and specificity are calculated for each of the five folds. CV allows the entire dataset to train and test one model/method, while being able to have a reasonable idea of how well it will generalize so the mean accuracy and standard deviation from 5-fold CV is estimated. By taking the mean we get an estimate of our out-of-sample accuracy of both the mean and the variation of the performance metrics of each model, with the aim of evaluating the average performance and the stability of different models. D. Performance metrics evaluation of the models The prediction results of the classifiers for each fold of the CV are presented in a confusion matrix and the performance metrics of each of the models are compared against to evaluate the best classifier.

5 Evaluation

A.Data Exploration results

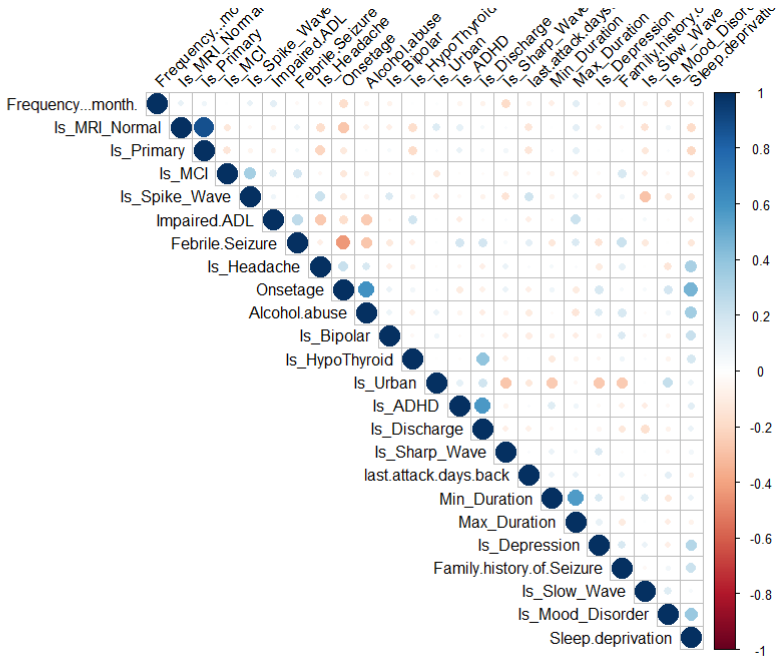


Figure 6: Correlation Matrix

At this stage we explored the variables one by one to understand the central tendency and spread of the variable. These are measured using various statistical metrics visualization methods as show below. The correlation among the variables were studied but as can be seen from the figure below there is hardly any correlation among the variables, the highest correlation being .87 . We tried PCA to reduce the dimensionality but PCA did not give any satisfactory results. Correlation matrix is used to highlight the most correlated variables in a data table. In this plot, correlation coefficients is colored according to the value. and degree of association between variables. The blue colored dot represents positive correlation and red colored dot represents negative correlation.The R corrplot package is used here. As can be seen from the histogram last attack days and duration of attack attributes are highly skewed. The onset age is more or less normally distributed.

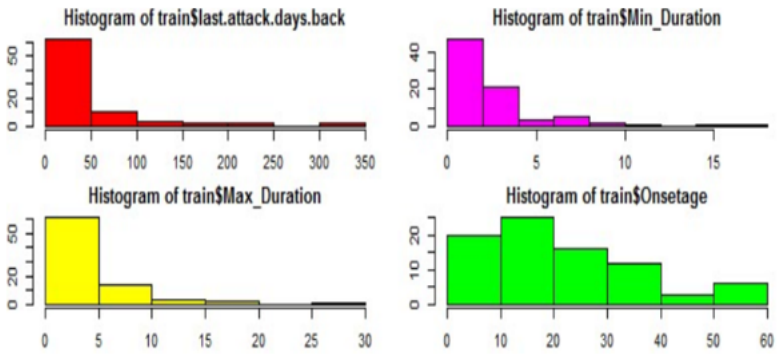


Figure 7: Histogram

B.Modelling results

The Random Forest algorithm in R works on the Gini Index which is same as entropy of information gain. It calculates the Gini based on probability of success and failure for each nodes and sub-nodes and take the weighted Gini for each nodes. The node split takes place based on Gini score. Since Random Forest works on growing multiple trees and each time the attributes are selected randomly the forest chooses the classification having the most votes over all the trees in the forest. The rationale of choosing Random Forest is that in general RF is less biased than the decision tree model. So we expect better accuracy when applying the model. Also random forest gives the variable importance as illustrated in the table below. The information gain in the attributes depend mostly on

Table 1: Variable Importance in Random Forest

Attributes	MeanDecreaseGini
Is Primary	7.35
Is MRI Normal	6.37
last.attack.days.back	2.49
Onsetage	2.37

the Etiology of seizure as primary or secondary, MRI and CT brain normal or not, the last attack days back and the seizure onset-age. The Random Forest gave the second highest in terms of accuracy and highest in terms of recall sensitivity among the models.

Table 2: 5-Fold CV Performance metrics in Random Forest

CV	Accuracy	Sensitivity	Specificity
Fold 1	0.84	0.90	0.73
Fold 2	0.81	0.86	0.72
Fold 3	0.85	0.92	0.75
Fold 4	0.86	0.92	0.76
Fold 5	0.86	0.92	0.75
Mean	0.84	0.90	0.74

The average accuracy across the 5 folds is almost same with less variation. Since all K repeats give nearly the same performance we can conclude there is less chance of the model being over fitted.

ANN provide slightly worse performance with greater variation among different ways of five-fold splitting of the training data set with the last fold giving 100% accuracy which is not good for the model. But the mean accuracy of the 5-fold CV is almost same as Random Forest. So there may be a possibility of overfitting while applying ANN model for this dataset.

K nearest neighbours is the widely used classification technique which is easy to interpret and low calculation time with good predictive power. To choose the optimum value of K is a challenge. So we choose the value of K based on the training error rate and the validation error rate as two parameters we to access on different K-value. The error rate initially decreases and reaches some minima. After that point it then increases

Table 3: 5-Fold CV Performance metrics in ANN

CV	Accuracy	Sensitivity	Specificity
Fold 1	0.75	0.75	0.75
Fold 2	0.87	0.87	0.87
Fold 3	0.68	0.75	0.62
Fold 4	0.87	0.90	0.80
Fold 5	1	1	1
Mean	0.83	0.85	0.81

with increasing value of k. To get the optimum value of K the training and test data is segregated from the initial dataset and then the validation error curve is plotted for the optimal value of K. This value of K is used across the 5-fold cross validation. In our dataset we choose the value of K as 9. The selection of K determines how well the model can be built to generalize the results of kNN algorithm and determine the efficacy of the model. A large value of K may reduce the variance due to noisy data but develop a bias due to which the learner tends to ignore the patterns which might give useful insights. Since kNN works on calculating Euclidian distance of the data points so we normalized the numeric data since the scale used for the values for each variable might be different. The original variables have been feature engineered to set as numeric values and normalized before applying the kNN algorithm to transform it to a common scale. By tweaking the value of k we calculated the accuracy of the model and finally set k as 9 across the 5-fold cross validation to get the increased accuracy of the model.

Table 4: 5-Fold CV Performance metrics in kNN

CV	Accuracy	Sensitivity	Specificity
Fold 1	0.81	0.75	1
Fold 2	0.93	0.87	1
Fold 3	0.75	0.75	0.75
Fold 4	1	1	1
Fold 5	1	1	1
Mean	0.90	0.87	0.95

We get the highest accuracy from kNN model though the last 2 folds may suffer from overfitting since accuracy for last 2 folds came around 100%. But the mean accuracy achieved in kNN is the highest among the models.

6 Conclusion and Future Work

The dataset for this study was limited to 150 patients with seizures, but it would have been desirable to run the same test on a significantly larger dataset. However, it needs to be noted that reliable and well maintained data on a narrow focussed clinical area such as epilepsy, that too, from a machine learning point of view, is extremely difficult to source due to issues such as patient confidentiality, unless the study itself has been initiated or

conducted by some major hospital or health board on its own accord.
How the model will perform on larger dataset can be a part of an ongoing or future study to ratify the outcomes from this particular study. Whether or not the accuracy of 85% is achievable on much larger datasets remains to be tested.Based on the discussion above, two tasks may be carried out in future studies.First, more machine learning methods including packed ensemble methods may be included as the level-0 models, in order to see if stacked generalization can push further in the improvement of the model ensembles. Second, more complex algorithms may be tested as the level-1 model, to understand how the non-linearity in the level-0 predictions can be utilized for better prediction performance.This study can be extended further for any future work to classify other complex type of seizures based on the clinical factors which are a major determinant of epileptic seizures.

Acknowledgements

The following thesis, whilst an individual piece of work, benefited from the insights and able direction of my thesis supervisor Mr.Vikas Sahni. I wish to thank him for his continuous support, knowledge, advice and encouragement throughout the entire process. Next I wish to acknowledge gratefully Dr.Joydeep Mukherjee and Dr.Biman Bose's kind permission for providing the permission to use data set for this study as well as their clinical insights.Finally I am grateful for the support of my colleague Mr. Sanjay Shende for his unflinching and patient help with the proof reading and insightful inputs.

References

[1] U Rajendra Acharya, Filippo Molinari, S Vinitha Sree, Subhagata Chattopadhyay, Kwan-Hoong Ng, and Jasjit S Suri. Automated diagnosis of epileptic eeg using entropies. *Biomedical Signal Processing and Control*, 7(4):401–408, 2012.

[2] Donna J Andrews and Warren H Schonfeld. Predictive factors for controlling seizures using a behavioural approach. *Seizure*, 1(2):111–116, 1992.

[3] Jason S Doescher, Beverly S Musick, David W Dunn, Andrew J Kalnin, John C Egelhoff, Anna Weber Byars, Vincent P Mathews, Joan K Austin, et al. Magnetic resonance imaging (mri) and electroencephalographic (eeg) findings in a cohort of normal children with newly diagnosed seizures. *Journal of child neurology*, 21(6):490–495, 2006.

[4] Sina Ghiassian, Russell Greiner, Ping Jin, and M Brown. Learning to classify psychiatric disorders based on fmr images: Autism vs healthy and adhd vs healthy. In *Proceedings of 3rd NIPS Workshop on Machine Learning and Interpretation in NeuroImaging*, 2013.

[5] Reese C Graves, Karen Oehler, Leslie E Tingle, et al. Febrile seizures: risks, evaluation, and prognosis. *Am Fam Physician*, 85(2):149–53, 2012.

[6] Torsten Hothorn and Berthold Lausen. Bundling classifiers by bagging trees. *Computational Statistics & Data Analysis*, 49(4):1068–1078, 2005.

[7] Yusuf Uzzaman Khan, Nidal Rafiuddin, and Omar Farooq. Automated seizure detection in scalp eeg using multiple wavelet scales. In *Signal Processing, Computing and Control (ISPCC), 2012 IEEE International Conference on*, pages 1–5. IEEE, 2012.

[8] DJ Koffler and J Gotman. Automatic detection of spike-and-wave bursts in ambulatory eeg recordings. *Electroencephalography and clinical Neurophysiology*, 61(2):165–180, 1985.

[9] Irene Kotsopoulos, Marc de Krom, Fons Kessels, Jan Lodder, Jaap Troost, Mascha Twellaar, Tiny van Merode, and André Knottnerus. Incidence of epilepsy and predictive factors of epileptic and non-epileptic seizures. *Seizure*, 14(3):175–182, 2005.

[10] VA Lacombe, M Mayes, S Mosseri, SM Reed, and TH Ou. Distribution and predictive factors of seizure types in 104 cases. *Equine veterinary journal*, 46(4):441–445, 2014.

[11] R Lopes, Jean-Marc Lina, F Fahoum, and J Gotman. Detection of epileptic activity in fmri without recording the eeg. *NeuroImage*, 60(3):1867–1879, 2012.

[12] BK MacDonald, AL Johnson, DM Goodridge, OC Cockerell, JWAS Sander, SD Shorvon, et al. Factors predicting prognosis of epilepsy after presentation with seizures. *Annals of neurology*, 48(6):833–841, 2000.

[13] Lavi Nigam, Deepthi Karnam, Sreerama K Murthy, Petro Fedorovych, Vasu Kalidindi, et al. Machine learning for seizure prediction: A revamped approach. In *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, pages 1159–1164. IEEE, 2015.

[14] Kanitpong Phabphal, Alan Geater, Kitti Limapichat, Pornchai Sathirapanya, and Suwanna Setthawatcharawanich. Risk factors of recurrent seizure, co-morbidities, and mortality in new onset seizure in elderly. *Seizure*, 22(7):577–580, 2013.

[15] Hao Qu and Jean Gotman. Improvement in seizure detection performance by automatic adaptation to the eeg of each patient. *Electroencephalography and clinical Neurophysiology*, 86(2):79–87, 1993.

[16] N Senanayake and Gustavo C Román. Epidemiology of epilepsy in developing countries. *Bulletin of the World Health Organization*, 71(2):247, 1993.

[17] Ali H Shoeb and John V Guttag. Application of machine learning to epileptic seizure detection. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 975–982, 2010.

[18] Guoqing Wang, Haiyan Jia, Chunfu Chen, Senyang Lang, Xinfeng Liu, Cheng Xia, Yan Sun, and Jun Zhang. Analysis of risk factors for first seizure after stroke in chinese patients. *BioMed research international*, 2013, 2013.

[19] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[20] Qi Yuan, Weidong Zhou, Shufang Li, and Dongmei Cai. Epileptic eeg classification based on extreme learning machine and nonlinear features. *Epilepsy research*, 96(1):29–38, 2011.