DEVELOPMENT OF ARITFICIAL INTELLIGENCE BASED APPLICATION FOR EARLY DIAGNOSIS OF ALZHEIMER'S DISEASE

A

THESIS

Submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

In

BIOINFORMATICS

Under the supervision of

Dr. Tiratha Raj Singh

(AssociateProfesor)

by

Muskan Kapoor (171501)
Mehak Kapoor (171519)

to

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN – 173234

HIMACHAL PRADESH,

INDIA May – 2021

STUDENT'S DECLARATION

We hereby declare that the work presented in the Project report entitled **"Development of AI based application for early diagnosis of AD"** submitted for fulfillment of the requirements for the degree of Bachelor of Technology in Bioinformatics at **Jaypee University of Information Technology, Waknaghat,** is an authentic record of our work carried out under the supervision of **Dr. Tiratha Raj Singh.** This work has not been submitted elsewhere for the reward of any other degree/diploma. We are fully responsible for the contents of our project report.

CERTIFICATE

This is to certify that the work which is being presented in the project report titled **"Development of AI based application for early diagnosis of AD"** in fulfillment of the requirements for the award of the degree of Bachelor of Technology in Bioinformatics submitted to the Department of Biotechnology and Bioinformatics, **Jaypee University of Information Technology, Waknaghat,** is an authentic record of work carried out by **Muskan Kapoor, 171501 and Mehak Kapoor, 171519** during a period from August, 2020 to May, 2021 under the supervision of **Dr. Tiratha Raj Singh** Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat.

The above statement made is correct to the best of our knowledge.

(Dr. Tiratha Raj Singh)
Date: 25 May 2021

ACKNOWLEDGEMENT

Foremost, we would like to express our sincere gratitude to our advisor **Dr. Tiratha Raj Singh (Associate Professor)** for the continuous support of our project study, for his patience, motivation, enthusiasm, and immense knowledge. His guidance has helped us in all the time of this study and writing of this report.

We are also very thankful to all the faculty members of Biotechnology and Bioinformatics Department, Jaypee University of Information and Technology for constant encouragement during the project. We also take this opportunity to thank Rohit Shukla (PhD Scholar) who have helped and guided us in our project work.

Last but not the least we would like to thank our parents, who taught us the value of hard work by their own example.

Muskan Kapoor *Muskankapoor*
Mehak Kapoor *MehakKapoor*
May - 2021

ABSTRACT

Alzheimer's disease is currently at the forefront of scientific research. The global epidemic burden of AD is expected to exceed $2 trillion by 2030, entailing early diagnosis. However, research revealed that one out of every three cases of dementia can be avoided if detected early enough before profound brain loss happens.

AD is a gradual, irreversible brain disease that deteriorates a patient's memory, cognitive functions and shrinks the brain's size, eventually leading to death. However, for the most part, symptoms do not escalate. In reality, the signs can be transient or reversible in some cases. Based on a person's early symptoms, it can be impossible to predict an Alzheimer's diagnosis. This makes it challenging to diagnose AD early, as subjects who are affected have the greatest probability of benefiting from the few interventions and drug therapies available.

Faced with these obstacles, a significant research work is focusing on developing technologies that can detect cases earlier and more accurately, particularly in the elderly who are undergoing some cognitive impairment. Thus, diagnosis of AD is mainly based on clinical evaluation as well as cognitive assessment using neuropsychological tests which might assist in advancing and assessing peculiar medication. Early detection and classification of divergent phases of AD are done through ML with AI techniques that can be applied to EHR to provide accurate and comprehensive diagnosis to improve the quality and productivity of healthcare. ML models that use numerous optimization and probabilistic approaches may be used to make this diagnosis.

In this article, a multi-modality procedure is followed. The data is acquired from ADNI-TADPOLE grand challenge and UNIPROT inclusive of cross-sectional, longitudinal (baseline) datasets and sequence datasets. The numerical data as a whole includes clinical scans (MRI, PET, FDG), cognitive scores (MMSE, ADAS-Cog, ADAS-11, ADAS-13) demographic attributes (Age, Race) and sequence of AD/non AD-related genes. The four diverse classifiers employed for classification and prediction of AD are RF, SVM, ANN, RF+SVM. The primary goal of our study is to identify the most relevant biomarkers and features that can contribute to reliable, accurate, effective and timely diagnosis of AD. Thus, to escalate patient’s quality of life and forbid high medical expenses, an automated, persistent, and unobtrusive pre - clinical detection system will be enforced.

Keywords: Alzheimer's Disease, Machine Learning, Bioinformatics, ADNI, Biomarkers.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS & ABBREVIATIONS

Acronym	Description
AD	Alzheimer's Disease
EHR	Electronic Health Record
ML	Machine Learning
AI	Artificial Intelligence
ADNI	Alzheimer's Disease Neuroimaging Initiative
TADPOLE	The Alzheimer's Disease Prediction of Longitudinal Evolution
fMRI	Functional Magnetic Resonance Imaging
PET	Positron Emission Technology
FDG	Flurodeoxyglucose
MMSE	Mini-Mental State Examination
ADAS	Alzheimer's Disease Assessment Scale
RF	Random Forest
SVM	Support Vector Machine
ANN	Artificial Neural Network
CDR	Clinical Dementia Rating
CSF	Cerebrospinal Fluid
CN	Cognitively Normal
MCI	Mild Cognitive Impairment
LMCI	Late Mild Cognitive Impairment
EMCI	Early Mild Cognitive Impairment
OASIS	Open Access series of Imaging Studies
AIBL	Australian Imaging Biomarker and Lifestyle Flagship study of aging
GA	Genetic Algorithm
AdaBoost	Adaptive Boosting

NCBI	National Centre for Biotechnology Information
APP	Amyloid Precursor Protein
PS-1	Preselin-1
PS-2	Preselin-2
KNN	K Nearest Neighbor
DT	Decision Tree
CVD	Cardiovascular Disease
SMC	Subjective Memory Complaint
CD-HIT	Cluster database at High Identity with Tolerance
RBF	Radial Basis Kernel
MLP	Multi-Layer Perceptron
ROC	Receiver Operating Curve
AUC	Area under ROC curve
FPR	False Positive Rate
TPR	True Positive Rate
NL	Normal individuals
PMC	Pubmed Central
TF	Transcription Factor
SNP	Single Nucleotide Polymorphisms
NFT	Neuro Fibiliary Tangles

CHAPTER 1
INTRODUCTION

1.1 General

Alzheimer's disease is a progressive neurodegenerative disorder that demolishes cognitive dysfunction. Recent research derived that AD is the third leading cause of death. According to recent estimates in 2021 about 6.2 million Americans aged 65 and older are already suffering from AD in both developed and developing nations [1]. It is also reported that the incidence of AD will increase to every 33 seconds by 2050 and the rate at which AD occurs currently is every 65 seconds in the USA [2].

Unfortunately, there is currently no treatment for AD. Though there are certain drugs available but, medications only help patient's conditions for a short time. The most recently accepted treatment is just a mixture of two previously approved anti-medications, Alzheimer's Donepezil and Memantine. Despite significant attempts to discover a treatment for AD, clinical trials for AD treatments have a 99.6% failure rate.

1.2 History

Alzheimer's syndrome was first identified in 1906 by German psychiatrist Dr. Alois Alzheimer, who described "a peculiar disease" characterized by severe cognitive impairment and subtle brain changes. [2]. The disease is an inoperable, irreversible and progressive neurodegenerative disease that disrupts neurons and their associations in areas of the brain that are important in memory processing and this includes the cerebral cortex, entorhinal cortex, hippocampus and lobes. Changes in cognitive skills in Alzheimer's patients often begin slowly and accelerate with time. As a result, affects regions of the cerebral cortex involved in grammar, reasoning, and social behavior. In due course, damaging supplementary parts of the brain and thus proving to be an inevitable disease. Some of the risk factors associated with AD include family history, cardiovascular disease, Down's syndrome, head injuries, age and sedentary lifestyle [3].

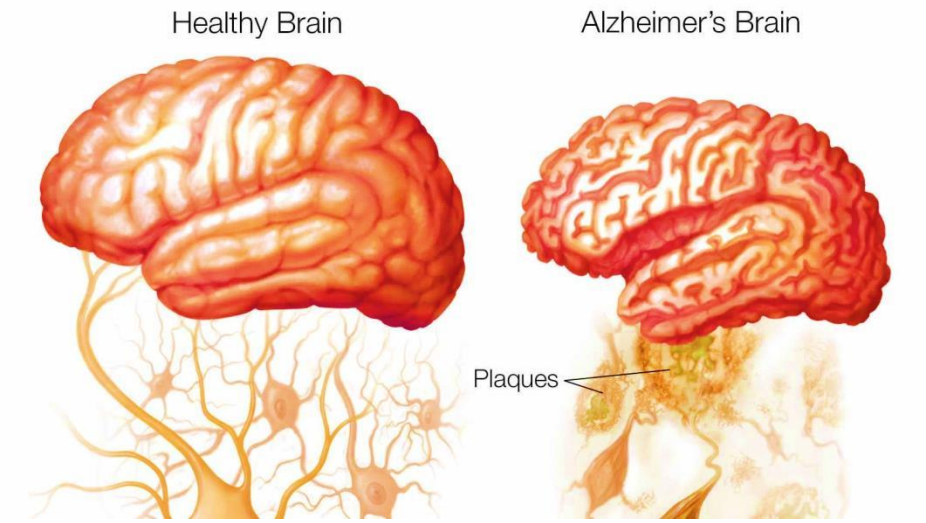

Figure 1.1: Representation of healthy and AD brain [4]

1.3 Diagnosis

Diagnosis of AD at primary onset may delay the progression of the disease. The diagnosis primarily relies on cognitive assessment and clinical evaluation using a neuropsychological test which includes Mini-Mental State Examinations(MMSE),

Alzheimer's disease Assessment Scale Cognitive (ADAS-11, ADAS-13), and Clinical Dementia Rating (CDR). The radical brain image acquisition techniques such as Computerized Tomography (CT), Functional Magnetic Resonance Imaging (fMRI), Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI) are used as clinical biomarkers that can effectively diagnose breakthrough of AD [5]. The genetic biomarkers and risk factors play a vital role as well.

PET imaging with precise tracers and cerebrospinal fluid biomarkers can be used to test in vivo the two pathological hallmarks of AD, amyloid beta plaques and neurofibrillary tau angles, which signify the existence of rare abnormal protein deposits [6]. Based on the gradual progression of AD, it is divided into different stages such as Dementia, MCI (mild cognitive impairment) and CN (cognitively normal). This distinction is important because patients with different stages of AD need different treatments and the same drug cannot be used by all of them [7]. Hence, classification of different stages is crucial for the achievement of the goal that it will increase the patient's quality of life by allowing for appropriate symptom care.

1.4 Machine Learning Techniques used in Alzheimer's disease

Machine Learning (ML) is computational and non-traditional approach that aims to organize, interpret and analyze different kinds of datasets. It is a subcategory of artificial intelligence that includes a number of techniques for making mathematical, probabilistic decisions based on prior knowledge. It classifies new events and predicts new trends based on previous learning. Features can be extracted and categorized without involving any experts thanks to the advancement of ML models.

The use of contemporary computing technology and resources is becoming a boon to new trends in healthcare and diagnosis. Electronic Health Record (EHR) is setting a gauge to record patient's data electronically through the replacement of conventional methods that comprises a collection of data in paper-based form [7]. Although diagnosing AD at an initial stage remains a challenging task. The trend of ML and EHR to anticipate and visualize illness is particularly prevalent toward predictive and personalized prescribing. ML algorithms that use numerous optimization and probabilistic techniques may be used to make this diagnosis.

The potential of ML techniques to acquire relevant patterns within data and provide automated classifications and forecasts makes them especially appealing. Wide databases of patients with multimodal data have been made public in recent years. The Alzheimer's Disease Neuroimaging Initiative (ADNI) is the most well-known publically accessible software, however there are other widely accessible databases, such as the Open Access Series of Imaging Studies and the Australian Imaging, Biomarker, and Lifestyle Flagship Study of Ageing. This would significantly accelerate the advancement of ML approaches to aid in the diagnosis and prognosis of AD. [6]

1.5 Single and Multiple Modality Approach

Despite substantial studies and advancements in clinical practice, only about half of AD patients had their pathology and disease development correctly diagnosed relying on their clinical symptoms. Clinical symptoms, demographic characteristics, neurological assessments and ratings, such as MMSE scores, are supplemented by imaging, molecular, and protein biomarkers in the study of AD. The majority of these studies use single-modality data to classify biomarkers, which limits a holistic evaluation of AD disease progression.

While the use of various single biomarkers yields positive outcomes, they are intended to characterize group distinctions rather than to classify individuals [8] developed a mechanism for distinguishing between stable and AD participants by combining the three biomarkers for Alzheimer's disease diagnosis, such as MRI, PET, and CSF. [9]

1.6 Stages of Progression of Disease

The four stages of AD are pre-dementia (up to 20 years), early, moderate (2-10 years) and advanced (1-5 years). A description of stages is explained in the context below.

Pre-Dementia Stage: It is marked by symptoms such as memory loss which are often mistaken as aging or stress. Early-stage Alzheimer's disease is characterized primarily by encoding and retrieval problems, which result in acute memory impairment and a reduced

ability to acquire new information. The cognitive symptoms include subtle episodic memory loss and executive abilities. The patients might face behaviors such as irritability, apathy and dysphoria.

Early and Moderate Stage: The increasing impairment of learning and memory eventually leads to the early stage. The stage feature limited vocabulary and decreased word fluency. Moreover, speech and motor skills are progressively lost. The cognitive symptoms include episodic memory loss, misplacing items, disorientation and decreased executive abilities such as problem-solving and decision making.

Advanced Stage: Mental function continues to deteriorate in the late stages of Alzheimer's disease, known as advanced-stage related to Alzheimer's disease. The disease has a rising influence on mobility and physical abilities. The major challenge is loss of "self". The patients might face behaviors such as agitation and insomnia.

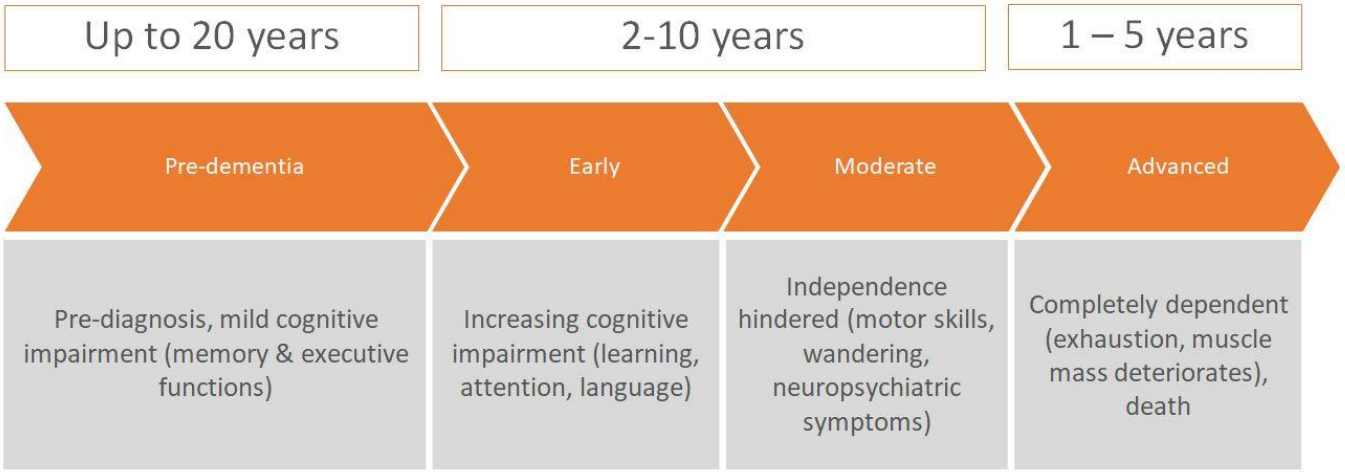

Figure 1.2: Different Stages of progression of AD.

1.7 Objectives

Prognosis of AD on available data is a cutting-edge challenge. The thesis presents the performance and analysis of AD through a multi-dimensional format which includes neuroimaging and sequence data. The main objectives of this study are:

I. It focuses on the prediction of progression of subjects which are at distinct stages of AD by applying machine learning techniques on the ADNI-TADPOLE challenge data

II. It uses ADNI-TADPOLE data with baseline as well as non-baseline features achieving considerable accuracies.

III. We used multi-modal approach which includes both ADNI-TADPOLE data (clinical, cognitive, and demographic) as well as sequence data (genetic).

IV. Target was to identify the most distinguishing and imperative biomarkers in both types of datasets used. The available data is explored in the subsections.

1.8 Research Structure

This research study is distributed into five chapters. A short description of each chapter is provided here.

Chapter 1 Introduction: This chapter provides insights about the topic, discusses the objective, provides the purpose and scope of this study, and defines the AD, stages of progression and diagnosis.

Chapter 2 Literature Review: This chapter can be pondered as the back bone of this report. It provides the facts and data gathered by former researchers on the topic of an amalgam of AD and AI published in diverse reputed journals papers, books and articles. The findings of these research articles provide knowledge of various techniques to detect AD at its initial phase.

Chapter 3 Methodology: Discussion of the pipeline implemented and its relative data collection used for this research methodology is enlisted in this particular chapter. Data

availability exploration, preprocessing, proposed pipeline, cross-validation and ML algorithms are thoroughly described in this chapter.

Chapter 4 Results and Discussions: From the analysis of the pipeline proposed, all the results of various datasets were embodied in tabular form in an overall manner as well as in their respective category. The ROC curves, accuracies of different classifiers were enlisted in the form of figures in this chapter.

Chapter 5 Conclusions and Future Scope: On the basis of methodology proposed by the analysis of data and literature review, conclusions and future scopes are formulated for all the available data. This chapter also concludes this entire study and provides meaningful information.

Appendix: In this section codes performed in Python using machine learning is provided in ipython notebook format which can be open in Jupyter-notebook.

CHAPTER 2
LITERATURE REVIEW

2.1 General

Evaluation of literature is considered as the foundation of any research study. Stronger the foundation, the more ambitious goals we can achieve. Similarly, if the literature review of the research study is thorough and systematic, it will define the aim of the research very clearly and makes the research very reliable. In this chapter a brief summary of various prestigious journals' research work is discussed on ML and AI techniques to forecast AD and its progression.

2.2 Review of Literature

There are significant studies involving usage of the TADPOLE repository generated from ADNI. In 2020, Thushara A. et al. [5] conducted a study on the data collected from the TADPOLE repository, which includes biomarkers for clinical and neuroimaging applications. Due to the partial nature of the data set, the RF ascription protocol was used to replace the missing parameters. Thus, enhancing the classification accuracy because this technique brimmed the high-dimensional incomplete data more precisely than the mean or median form. 486 attributes were chosen at random manually through feature selection, and the Gini index was calculated for each one. The characteristics are rated based on the final value, and the top 21 features were chosen using the RF algorithm. The data collection obtained after the feature selection procedure was subjected to the RF classification technique. The RF classification model was operated to analyze the records acquired during the attribute selection procedure. However, It was reported that the classifier had an accuracy of 69.33%, as well as macro averaged precision, sensitivity, and F-score value of 69.33%, 40.9%, and 51.4%, respectively. The best results were obtained by the NC class (F-score 65.15%, sensitivity 58.09%, and precision 74.17%), while the lowest results were obtained by the MCI class (F-score 1.36%, Sensitivity 0.79% and, precision 5%). In contrast, the RF Classification model was compared with the reference dataset obtained after using the Genetic algorithm

with a 66 percent break and 20 trees to pick features. Overall precision is 57.54 percent. The best and worst classification accuracy is attained by NC and MCI, respectively. The poor precision in MCI is due to a lower number of patient's data in that class, as per the data set analysis. The results revealed that the RF selection and prediction model outperformed the GA-based selection and prediction model in terms of classification accuracy. Limitations were also highlighted in this article such as the accuracy in the multiclass grouping is a problem in the analysis of AD. In this paper, a four-class grouping challenge for AD is investigated. OVR, SAEZEROMASK, SVM are compared to the proposed system. The accuracy of the data, however, is on the low scale. In contrast to the proposed approach, the range of accuracies in OVR, SAEZEROMASK and SVM was smaller. As a result, the suggested approach is a viable option for AD classification (multiple) problems.

A study was conducted by Ji. Hwan Park [10] et al. in 2020 on the data collected from the Korean National Health Insurance Service database between 2002 and 2010, consisting of 4,894 clinical features such as ICD-10, drug codes, laboratory data, the background of family ,personal disease, and demographics of individuals. To define incident AD, they used definite AD with codes for diagnosis and drugs for dementia, and probable AD with just a diagnosis. RF, SVM, and logistic regression were used to train and test random forest, and forecast event AD in 1, 2, 3, and 4 years. The ML models demonstrated fair results in the first year of prediction in bootstrapping having AUCs of 0.775 (definite AD) and 0.759(probable AD). In second year, 0.730(definite AD) and 0.693(probable AD); in third year, 0.677(definite AD) and 0.644(probable AD); in fourth year, 0.725(definite AD) and 0.683(probable AD). When the whole unbalanced samples were utilized, the findings in the study were identical. Urine protein level, hemoglobin level, and age were all mandatory clinical features chosen in the model of logistic regression. RF, SVM with linear kernel, and logistic regression were carried out as ML algorithms. Nested stratified 5-fold cross-validation with 5 iterations was implemented for model preparation, validation, and testing of data. The variance threshold method38 was used to pick features inside train sets. Validation sets were used to optimize hyper parameters. Tuned hyper-parameters includes the least number of data samples wanted at a leaf node and the number of trees in the forest for RF;

linearization strength for SVM; and the converse of linearization strength for logistic regression. RF performed best in predicting the zero-year incidence of AD, with a precision of 0.823 and an AUC of 0.898. The features that were positively linked to incident AD were defined using logistic regression. This included age, increased urine protein, Zotepine prescriptions, and characteristics that were negatively linked to attribute of incident AD, like reduced hemoglobin, Nicametate, Citrate prescriptions, diagnosis of the external ear and other degenerative disorders. This research sheds light on the usefulness of a data-driven ML model based on large-scale health data in predicting the risk of AD, which could ameliorate patient selection.

In another study held out by Shahbaz.et.al [7] in the year 2019 TADPOLE data were collected from both male and female participants, including those with a moderate cognitive disability, the elderly, and Alzheimer's patients; there are 1,907 features in the dataset for 1,737 respondents. These individuals were classified into five categories depending on their diagnoses: SMC, LMCI, AD, EMCI, CN. The 41 features with sufficient data were subjected to an evaluation to eliminate redundancy. For this case, 28 attributes were chosen after pre-processing and data analysis. The dataset is split into two parts: a training dataset (which accounts for 70% of the total) and test datasets (which account for 30% of the total) (30 percent of the entire dataset). The 2,164 investigation records have been employed for data mining model preparation, while the 927 examination records have been used for model research, according to this distribution. Six different data mining algorithms and machine learning including DT, KNN, DL algorithm, Naive Bayes, GLM, and rule induction, are used to identify the five stages of AD in this study based on the ADNI dataset. Rapidminer studio, a well-known data mining method, is used to execute all of these techniques in this research. Herein, all classification models in this analysis were trained using a 10-fold cross-validation method on the preparing dataset. The CDRSB cognitive exam, which emerges at the cork of the decision tree, is the most distinguishing trait for the five phases of AD, according to the rule induction models and decision tree. The patient's lifespan is the most distinctive demographic feature, while the quantity of the entire brain is the most characteristic clinical evaluation feature. The generalized linear model correctly categorized the majority of the

predicted unknown occurrences of LMCI, AD, EMCI, and CN classes, with division recalls of 86.01%, 100.00%, 90.62%, 94.44 and rank accuracy of 79.45%, 100.00%, 98.12%, 889.16% respectively, out of all classifiers.

In another study conducted by A. Kumar.et.al [11] in the year 2019 Alzheimer's Biomarkers Comprehensive Database was developed. ABCD is built and written in MySQL, with a web interface refined in PHP on the phpMyAdmin framework and hosted on an Apache web server. It is a repository of choice for the scientific community due to its comprehensiveness, uniformity, free availability, ease of use, and support for various user accounts. The data in the database was gathered from the previous publications and online sources such as NCBI, Google Scholar, PubMed, Medical Literature Analysis and Retrieval System Online (MEDLINE), and PMC. ABCD is made up of SNPs, proteins, genes, and microRNAs (miRNAs), all of which work together to offer gene regulatory input to researchers. The user can access verifiable data on miRNAs from miRBASE software. A bespoke PHP script that used the Entrez API to retrieve molecular data via NCBI sites was used here. Chembank as well as other databases were used to get drug information. The latest version of the ABCD has data availability from 843 literature sources that describe 404 medications, 499 genes, and 767 miRNAs. Advanced search, gene regulatory information related search, protein search, and gene search are the primary ABCD interface components. Users can query the system using genes and proteins in the search area. Users may search ABCD by miRNA, mitochondrial gene, pathways, co-expressed gene, Transcription factors (TF), Single Nucleotide Polymorphisms (SNPs), and pharmacological target in the progressive search. The data distribution and frequency are as follows: biological processes (1473), clinical trials (1608), miRNA (767), mitochondrial genes (36), SNP's (132), drugs (404), TF's (1538), proteins (259), genes (499), pathway (8), animal models (17), brain images (35), molecular functions (125), cellular components (148), co-expressed genes (4428).

G. Virar.et.al [12] in 2018 conducted their study on multimodal attributes from CSF, PET, MRI, and DTI. To construct the graph, the author mounts all quantitative aspects from this repository with the markers and encompasses age and gender. The problem of matrix completion is used in the methodology. This article explains the use of meta-information from these rows or columns to create the graph, or it can employ the row or column paths of the matrix to measure an identical metric among the set of vertices. Therefore, they suggest solving the multi-modal disease classification as a geometric matrix completion problem in this paper. On a dataset with partly observed features and marks, the author uses a Separable Recurrent GCNN (sRGCNN) to forecast the disorder and ascribe missing attributes at the same time. Thus, the author created the graph using meta-data from the subjects, as in their gender and age, since these factors are considered to be possible causes for AD. This study contrasts two approaches to creating row graphs using a stratified 10-fold cross-validation approach: one using a similarity metric to combine gender and age information, and the other using Euclidian distance-based KNN to combine gender and age information only. Thus, Hyper parameters were amended using Hyperopt and nested cross-validation on a hold-out validation collection which was set at 10% in every fold of training input, aiming for classification failure (binary cross-entropy). The author concluded that the finest performing technique, which uses a graph framework based on gender and age, attains classification accuracy of 87% with an AUC value of 0.950 at the baseline. Moreover, at all stages of matrix completeness, this approach surpasses definitive classifiers SVC, RF, and MLP.

C. Krittanawong.et.al [13] done the study in 2017 highlighting the upcoming future of AI techniques like ML, DL, and Cognitive assessment which may play a crucial part in CVD medicine and ameliorate patient care quality, increase cost efficiency, and lower readmission and death rates. This would further result in a paradigm shift towards accurate CVD medicine. Big data analytics using AI and their algorithms can allow precision medicine. Big data includes genetics, social media, environmental, and lifestyle-related factors or "omics" data that can be preserved through EHR's or precision medicine frameworks and thus accessed with other physicians or researchers through safe cloud systems. In the big data realm, ML revolutionizes CVD risk thus automating score prediction.ML can be divided into 3 categories: 1. Supervised Learning which involves the use of a dataset labeled by subjects

to forecast the desired outcome. It can be used in both classification and regression problems. It includes (i) ANN (ii) SVM (iii) DT (iv) RF (v) Naïve Bayes (vi) Fuzzy logic (vii) KNN. 2. Unsupervised Learning aims to find new disease pathways, genotypes, or phenotypes in data that are obscured or complex. It can be divided into the following categories: (I) Algorithms for clustering (ii) Algorithms for learning association rules. 3. Reinforcement learning combines supervised and unsupervised learning techniques. Its objective is to use trial and error to boost algorithm accuracy.

A study in India by Rao et al. [14] in 2008 gathered 74 proteins that are thought to be linked in AD pathogenesis. NCBI was used to extract functional protein sequence data in FASTA format for these proteins. ClustalW is given these sequences for Multiple Sequence Alignment, which determines the best fit for the chosen sequences and sets them up so that the names, distinctions, and variations can be seen. Out of 74 proteins identified as essential pathological proteins during the progression of AD, bioinformatics research demonstrates three significant proteins. The current bioinformatics research discovered that the proteins PS-1 and PS-2, as well as APP, play significant parts in the pathogenesis of AD.

Another study in the USA conducted in 2007 by Hamed et al. [15] with emphasis on biomarker selection for prognosis of AD is discussed. The Alzheimer's disease Prediction of Longitudinal Evolution (TADPOLE) challenge provided the record for this investigation. The goal of this challenge is to evaluate several algorithms for future Alzheimer's disease predictions of individuals. The study combined the findings of cognitive tests with quantitative biomarkers such as MRI measures. Cognitive exams are neuropsychological exams that are overseen by clinical professionals. The study uses 1568 samples from each of the three distinct groups (AD, EMCI, and CN) for a total of 4704 samples and 11 characteristics (6 MRI measures and 5 cognitive tests). The utilized mean imputation in this study, which replaces absent values with the mean of a single attribute among all subjects. Normalization techniques or standardization are required after assigning the NaN values to prevent any impartialities that might also alter the final results due to the various ratios of the attributes in the TADPOLE dataset. The authors employed the normalizing approach to

rescale all the features between 0 and 1 in this article since one of the overall feature selection methods deals with nonnegative estimates. In this study, six feature selection procedures across three different categories are used. Tree-based, RFE, and Chi-Square methodology can categorize features by significance. The forward feature selection approach, on the other hand, is unable to rate feature significance directly. Modifying the number of chosen characteristics from 1-11 allows the attributes to be sorted in order of significance to the algorithm. The AdaBoost classifier with multiclassification capacity is employed in this work since the major purpose is to choose the relevant characteristics that might cause the distinction between three distinct groups (AD, EMCI, and CN). AdaBoost is a machine learning meta-algorithm that may be employed to improve the execution of other learning algorithms. By turning weak learners into strong learners, this technique might provide higher accuracy than traditional machine learning methods. To train and test the classifier in the work, a multiclass categorization strategy with 5-fold cross-validation was utilized. This implies that the classifier is instructed with 80% of the data in each fold and evaluated on the remaining data (i.e., 20 percent). The accuracy measure was used to assess the relevance and divisibility of the selected characteristics in the three distinct groups (AD, CN and EMCI). The various approaches used are- The Pearson correlation coefficient heatmap, which is among the filter feature selection methods inclusive of all the cognitive tests and the biomarkers, indicates that ADAS11 and ADAS13 are very connected, therefore one of them may be excluded from the dataset. Finally, the nine chosen characteristics RAVLT, MMSE, CDRSB, MidTemp, Fusiform, Whole-Brain, Entorhinal, Hippocampus, and Ventricles remain. Each feature is assigned a score by the Chisquare, and the greater the value, the more important the feature is. Four features, ADAS11, RAVLT, CDRSB, and ADAS13, are considered more essential features using the Chi-Square approach. Grounded on these findings and comparability with the Pearson's correlation approach, Chisquare doesn't provide information regarding the dataset's redundancy. Both ADAS 11 and ADAS 13 are considered crucial characteristics, as can be seen.

All characteristics are prioritized in the Wrapper-based technique based on their coefficients and significance. The Entorhinal cortex was selected as the best characteristic for classifying AD, EMCI, and CN patients. The other four key features are the MMSE, ADAS13, CDRSB and Hippocampus, which are ranked from 2 - 5. With the LASSO

regularization strategy, there is no clear way to prioritize the relevance of the features in the Embedded technique. The sparsity parameter is the sole factor that may be used to choose lasso features. The lower the sparsity, the lesser features are chosen. Only the MMSE is used as a feature in this investigation since the sparsity is set at 0.001. As a result, we may deduce that Lasso deems this attribute to be more important than the others. When the sparsity was reduced to 0.02, the additional characteristics picked were RAVLT, ADAS13, and CDRSB, but no ranking information was obtained.

The Gini importance is used by the extra tree classifier to assign a rating to each feature. MMSE, ADAS13, CDRSB, RAVLT, and ADAS11 are the top five characteristics chosen, in that order. ADAS11 and ADAS13 were also regarded as essential features by the additional tree classifier; nevertheless, these two features are closely connected. In the feature combination approach, each feature must be evaluated individually, as well as all conceivable combinations, before being sent to a ML algorithm. For classification between the AD, EMCI, and CN groups, MidTemp, which wasn't picked as an essential feature via standard feature selection techniques, but had the best accuracy (91.12%), even, surpassing the feature CDRSB. Furthermore, among all possible combinations for distinguishing between the AD, EMCI, and CN, the amalgam of the CDRSB and MidTemp obtained an accuracy of 92.86%. The findings show that traditional feature selection approaches aren't always trustworthy, and that they might overlook certain critical biomarkers throughout the feature selection route. As a result, fusion-based feature selection may be thought of as a beyond more trustworthy strategy for identifying the utmost important biomarkers in AD research.

One of the studies conducted in 2004 by Ramesh.et.al [16] was conducted in the UK, in which the author concluded AI is capable of analyzing complex medical data, with the ability to be harnessed in each and every field of medicine. It's also used to derive concrete relationships from a data base, which can be used for prediction, treatment, and diagnosis in a variety of clinical scenarios. In this review paper published, an outline of various AI algorithms is explained along with mandatory clinical applications. There are mainly 4 kinds of techniques used in the research article (i) ANN (Artificial Neural Network (ii) Fuzzy expert system (iii) Evolutionary Computation (iv) Hybrid Intelligent Systems.

2.3 Summary of Literature

The development of software that exploits the ability of human intelligence, including certain making decisions, thought, and training, has been aided by advances in computer technology. The article highlights the potential of AI techniques for web-based medical diagnosis, prediction, and implementation. Combining AI and medicine has the potential to minimize costs, time, human expertise, and medical error. As a result, physicians can use it to help with decision-making, diagnosis, and prediction. Telemedicine and electronic health records (EHR) are used to provide health care over the internet. These advances may improve conventional records, allow for faster storage and retrieval, facilitate telemedicine, and encourage medical research. For the interface between the database and the clinician, electronic records could be stored and updated on a regular basis. The author's model is a web-based medical diagnosis and prediction system with four parts: databases, prediction module, diagnosis module, and user interface. [17]

The databases are of two types: patient database which stores patient information and patients-disease database which comprises diseases, treatments and other attributes about tests and therapies. The training and test datasets are taken from the patients-disease database, and the training weight is saved to predict new data fed into the system. The second main feature of the diagnosis module is a hybrid (expert-fuzzy system) that performs diagnosis tasks by combining expert systems and fuzzy logic. AD is a progressive neurological disorder marked mainly by amyloid-beta plaques and NFT's resulting in the death of neurons [18].

The protein-forming deposits are central to disease and its progression. The three genes are beta-amyloid precursor gene, PS-1; PS-2 located on chromosome number 21, 14 and 1 respectively. Artificial intelligence in medicine chiefly involves the use of computer techniques to diagnose, treat and predict the results. The art of AI has highly measurable improvements as compared to previous systems. The authors devised findings of 11 papers that demonstrated the performance of medical AI [19].

The purpose of this article is to present best practices for writing and reviewing AI for medical image analysis articles [20]. Proof of technological viability, expert-level results, and

clinical performance are all possible evidence goals for accepting conclusions. Tasks involving image processing include the classification which is as follows: (binary, multiclass). Regression, Localization: Return localization to the particular image. Segmentation: Classifies pixel as or not as a part of an object. Learning approach involves- Supervised Learning, Unsupervised Learning. Data collection and processing- Collection: Description of data includes whether the collection was retrospective or prospective, sampling is consecutive or convenience and Processing. The division into subsets: Division of datasets includes training and testing set of the algorithm. Since there are a small number of experts available to provide correct picture labels, data labeling is performed on test sets. Radiology reports, expert consensus, and reference standards are examples of these. Model training involves hardware and software's. Hyper parameters include numeric parameters; examples of this include learning rate, regularization variables. Quality evaluation. The -F1 score is used to determine summary measurements for binary classification: Precision is the average predictive value, while recall is the sensitivity, with a range of 0-1. The Youden J index considers both false negatives and positives to be unfavorable. ROC AUC curve: used as a single metric to summarize a classifier's entire output range. Summary measures for Multi-class classification-By averaging macro average and micro average. Thus, AI in the infant stages for application of medical imaging, patient's safety demands.

The goal of this study was to look for direct components and associated qualities that are believed to be entangled in AD pathogenesis. The PSEN2, PSEN1 genes and APP are presently thought to be implicated in genetic forms of AD. However, only the APOE gene has been firmly linked to the exposure of AD. Weka 3.6.9, RapidMiner Studio 6.2.0, and enrichment analysis for functional gene categorization with the David tool were used to classify gene dataset. The web-based GEne SeT AnaLysis Toolkit (WebGestalt) was used to conduct the gene ontology (GO) analysis. The C4.5 algorithm, which is used in the RapidMiner toolset, interprets missing data differently from regular data. Weka Tool was used to build association rules, which were decided by numerous attributes. The author used Alzheimer's genome data from a variety of typical internet resources, including NCBI, AlzGene, ensembl gene and GenCard. They chose 2111 raw genes that are known to be significant to AD. The dataset comprises 14 features that describe distinct aspects of AD

genes. Chi-squared characteristic analysis and gain ratio were used. The model was validated using the 10-fold cross-validation deployment operator. Descriptions of gene data include MMSE score, chromosomal location, association score, and gene name. This gene collection was also subjected to enrichment analysis in order to classify genes based on similarity that used a clustering technique. Stringency was retained for more strongly related genes in each group by maintaining the kappa threshold at 0.3, as anything underneath this level has a high likelihood of being noisy. The categorization of data was accomplished via a J48 algorithm developed in Weka, while C4.5 was employed in RapidMiner. This study yields the accurate categorization of 950 of 2111 genes. The information gain criteria revealed that Huge navigator and MMSE were the most informative variables in this gene collection. The specificity, sensitivity, and DT accuracy, were calculated to be 81%, 86% and 71% respectively. Enrichment analysis affects gene function in terms of GO, phenotype analysis, illness association, pathway analysis, drug association, and gene functional categorization. According to gene functional categorization assessment, the major genes that are highly related with AD include APOE, ACE, GRN, PSEN1, BCHE, IL1A, and PRNP [21].

CHAPTER 3

METHODOLOGY

3.1 General

This chapter discusses the methodology adopted for this research work. In this section we will discuss about the data collection which encompasses availability description and preprocessing of data followed by pipeline, analysis of data, cross-validation and methods used in the analysis of data. This chapter thoroughly explains each and every step taken for the completion of this research study and also explains the analysis methods used in this study.

3.2 Data Availability

In this research work, we utilized the TADPOLE grand challenge data generated from ADNI. ADNI is a multicenter study aimed to improve the genetic biomarkers and neuroimaging data for early diagnosis of AD. A standard set of procedures and protocols is followed during ADNI data collection, to avoid any inconsistencies in the data [7]. The data consists of demographic attributes, clinical assessment and cognitive scores, genetic biomarkers and multi-modal MRI data.

The collection of sequence data is performed via UNIPROT. The AD-related genes are labeled as positive dataset and Non AD-related genes are labeled as the negative dataset. We retrieved sequences from each set of 72 AD and Non-AD related genes.

3.3 Data Description and Exploration

The TADPOLE challenge is categorized into training, prediction and testing dataset namely D1, D2 (Longitudinal data) and D3 (cross-sectional data). The number of patients recorded in longitudinal data and cross-sectional data is 12741 and 896 respectively. The D1 and D2 longitudinal training and prediction sets contain data from rollover individuals who

were asked to provide forecasts. The number of patients recorded here are 12741 with an age range of 54.4-94.4. The stages of progression of disease provided in the dataset are: LMCI (Late mild cognitive impairment), EMCI (Early mild cognitive impairment), SMC (Subjective Memory Complaint), CN (Cognitively Normal), and AD (Alzheimer's disease) with a total count of 4644, 2319, 3821, 1568 and 389 respectively. [22].

In contrast, the cross-sectional prediction set (D3) consists of a limited set of variables and single time point from each rollover participant of the training dataset. The number of individuals here is 896 with an age range of 55.0-99.3. The stages of progression of disease provided in the dataset are NL to MCI, NL, MCI, MCI to NL, MCI to Dementia, Dementia to MCI and Dementia with values of 8, 292, 250, 8, 39, 3 and 142 respectively.[22]

The raw sequence data consists of 72 sets of AD-related and NON-AD related genes. The initial number of sequences in the positive dataset is 493. While, negative dataset holds 610 sequences.

3.4 Data Preprocessing

3.4.1 Feature Selection

When generating a predictive model, attribute selection is the process of minimizing the amount of input variables. The quantity of input variables should be condensed to lower the expense of modelling computation and, in some situations, to escalate the model's performance. By improving the model's generalization, this strategy lowers the problem of over fitting. As a result, it aids in improved data interpretation, enhances prediction performance, and reduces the computing time and space necessary to execute the algorithm.

The first group of attributes is demographics, which includes individual's broad measurable traits. Clinical evaluation attributes, which include important biomarkers for Alzheimer's disease, constitute the second group. Following each participant's clinical assessment, these characteristics were noted. Finally, there is the cognitive assessment

attributes category, which includes qualities that indicate a participant's cognitive behavior. Various cognitive evaluation tests are conducted for this purpose, and scores are issued to individuals depending on their cognitive ability.

The features extracted from TADPOLE data are selected manually through a profound understanding of the literature. The demographic, clinical and cognitive attributes along with description are illustrated in table 3.1, 3.2, 3.3. The features selected from longitudinal and cross-sectional data are 36 and 18 respectively

Table 3.1: Demographic Features and their respective description.

DEMOGRAPHIC ATTRIBUTES	
FEATURES	**DESCRIPTION**
RID	Participant Roster ID
VISCODE	Visit Code
SITE	Where visit took place
DX_bl	Baseline diagnosis from the first visit
AGE	Individual Age at baseline(bl)
PTGENDER	Individual Sex
PTRACCAT	Race
PTETHCAT	Individual Ethnicity
PTMARRY	Individual Marital status at baseline
DX	Clinical diagnosis(current and baseline)
Years_bl	Years from baseline

Month_bl	Months from baseline
Month	Months from baseline(to nearest 6 months, as a factor)

Table 3.2: Clinical Assessment and their respective description.

CLINICAL ATTRIBUTES	
FEATURES	**DESCRIPTION**
APOE4	Gene APOE4(Risk Factor)
Ventricles	Measures of brain structural integrity(MRI ,DTI measures)
Hippocampus	
WholeBrain	
Entorhinal	
Fusiform	
Midtemp	
ICV	
Midtemp_bl	Measures of brain structural integrity(MRI , DTI measures) at baseline
ICV_bl	
Fusiform_bl	
Entorhinal_bl	
WholeBrain_bl	
Ventricles_bl	
Hippocampus_bl	
FDG	

AV45_bl	
FDG_bl	Avg FDG-PET(angular , temporal , posterior) at baseline
PIB_bl	Avg PIB (cortex, anterior , parietal, frontal) at baseline

Table 3.3: Cognitive Scores and their respective description.

COGNITIVE ATTRIBUTES	
FEATURES	**DESCRIPTION**
CDRSB	Clinical Dementia Rating scale Sum of Boxes score
ADAS11	AD Assessment Scale - 11
ADAS13	AD Assessment Scale – 13
MMSE	Mini Mental State Examination
RAVLT	Rey Auditory Verbal Learning Test
RAVLT_immediate	Rey Auditory Verbal Learning Test(sum of 5)
FAQ	Functional Activities Questionnaire
MOCA	Montreal Cognitive Assessment
EcogPtMem	Everyday Cognition Participant Memory
EcogPtLang	Everyday Cognition Participant Language
EcogPtVisspat	Everyday Cognition Participant finding ability
EcogPtPlan	Everyday Cognition Participant Plans
EcogPtTotal	Everyday Cognition Participant Total Score
EcogPtMem_bl	Everyday Cognition Participant Memory at baseline

EcogPtLang_bl	Everyday Cognition Participant Language at baseline
EcogPtVisspat_bl	Everyday Cognition Participant finding ability at baseline
EcogPtPlan_bl	Everyday Cognition Participant Plans at baseline
EcogPtTotal_bl	Everyday Cognition Participant Total Score at baseline
CDRSB_bl	Clinical Dementia Rating scale Sum of Boxes score at baseline
ADAS11_bl	AD Assessment Scale – 11 at baseline
ADAS13_bl	AD Assessment Scale – 13 at baseline
MMSE_bl	Mini Mental State Examination at baseline
RAVLT_bl	Rey Auditory Verbal Learning Test at baseline
RAVLT_immediate_bl	Rey Auditory Verbal Learning Test(sum of 5) at baseline
FAQ_bl	Family history Questionnaire at baseline

On the other hand, the features selected for sequence data involve the use of Cluster Database at High Identity with Tolerance using a sequence identity of 80%. CD-HIT is employed to lower the overall content of the database by only eliminating 'redundant' sequences. As a result, from a given FASTA sequence database, it produces a collection of closely related protein families. The total number of retrieved sequences from CD-HIT from set of 72 AD and Non AD genes is 110 (for each positive and negative dataset). Further, the obtained sequences are used as a file to run on a python package; iFeature.

iFeature is a toolkit for creating numerical feature representation schemes for protein sequences that is written in Python3. To reflect numerical sequence profiles, it primarily extracts structural and physiochemical properties from sequence data. We applied 19 different types of descriptors on both positive and negative data, which is demonstrated in table 3.4.

Table 3.4: Descriptors and their respective description.

I-FEATURE DESCRIPTORS	DESCRIPTION
AAC	Amino acid Comp
APPAC	Amphillic Amino Acid Comp
CKSAAGP	Comp of K-spaced Amino Acid Group pair
CTDC	Comp Descriptor
CTDD	Distribution Descriptor
CTDT	Transition Descriptor
CTriad	Conjoint Triad
DDE	Di-peptide Deviation from Expected Mean
DPC	Di-Peptide Comp
GAAC	Grouped Amino-Acid Comp
GDPC	Grouped Di-Peptide Comp
Geary	Autocorrelation Descriptors
Moran	Autocorrelation Descriptors
NMBroto	Normalized Moreau-Broto
PAAC	Pseudo Amino Acid Comp
QSOrder	Quasi-Seq Order
SOCNumber	Seq Order Coupling Number
TPC	Tri-Peptide Comp
CKSAAP	Comp of K-spaced Amino Acid Pairs

3.4.2 Data Normalization

The original data had various missing and redundant values. Since, high dimensional data and extreme variance of features are loopholes for prediction in machine learning, we followed up with the approach of data normalization.

To overcome the sparseness in provided data, we performed Feature Scaling techniques such as Label Encoder and Min-Max Scalar. Similarly, to load the null values a mean for the particular cell was generated from the overall column. Subsequently, an additional procedure was followed for sequence data named 'bfill ()' technique. The function bfill () is used to backfill the dataset's redundant data. It will fill the NaN values in the pandas data frame in reverse order.

Further, the TADPOLE and Sequential dataset is divided into training (80% of the entire dataset) set and test (20% of the entire dataset) set. To make the algorithms consistent and validate their performance similarity 5-fold cross-validation is deployed in both TADPOLE and sequence dataset.

3.4.2.1 Min-Max Scalar

The data is scaled to a defined range - generally 0 to 1 - in this method. The cost of having this constrained range is that we will end up with lower standard deviations, which can dampen the influence of outliers. Min-Max Scalar is denoted by the following equation.

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

3.4.2.2 Label Encoder

We frequently work with datasets in machine learning that include numerous labels in one or more columns. Labels in the set of words or numbers can be used. The training data is frequently labeled in words to make it intelligible or human-readable. The method of

converting labels into the numeric form so that computers can read them is known as label encoding. Machine learning algorithms, on the other hand, will make better choices about how to use those names. It is a key pre-processing stage for the organized dataset in supervised learning.

3.5 Proposed Pipeline

In our research work, we implemented four different types of algorithms comprising of RF, SVM, RF+SVM and ANN on TADPOLE to facilitate the classification of different stages of AD. While, the purpose of algorithms used in Sequence data measures the best-indicated descriptor to classify AD progression. The variation of datasets cleaves the pipeline into two sections i.e., the framework of TADPOLE data and the Sequence data which further assist in the early prediction of AD.

3.5.1 TADPOLE Data

The pipeline proposed in our study of TADPOLE data comprises of assorted biomarkers which aid in the prognosis of AD. To express the pipeline, Fig. 3.1 is illustrated.

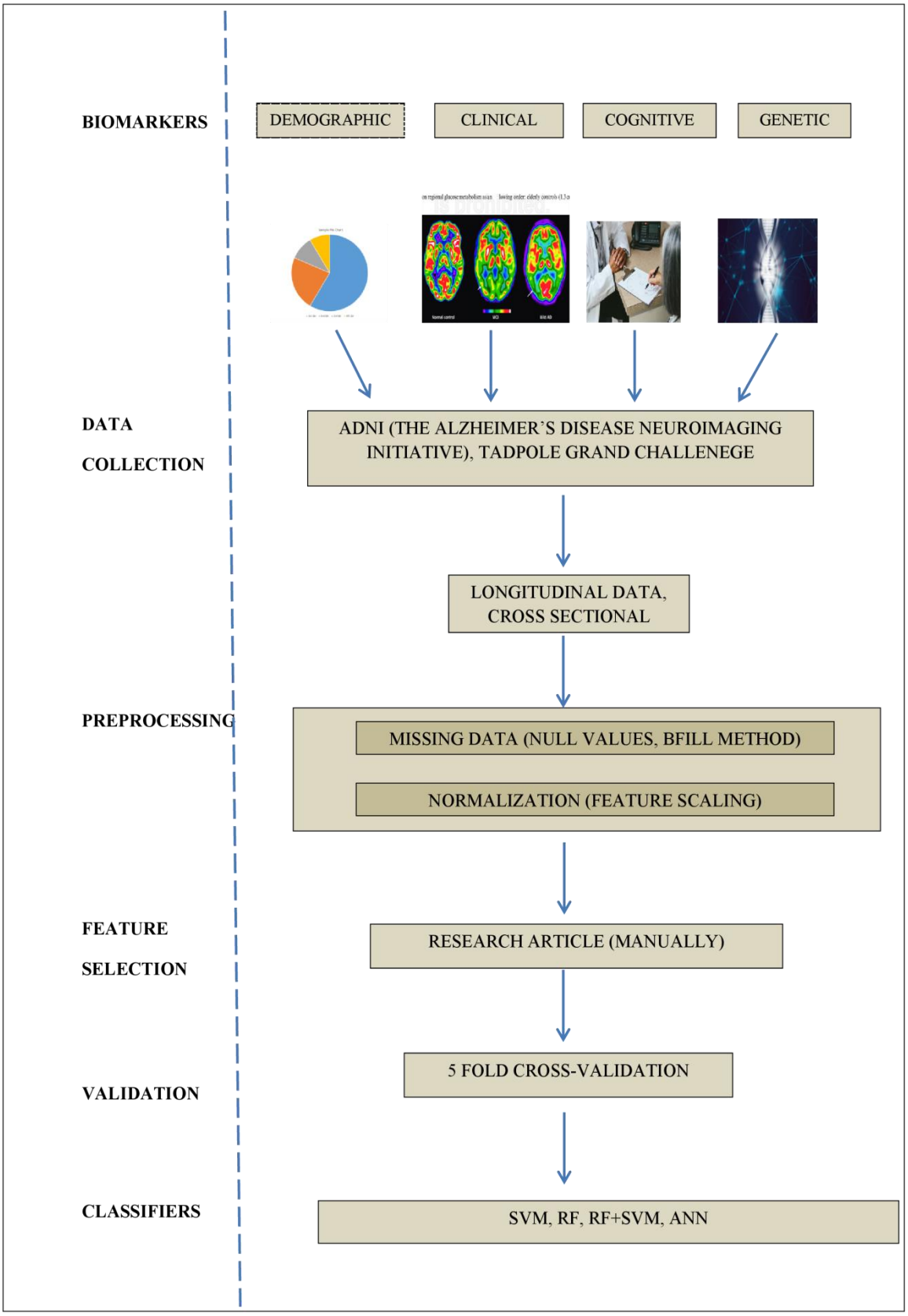
BIOMARKERS
DEMOGRAPHIC
CLINICAL
COGNITIVE
GENETIC
DATA COLLECTION
ADNI (THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE), TADPOLE GRAND CHALLENEGE
LONGITUDINAL DATA, CROSS SECTIONAL
PREPROCESSING
MISSING DATA (NULL VALUES, BFILL METHOD)
NORMALIZATION (FEATURE SCALING)
FEATURE SELECTION
RESEARCH ARTICLE (MANUALLY)
VALIDATION
5 FOLD CROSS-VALIDATION
CLASSIFIERS
SVM, RF, RF+SVM, ANN

Figure 3.1: Pipeline suggested for TADPOLE challenge Data.

3.5.2 Sequence Data

The framework of the available gene sets and classifiers precisely predicts the stages of AD. To emphasize the plan of sequence data, Fig 3.2 is illustrated below.

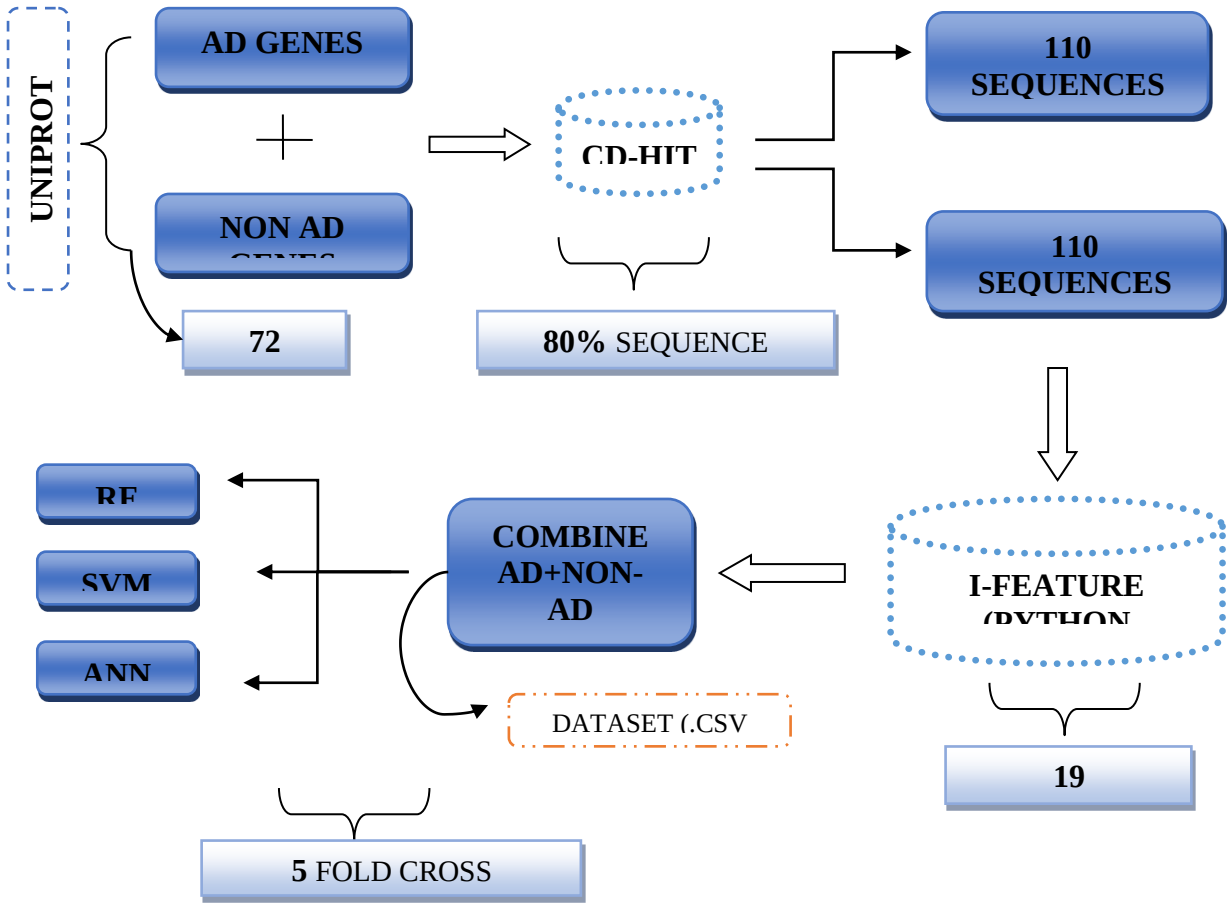

Figure 3.2: Pipeline suggested for Sequence data extracted from UNIPROT.

3.6 Cross Validation

Cross-validation is a series of techniques for measuring a method's efficiency on unknown data while minimizing the estimation's bias. This is accomplished by using two separate sets of data: a training set and a test set. The most common cross-validation approach is to divide the data into k partitions (generally, k ranges from 2 to 10). We utilized the 5-fold validation on each dataset. The output of each of these k-folds is then averaged through folds, with the union of the other folds serving as the test set.

3.7 Random Forest

RF is the category of supervised learning technique. Ensemble learning intends to generate an approach that integrates a range of simpler models of varying strengths. From the available dataset, a set of classifiers is trained, and then an amalgamation of them is built. Each of the distinct classifiers sheds a weighted vote, and the sum of the outcomes decides the subject's expected period. Moreover, it holds an upper hand over handling missing values and reduces over fitting in the model. Additionally, the algorithm classifies the categorical values of the dataset precisely.

3.8 Support Vector Machines

SVM is a type of supervised learning technique that can be exercised in both classification and regression problems. Mostly SVM is availed for classification problems. It can perform both linear and non-linear classification (through Kernel). The Kernel Function is used to convert n-dimensional input to m-dimensional input, where m is much greater than n, and then efficiently find the dot product in higher dimensions. The aim of SVM is to pick the ideal separating hyperplane for setting a threshold between points of separate categories. In high dimensions, the method is considered to be reliable.

The RBF (Radial Basis Kernel) forms the basis of the SVM classification in this study. It is used to find a non-linear classifier or regression line in machine learning. The mathematical expression of the kernel is as follows:

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right)$$

Here, x' and x are vectors in fixed dimension space.

3.9 RF+SVM

RF+SVM hybrid algorithm is employed to refine the results of SVM classifiers. Here RF is used as an auxiliary algorithm. The comprehensive analysis of the classifiers depicts their feasibility to increase the efficiency of SVM algorithm. RF-SVM can effectively predict data with very high dimensions and thus is highly accurate, generalizes better, and interpretable.

3.10 Artificial Neural Network

ANNs are useful data-driven modeling tools broadly employed for nonlinear systems and thus are a type of both supervised and unsupervised learning algorithm. The application includes solving obstacles related to text, image and tabular data. A neural network's general purpose is generating an output outline in response to a specific input design that is closely connected to how the brain works. These mappings are learned in the same way that the brain does. There are many types of neural networks, but the multi-layer perceptron has been encountered in this study. Multilayer Perceptron Artificial Neural Networks are computational prototypes that are widely employed to model and find designs in complex relationships amid inputs and outputs. The MLP period of neural networks includes a collection of training models in order to deduce a classifier that can anticipate a valid output cost in supervised learning. Mathematically, MLP can be written as:

$$y = \varphi\left(\sum_{i=1}^{n} \omega_i x_i + b\right) = \varphi(w^T x + b)$$

Where, w is the weights vector, x is the inputs vector, b denotes the bias, and φ represents the activation function.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 General

This chapter reviews the outcomes of this study. After analysis the results of ML algorithms using neuroimaging, sequence and clinical data, their respective accuracies, sensitivity (recall), specificity, precision, recall F1-score are listed in this chapter. Relatively their feature importance and ROC curve, AUC score is also indicated herewith. This chapter also discusses the most relevant biomarker in the early diagnosis of AD.

4.2 Feature Importance

To overcome high dimensionality data and predict the relevant features, a score is assigned to input features, which mark the feature importance of a model. This strategy aids in the better performance of models. Fig 4.1 and Fig 4.2 demonstrates the feature importance. From the results obtained, we conclude that the most relevant feature used for early detection of AD is cognitive scores in TADPOLE datasets.

4.2.1 Results of TADPOLE Longitudinal Data

The results below indicate that the most vital attribute for early diagnosis of AD using TADPOLE longitudinal data is Cognitive attribute in contrast to Demographic and Clinical. Here, in Figure 4.1 CDRSB_bl has the highest feature importance of 0.27, in comparison to all features. Second, the most essential feature is ADAS13_bl followed by RID and MMSE_bl respectively.

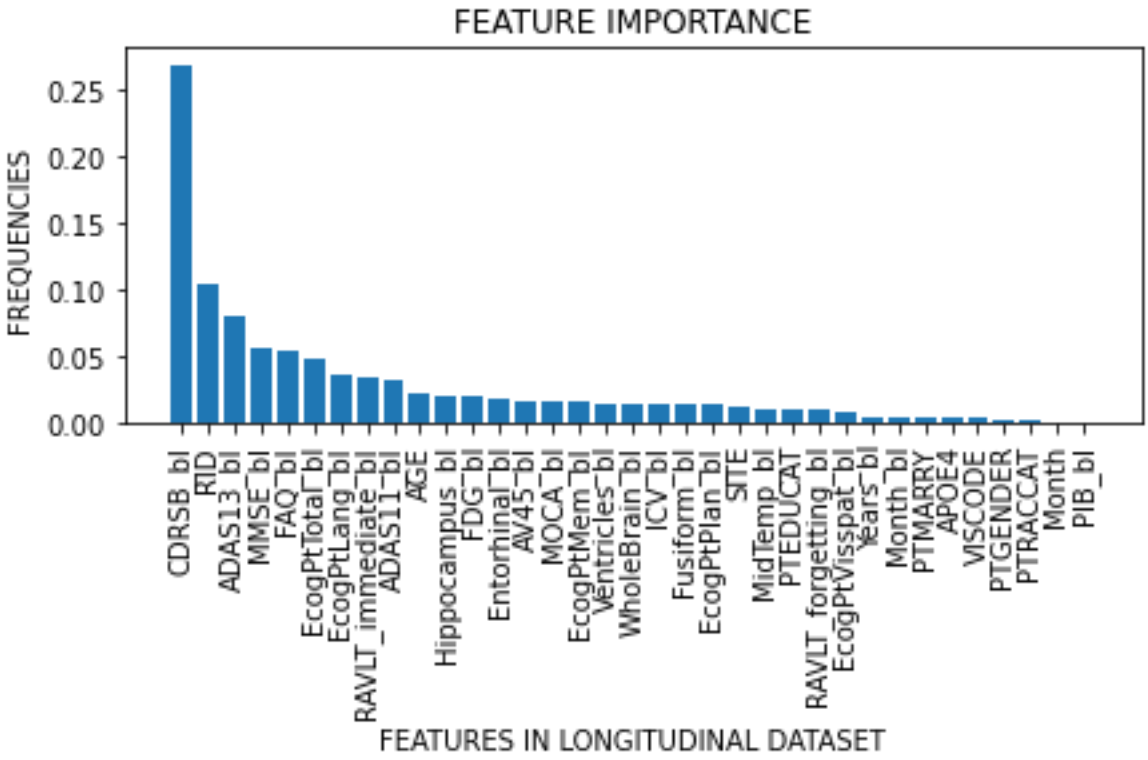

Figure 4.1: Feature Importance for Longitudinal Dataset (baseline).

4.2.2 Results of TADPOLE Cross-Sectional Data

The results below indicate that the most crucial attribute for early diagnosis of AD using TADPOLE Cross-Sectional data is Cognitive attribute rather than Demographic and Clinical. Here, in Figure 4.2 ADAS13 has the highest feature importance of 0.23, in contrast to all features. Second, the most essential feature is RID followed by RID and AGE respectively.

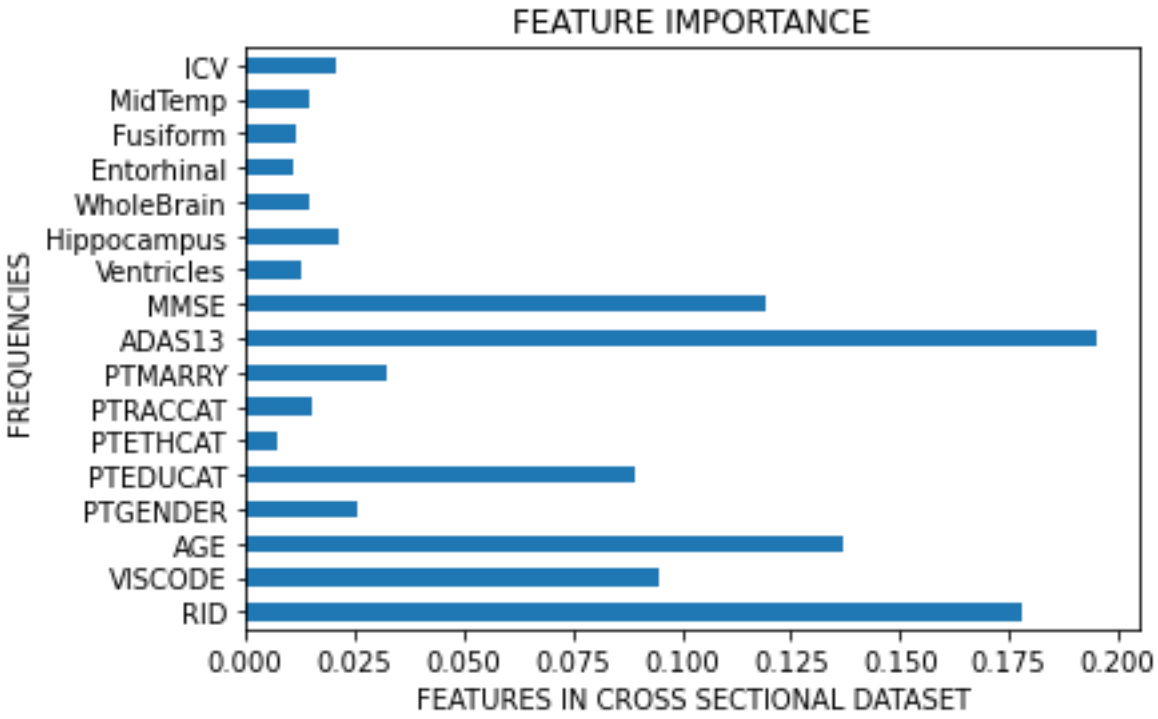

Figure 4.2: Feature Importance for Cross-Sectional Dataset (non-baseline).

4.3 Model Evaluation

In this study, four different models are evaluated based on their performance and accuracy scores. The model with maximum accuracy, precision and recall score is considered ideal for AD prognosis. The evaluation of models and respected scores are demonstrated in Fig 4.3 and Fig 4.4.

4.3.1 Results of Model Evaluation of TADPOLE Longitudinal Data

The results below indicate that the most decisive algorithm for early diagnosis of AD using TADPOLE Longitudinal data is the Random Forest algorithm rather than SVM, ANN and RF+SVM combined. Here, Figure 4.3 RF has the achieved the highest accuracy of 99.8%, in contrast to all techniques. Second, the most essential method is RF+SVM followed by ANN and SVM with accuracies of 97%, 96%, and 55.6% respectively.

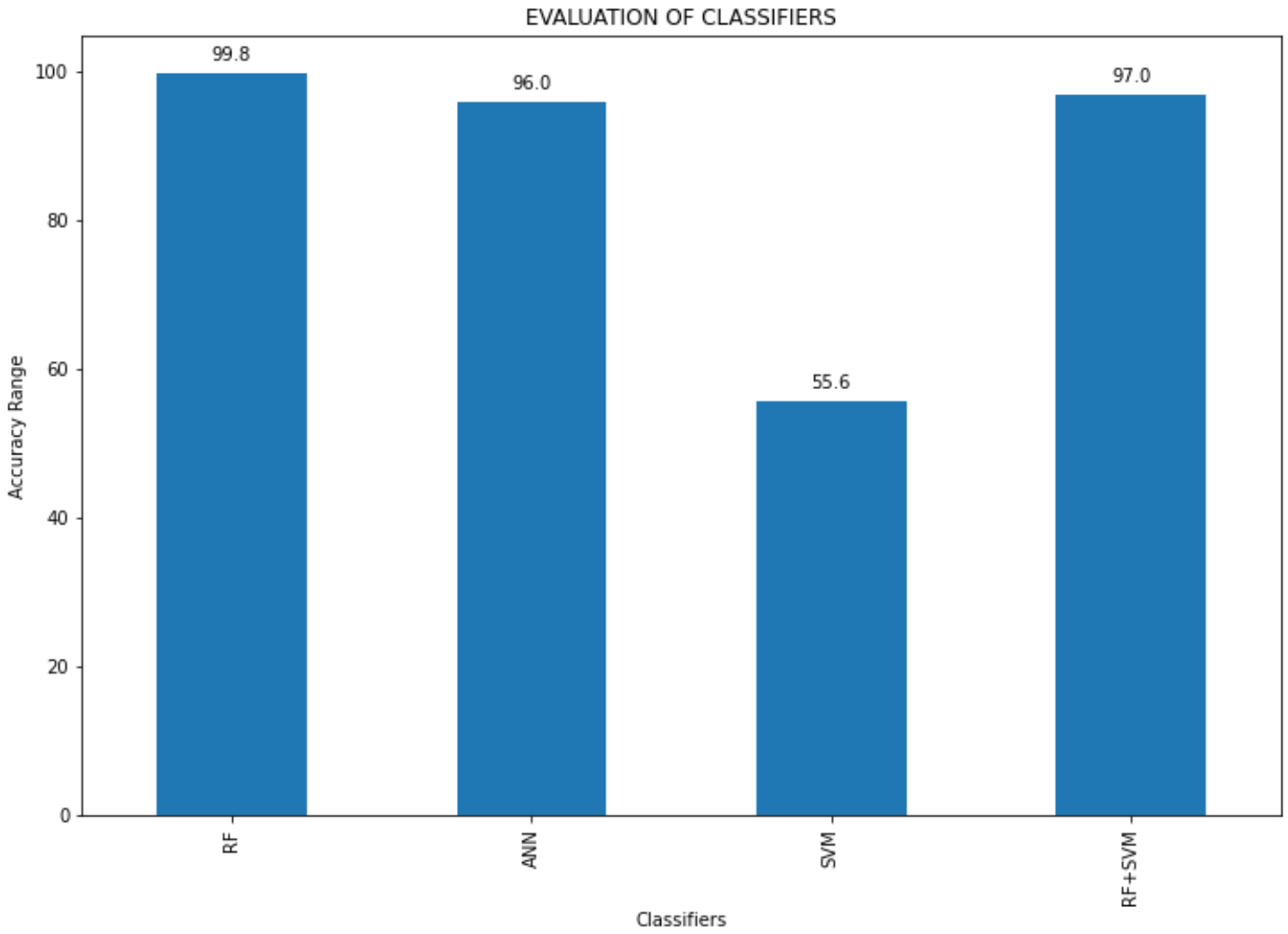

Figure 4.3: Evaluation of different classifiers in Longitudinal Dataset.

4.3.2 Results of Model Evaluation of TADPOLE Cross-Sectional Data

The results below indicate that the most imperative algorithm for early diagnosis of AD using TADPOLE Cross-Sectional data is the Random Forest algorithm rather than SVM, ANN and RF+SVM combined. Here, in the Figure 4.4 RF has achieved highest accuracy of 67.2%, in contrast to all techniques. Second, the most essential method is RF+SVM followed by SVM and ANN with accuracies of 66.7%, 58.3% and 48.3% respectively.

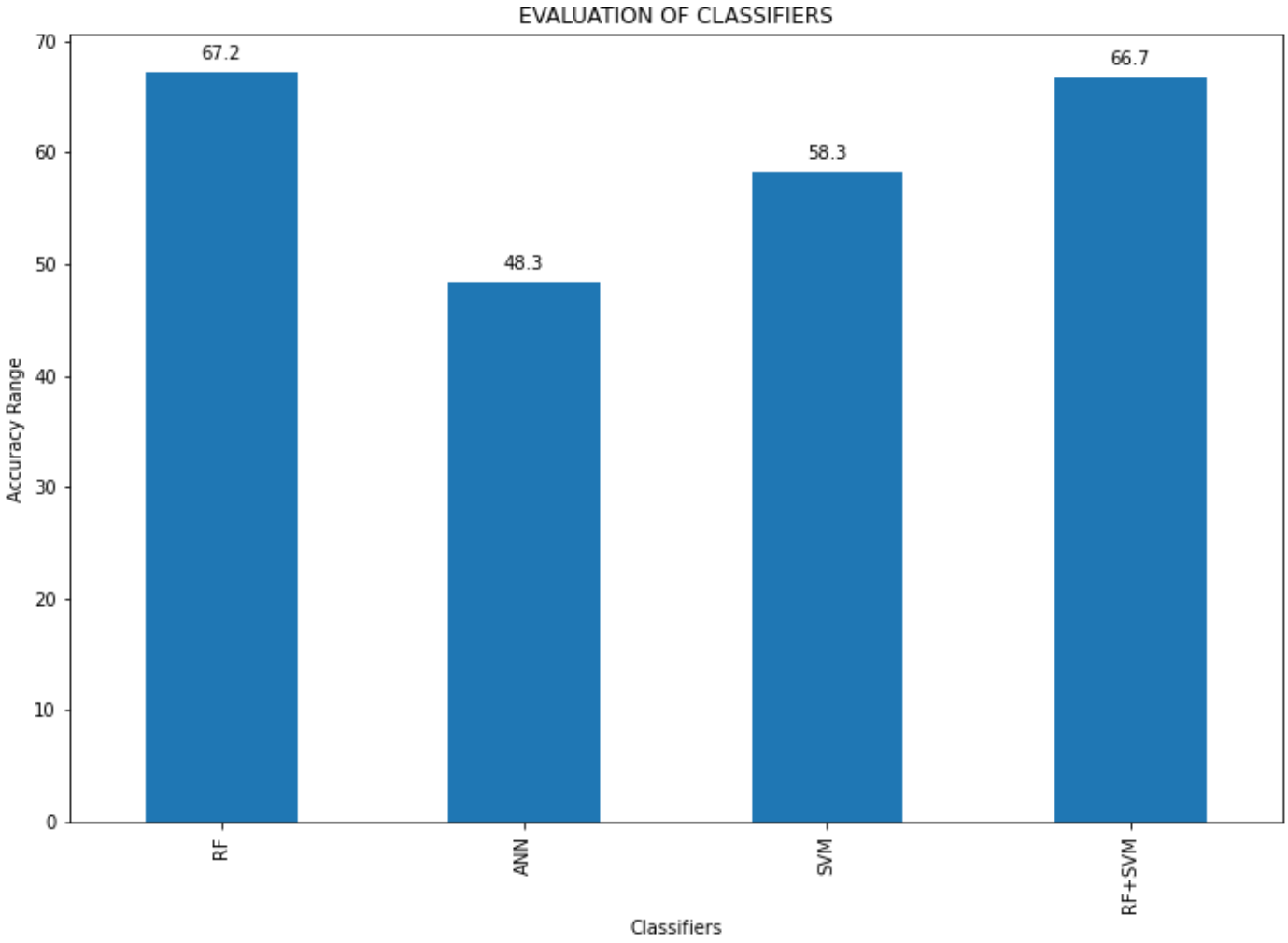

Figure 4.4: Evaluation of different classifiers in Cross-Sectional Dataset.

4.4 Performance Metrics

As an output of the classification, we report the balanced accuracy, area under the ROC curve (AUC), accuracy, sensitivity and specificity. Additionally, to calculate other desired metrics with this data, the expected class for each subject is computed.

$$Precision = \frac{True\ Postive(TP)}{TP(True\ Positive) + FP(False\ Positive)}$$

$$Accuracy = \frac{TP(True\ Positive) + TN(True\ Negative)}{TP(True\ Positive) + TN(True\ Negative) + FP(False\ Positive) + FN(False\ Negative)}$$

$$Recall/Sensitivity = \frac{TP(True\ Postive)}{TP(True\ Postive) + FN(False\ Negative)}$$

$$Specificity = \frac{TN(True\ Negative)}{TN(True\ Negative) + FP(False\ Positive)}$$

$$F1\ Score = 2 \times \frac{Precison\ \times Recall}{Precision + Recall}$$

4.4.1 Results of Classifiers in Longitudinal data with their Relative Performance Metrics

The results below indicate that the most imperative algorithm to forecast distinct stages for early diagnosis of AD using TADPOLE Longitudinal data. The Random Forest algorithm has attained the utmost accuracy rather than SVM, ANN and RF+SVM combined. Here, table 4.1 represents distinctive stages along with their respective accuracies, precision, recall, F1 score for all the classifiers employed for baseline attributes.

Table 4.1: Classifiers and their respective Stages, Precision, Recall, F1 score, Accuracy.

CLASSIFIERS	STAGES	PRECISION	RECALL	F1-SCORE	ACCURACY
RANDOM FOREST	AD	1.0	1.0	1.0	99.88%
	CN	1.0	1.0	1.0	
	EMCI	1.0	1.0	1.0	
	LMCI	1.0	1.0	1.0	
	SMC	1.0	1.0	1.0	
ANN	AD	0.99	0.96	0.98	97.01%
	CN	0.97	0.97	0.97	
	EMCI	O.97	0.96	0.96	
	LMCI	0.96	0.98	0.97	
	SMC	0.99	0.88	0.93	
SVM	AD	1.0	0.09	0.16	55.00%
	CN	1.0	0.45	0.62	
	EMCI	1.0	0.18	0.3	
	LMCI	0.45	1.0	0.62	
	SMC	1.0	0.01	0.02	
RF+SVM	AD	1.0	1.0	1.0	96.58%
	CN	1.0	1.0	1.0	
	EMCI	1.0	1.0	1.0	
	LMCI	0.92	1.0	0.96	
	SMC	1.0	0.01	0.02	

4.4.2 Results of Classifiers in Cross-Sectional data with their Relative Performance Metrics

The results below indicate that the most significant algorithm to forecast distinct stages for early diagnosis of AD using TADPOLE Cross-Sectional data. The Random Forest algorithm has attained the utmost accuracy rather than SVM, ANN and RF+SVM combined. Here, table 4.2 represents distinctive stages along with their respective accuracies, precision, recall, F1 score for all the classifiers employed for baseline attributes.

Table 4.2: Classifiers and their respective Stages, Precision, Recall, F1 score, Accuracy.

CLASSIFIERS	STAGES	PRECISION	RECALL	F1-SCORE	ACCURACY
RANDOM FOREST	NL	0.71	0.82	0.76	67.22%
	MCI	0.65	0.63	0.64	
	DEMENTIA	0.62	0.71	0.67	
	MCI to DEMENTIA	0	0	0	
	MCI to NL	0	0	0	
	NL to MCI	0	0	0	
	DEMENTIA to MCI	0	0	0	
ANN	NL	0.67	0.34	0.45	48.33%
	MCI	0.42	0.69	0.52	
	DEMENTIA	0.52	0.76	0.62	
	MCI to DEMENTIA	0	0	0	
	MCI to NL	0	0	0	
	NL to MCI	0	0	0	
	DEMENTIA to MCI	0	0	0	
SVM	NL	0.59	0.84	0.7	58.33%
	MCI	0.55	0.48	0.51	
	DEMENTIA	0.67	0.29	0.4	

	MCI to DEMENTIA	0	0	0	
	MCI to NL	0	0	0	
	NL to MCI	0	0	0	
	DEMENTIA to MCI	0	0	0	
RF+SVM	NL	0.75	0.77	0.76	66.66%
	MCI	0.61	0.68	0.64	
	DEMENTIA	0.58	0.71	0.64	
	MCI to DEMENTIA	0	0	0	
	MCI to NL	0	0	0	
	NL to MCI	0	0	0	
	DEMENTIA to MCI	0	0	0	

4.4.3 Results of Classifiers in Sequence data with their Relative Performance Metrics

The results below indicate that the most significant algorithm to estimate distinct stages for early diagnosis of AD using UNIPROT Sequence data. The CKSAAP, DPC, NMBroto, TPC have attained the utmost accuracy rather than other 19 descriptors used.

Here, the table 4.21 displays maximum accuracy of TPC descriptor using ANN, SVM, and RF which are 86.36%, 84.09%, 81.81%. Second maximum accuracy is achieved by CKSAAP descriptor using ANN, RF which is 84.09% and 81.81%. Relatively, third highest accuracy is acquired by NMBroto descriptor using RF which is 81.81%. Similarly, fourth accuracy is of DPC descriptor with 79.54% accuracy using ANN technique.

The results are shown in the tables (4.3-4.21) below with their classifier, descriptors, F1 score, precision, recall (sensitivity), and specificity. The 0 stage represents AD related genes , while 1 indicates non AD related genes.

Table 4.3: Classifiers and their respective AAC Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
AAC	SVM	0	0.69	0.41	0.51	0.409	61.36%
		1	0.58	0.82	0.68		
	RANDOM FOREST	0	0.7	0.64	0.67	0.636	68.18%
		1	0.67	0.73	0.7		
	ANN	0	0.68	0.59	0.63	0.59	65.90%
		1	0.64	0.73	0.68		

Table 4.4: Classifiers and their respective APAAC Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
APAAC	SVM	0	0.40	0.18	0.25	0.181	45.45%
		1	0.47	0.73	0.57		
	RANDOM FOREST	0	0.65	0.5	0.56	0.5	61.36%
		1	0.59	0.73	0.65		
	ANN	0	0.72	0.59	0.65	0.59	68.18%
		1	0.65	0.77	0.71		

Table 4.5: Classifiers and their respective CKSAAGP Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
CKSAAGP	SVM	0	0.63	0.55	0.59	0.545	61.36%
		1	0.6	0.68	0.64		
	RANDOM FOREST	0	0.75	0.68	0.71	0.681	72.72%
		1	0.71	0.77	0.74		
	ANN	0	0.74	0.77	0.76	0.772	75%
		1	0.76	0.73	0.74		

Table 4.6: Classifiers and their respective CKSAAP Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
CKSAAP	SVM	0	0.57	1.00	0.73	1.00	72.72%
		1	1.00	0.57	0.73		
	RANDOM FOREST	0	0.70	0.88	0.78	0.875	81.81%
		1	0.92	0.79	0.85		
	ANN	0	0.71	0.94	0.81	0.9375	84.09%
		1	0.96	0.79	0.86		

Table 4.7: Classifiers and their respective CTDC Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
CTDC	SVM	0	0.46	0.27	0.34	0.272	47.72%
		1	0.48	0.68	0.57		
	RANDOM FOREST	0	0.75	0.55	0.63	0.545	68.18%
		1	0.64	0.82	0.72		
	ANN	0	0.68	0.59	0.63	0.590	65.90%
		1	0.64	0.73	0.68		

Table 4.8: Classifiers and their respective CTDD Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
CTDD	SVM	0	0.00	0.00	0.00	0.00	43.18%
		1	0.43	1.00	0.60		
	RANDOM FOREST	0	0.62	0.52	0.57	0.52	54.54%
		1	0.48	0.58	0.52		
	ANN	0	0.00	0.00	0.00	0.00	38.63%
		1	0.40	0.89	0.56		

Table 4.9: Classifiers and their respective CTDT Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
CTDT	SVM	0	0.87	0.52	0.65	0.52	68.18%
		1	0.59	0.89	0.71		
	RANDOM FOREST	0	0.73	0.76	0.75	0.76	70.45%
		1	0.67	0.63	0.65		
	ANN	0	0.85	0.68	0.76	0.68	75.00%
		1	0.67	0.84	0.74		

Table 4.10: Classifiers and their respective CTriad Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
CTriad	SVM	0	0.59	0.56	0.57	0.55	65.90%
		1	0.7	0.73	0.72		
	RANDOM FOREST	0	0.44	0.61	0.51	0.61	52.27%
		1	0.63	0.46	0.53		
	ANN	0	0.63	0.67	0.65	0.66	70.45%
		1	0.76	0.73	0.75		

Table 4.11: Classifiers and their respective DDE Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
DDE	SVM	0	0.68	0.72	0.70	0.72	75.00%
		1	0.80	0.77	0.78		
	RANDOM FOREST	0	0.62	0.72	0.67	0.72	70.45%
		1	0.78	0.69	0.73		
	ANN	0	0.67	0.78	0.72	0.77	75.00%
		1	0.83	0.73	0.78		

Table 4.12: Classifiers and their respective DPC Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
DPC	SVM	0	0.81	0.59	0.68	0.59	72.72%
		1	0.68	0.86	0.76		
	RANDOM FOREST	0	0.74	0.64	0.68	0.63	70.45%
		1	0.68	0.77	0.72		
	ANN	0	0.81	0.77	0.79	0.77	79.54%
		1	0.78	0.82	0.80		

Table 4.13: Classifiers and their respective GAAC Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
GAAC	SVM	0	0.69	0.38	0.49	0.37	56.81%
		1	0.52	0.8	0.63		
	RANDOM FOREST	0	0.68	0.71	0.69	0.70	65.90%
		1	0.63	0.60	0.62		
	ANN	0	0.67	0.58	0.62	0.58	61.36%
		1	0.57	0.65	0.6		

Table 4.14: Classifiers and their respective GDPC Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
GDPC	SVM	0	0.41	0.39	0.4	0.38	52.27%
		1	0.59	0.62	0.6		
	RANDOM FOREST	0	0.4	0.56	0.47	0.55	47.72%
		1	0.58	0.42	0.49		
	ANN	0	0.53	0.44	0.48	0.44	61.36%
		1	0.66	0.73	0.69		

Table 4.15: Classifiers and their respective Geary Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
Geary	SVM	0	0.75	0.78	0.77	0.78	75.00%
		1	0.75	0.71	0.73		
	RANDOM FOREST	0	0.70	0.70	0.70	0.69	68.18%
		1	0.67	0.67	0.67		
	ANN	0	0.67	0.78	0.72	0.78	68.18%
		1	0.71	0.57	0.63		

Table 4.16: Classifiers and their respective Moran Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
Moran	SVM	0	0.70	0.86	0.78	0.86	75.00%
		1	0.82	0.64	0.72		
	RANDOM FOREST	0	0.67	0.73	0.70	0.72	68.18%
		1	0.70	0.64	0.67		
	ANN	0	0.74	0.77	0.76	0.77	75.00%
		1	0.76	0.73	0.74		

Table 4.17: Classifiers and their respective NMBroto Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
NMBroto	SVM	0	0.62	0.75	0.68	0.75	68.18%
		1	0.75	0.62	0.68		
	RANDOM FOREST	0	0.80	0.80	0.80	0.80	81.81%
		1	0.83	0.83	0.83		
	ANN	0	0.75	0.75	0.75	0.75	77.27%
		1	0.79	0.79	0.79		

Table 4.18: Classifiers and their respective PAAC Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
PAAC	SVM	0	0.50	0.85	0.63	0.85	54.54%
		1	0.70	0.29	0.41		
	RANDOM FOREST	0	0.52	0.60	0.56	0.60	56.81%
		1	0.62	0.54	0.58		
	ANN	0	0.69	0.90	0.78	0.90	77.27%
		1	0.89	0.67	0.76		

Table 4.19: Classifiers and their respective QSOrder Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
QSOrder	SVM	0	0.82	0.32	0.46	0.32	52.27%
		1	0.42	0.88	0.57		
	RANDOM FOREST	0	0.83	0.54	0.65	0.53	63.63%
		1	0.50	0.81	0.62		
	ANN	0	0.88	0.54	0.67	0.53	65.90%
		1	0.52	0.88	0.65		

Table 4.20: Classifiers and their respective SOCNumber Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
SOCNumber	SVM	0	0.00	0.00	0.00	0.00	40.90%
		1	0.41	1.00	0.58		
	RANDOM FOREST	0	0.82	0.69	0.75	0.69	72.72%
		1	0.64	0.78	0.70		
	ANN	0	0.44	0.15	0.23	0.15	38.63%
		1	0.37	0.72	0.49		

Table 4.21: Classifiers and their respective TPC Descriptor, Stages, Precision, Recall, F1 score, Specificity Accuracy

DESCRIPTOR	CLASSIFIERS	GENES	PRECISION	RECALL	F1-SCORE	SPECIFICITY	ACCURACY
TPC	SVM	0	0.75	1.00	0.86	1.00	84.09%
		1	1.00	0.70	0.82		
	RANDOM FOREST	0	0.74	0.95	0.83	0.95	81.81%
		1	0.94	0.70	0.80		
	ANN	0	0.80	0.95	0.87	0.95	86.36%
		1	0.95	0.78	0.86		

4.5 Biomarker

A biomarker is a set of quantitative features that signal the body's biological processes and can be used to detect disease early. It can be used as a valuable indication for tracking the development of MCI to AD. Biomarker identification is critical for each dataset. Thus, we found that PS-1, PS-2 and APP are the most relevant biomarkers in sequence datasets through a profound understanding of the literature.

In addition, cross-sectional dataset feature importance reveals that ADAS-13 is the most relevant biomarker followed by RID. Although, the cognitive score holds the highest relevance, it seems that demographics are also identified as important biomarkers.

In contrast, the longitudinal dataset shows that CDRSB_bl and ADAS-13_bl are significant biomarkers. Although, the cognitive scores outperform in biomarker relevance, the clinical attributes are also imperative biomarkers.

4.6 ROC Curves and AUC Scores

ROC (Receiver Operating Characteristic) Curve is a plot of the FPR (x-axis) versus the TPR (y-axis) for numerous threshold values amid 0.0 and 1.0. It is a statistical estimator which is used to analyze the performance of classification models. Thus, ROC curves are a very robust and intuitive choice for illustrating and analyzing classification models, as well as supporting the selection of cut-off points to maximize phenomenon categorization.

AUC (Area under ROC curve) compares different classifiers and further summarizes the performance of each classifier into a single measure. Additionally, the AUC is the product of integrating all of the points along the curve's course, and it computes sensitivity and specificity at the same time, providing an estimator of a test's performance characteristics and accuracy.

4.6.1 ROC Curve for Longitudinal Data

The results below display the ROC curve extracted from longitudinal data using ML techniques. The ROC represented, consists of four different algorithms. Here, the x-axis signifies a false-positive rate, whereas the y-axis denotes true positive rate. The figure 4.5 clearly indicates that RF and RF+SVM possess better performance followed by ANN and SVM.

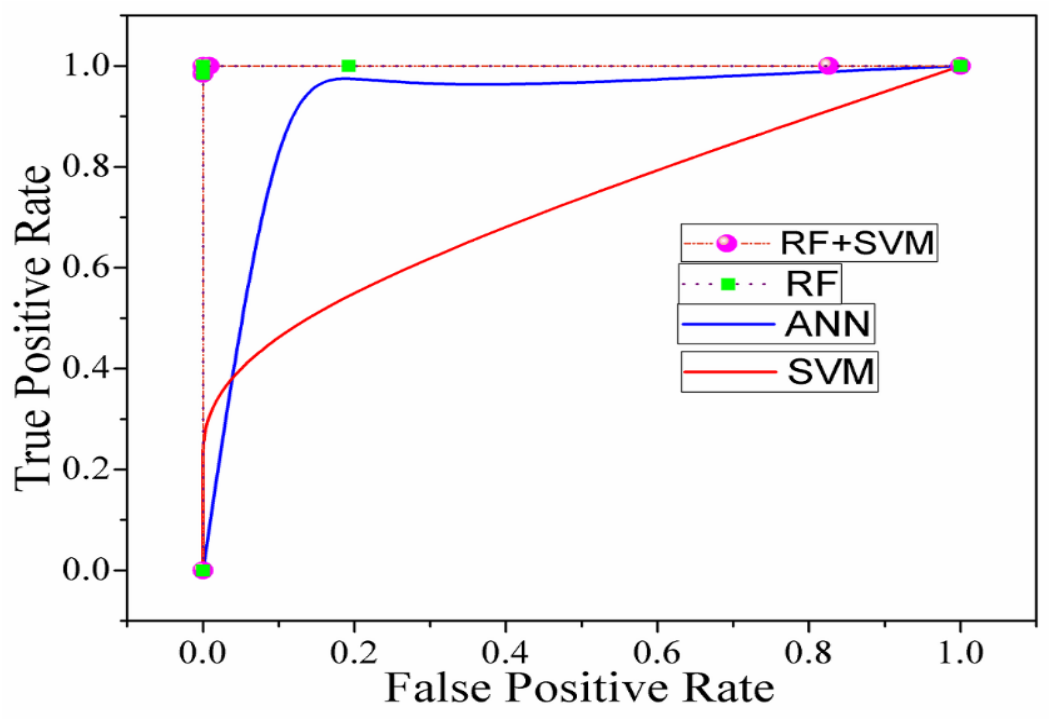

Figure 4.5: ROC for Longitudinal Data.

4.6.2 ROC Curve for Sequence Data using 19 Descriptors

The results below display ROC curves extracted from sequence data using ML techniques. The ROC represented, consists of three different algorithms. Here, the x-axis signifies a false-positive rate, whereas the y-axis denotes true positive rate.

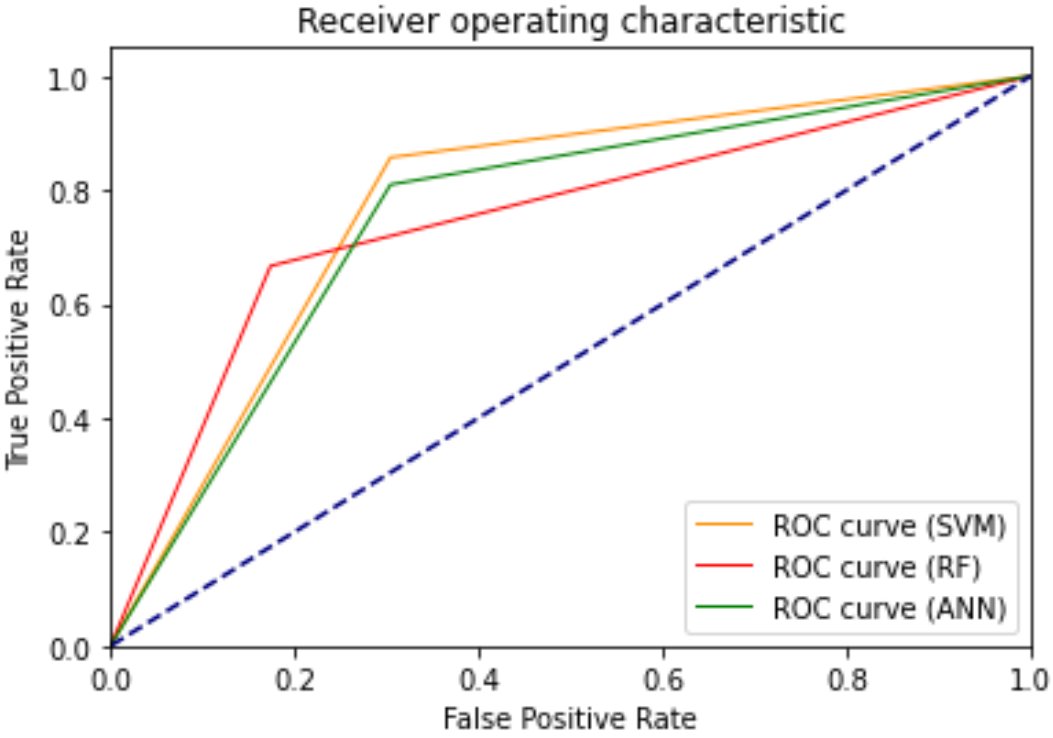

Figure 4.6: ROC Curve for AAC Descriptor

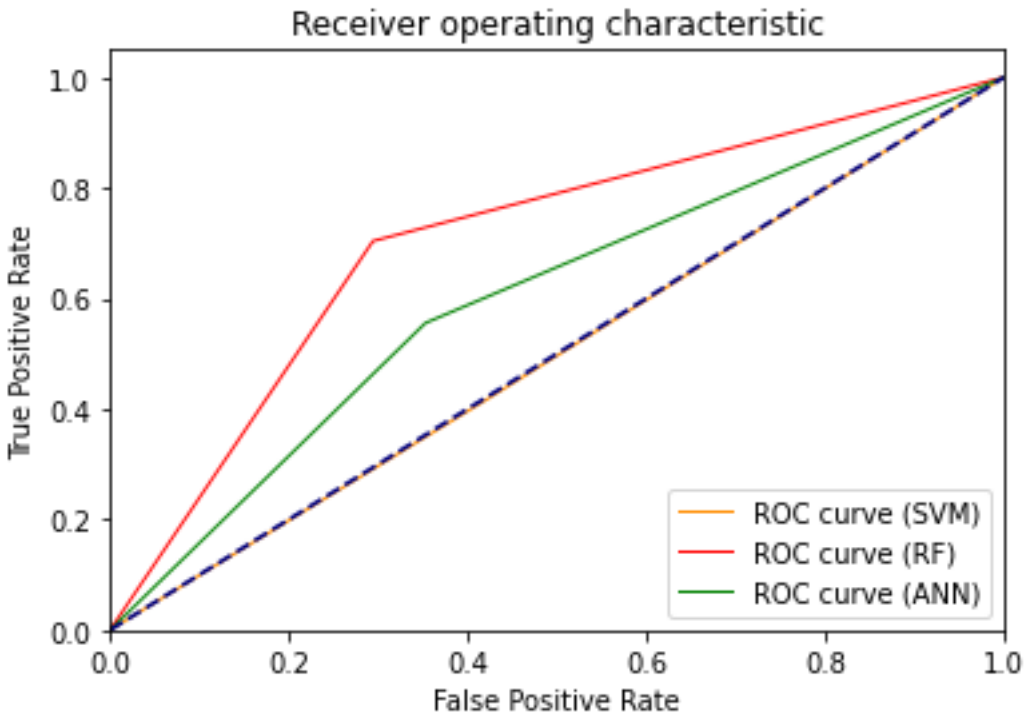

Figure 4.7: ROC Curve for APAAC Descriptor

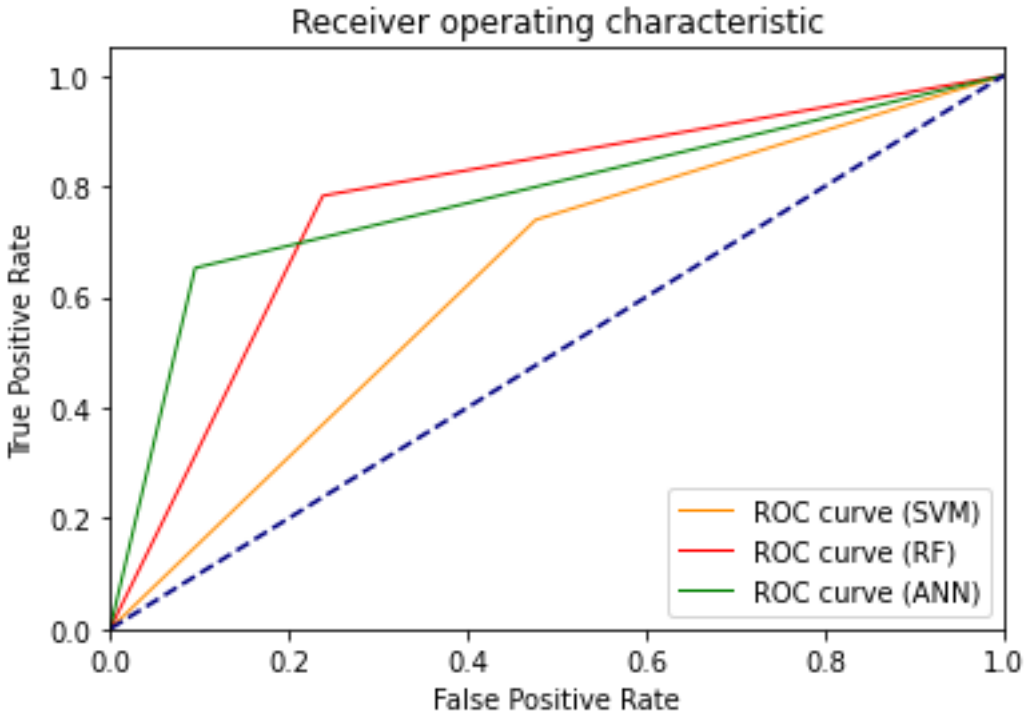

Figure 4.8: ROC Curve for CKSAAGP Descriptor

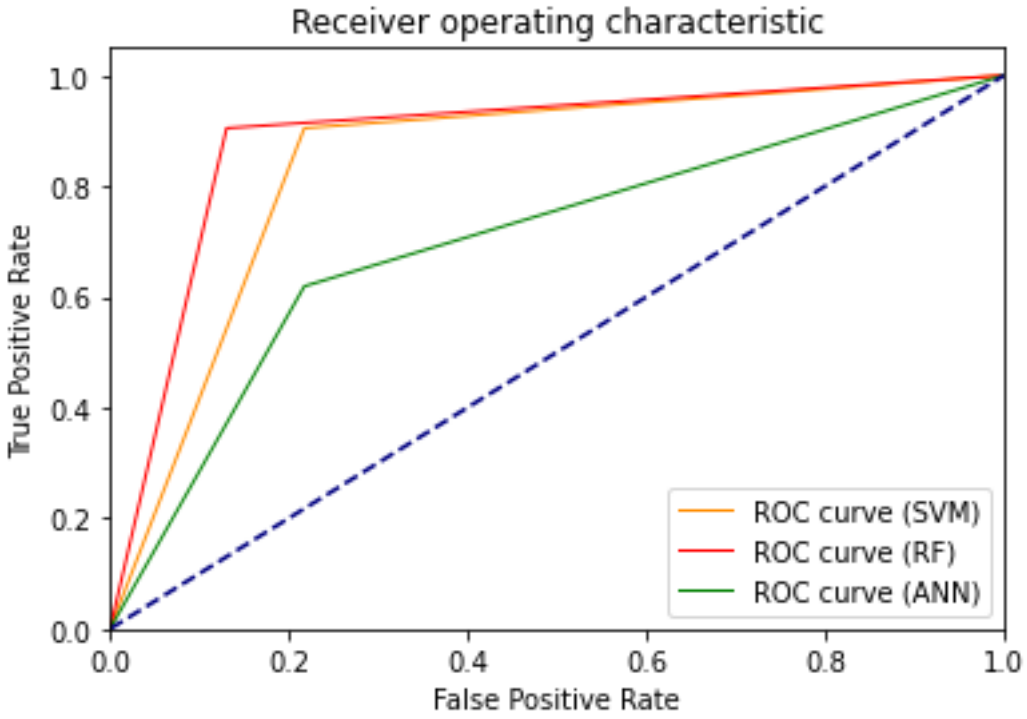

Figure 4.9: ROC Curve for CKSAAP Descriptor

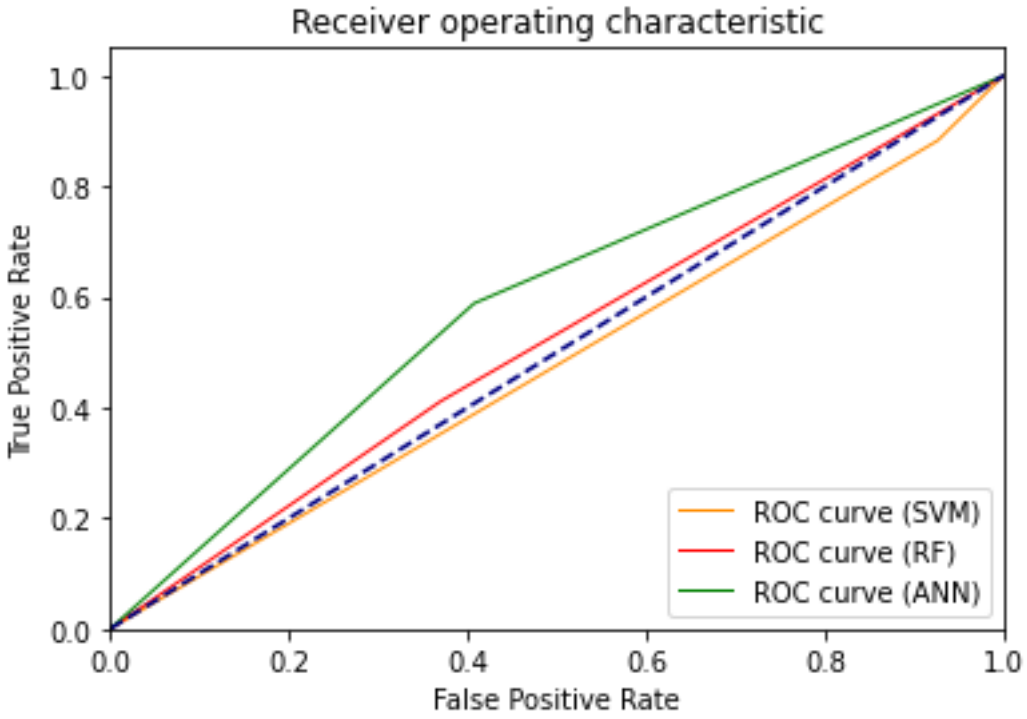

Figure 4.10: ROC Curve for CTDC Descriptor

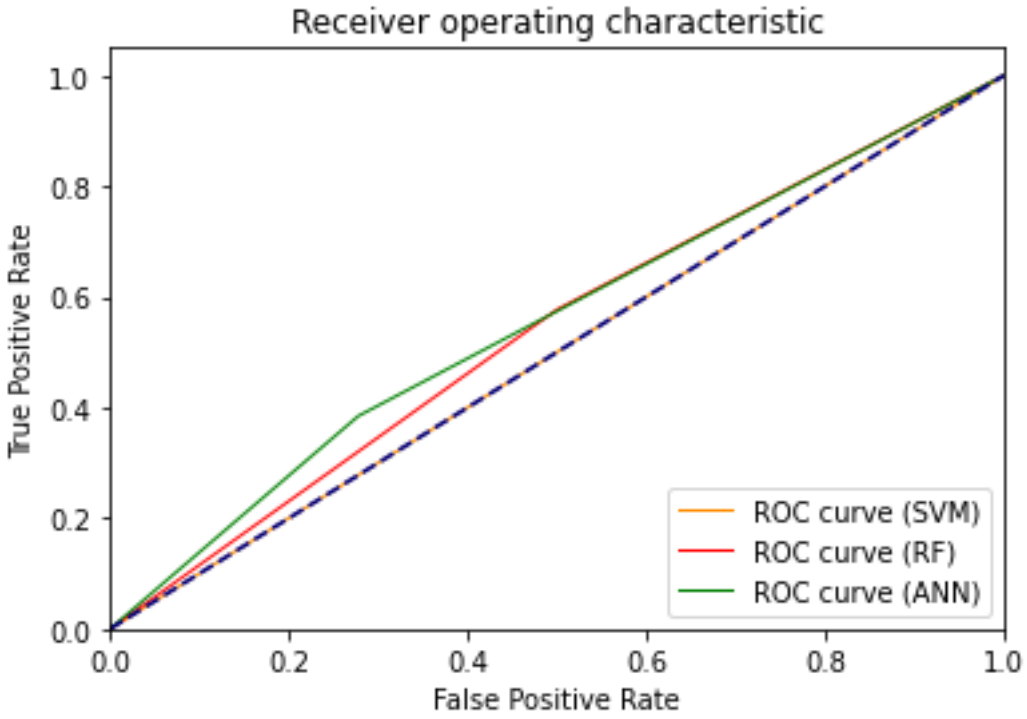

Figure 4.11: ROC Curve for CTDD Descriptor

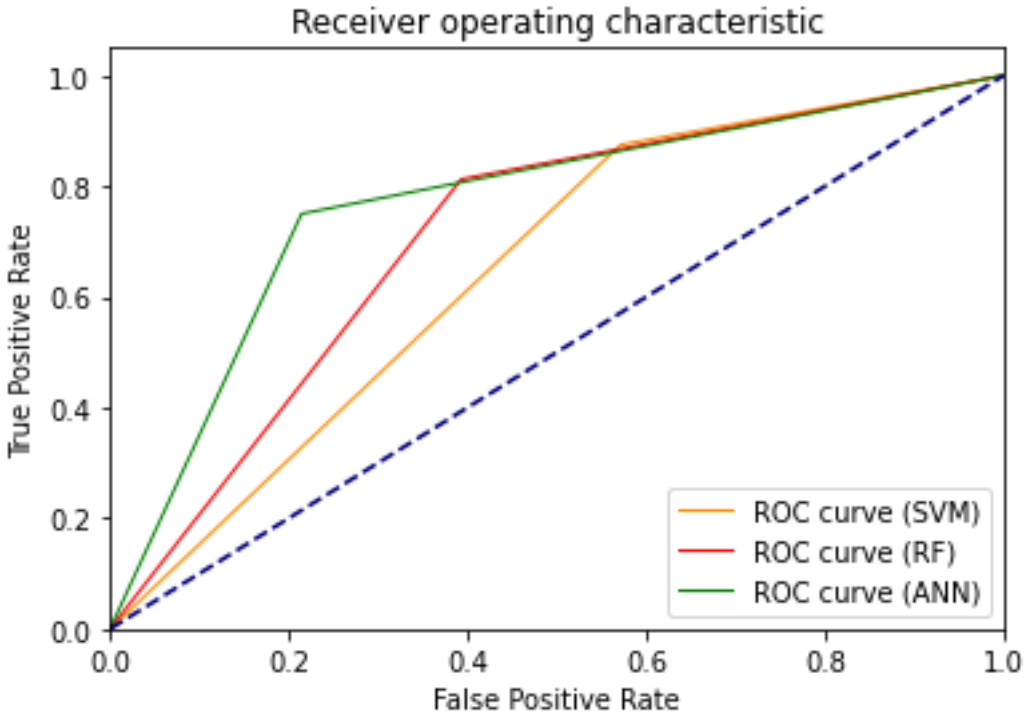

Figure 4.12: ROC Curve for CTDT Descriptor

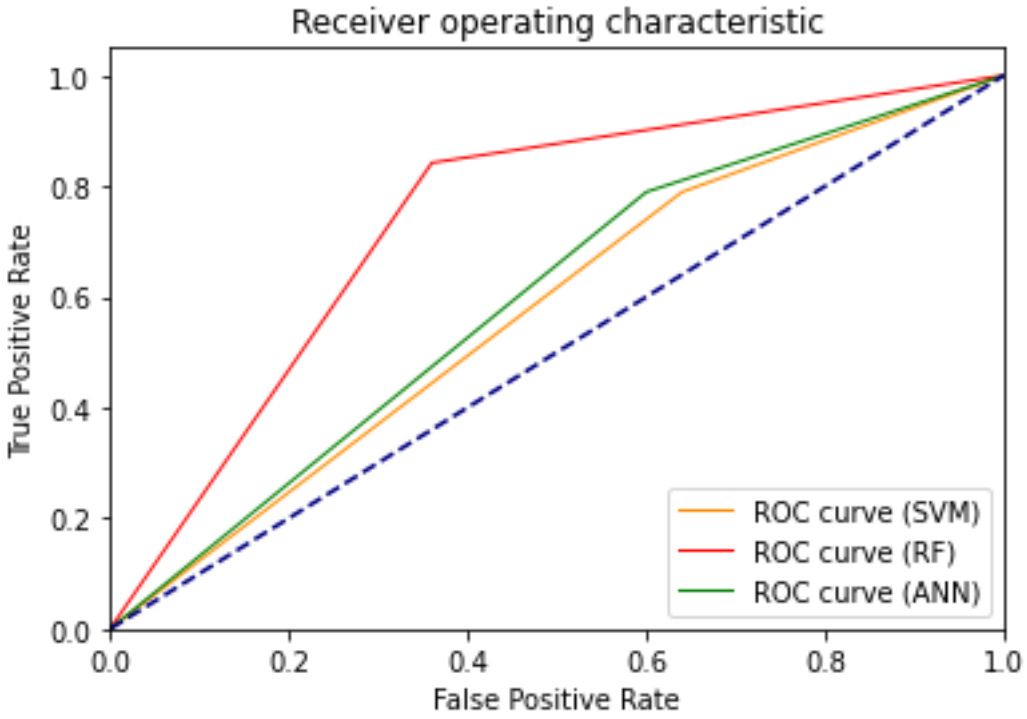

Figure 4.13: ROC Curve for CTriad Descriptor

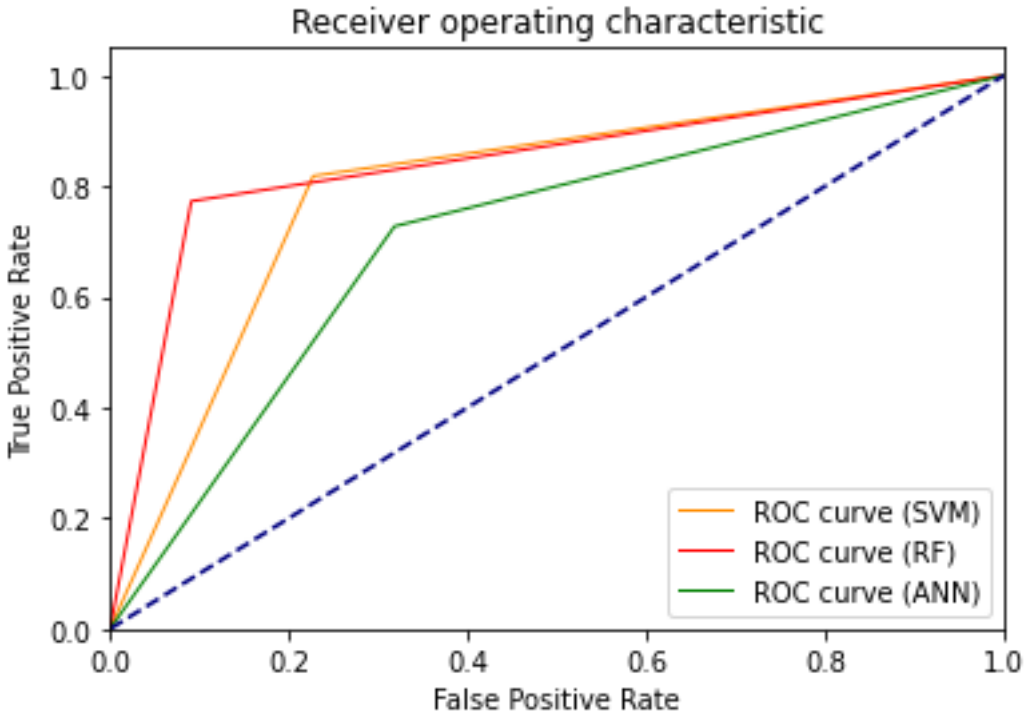

Figure 4.14: ROC Curve for DDE Descriptor

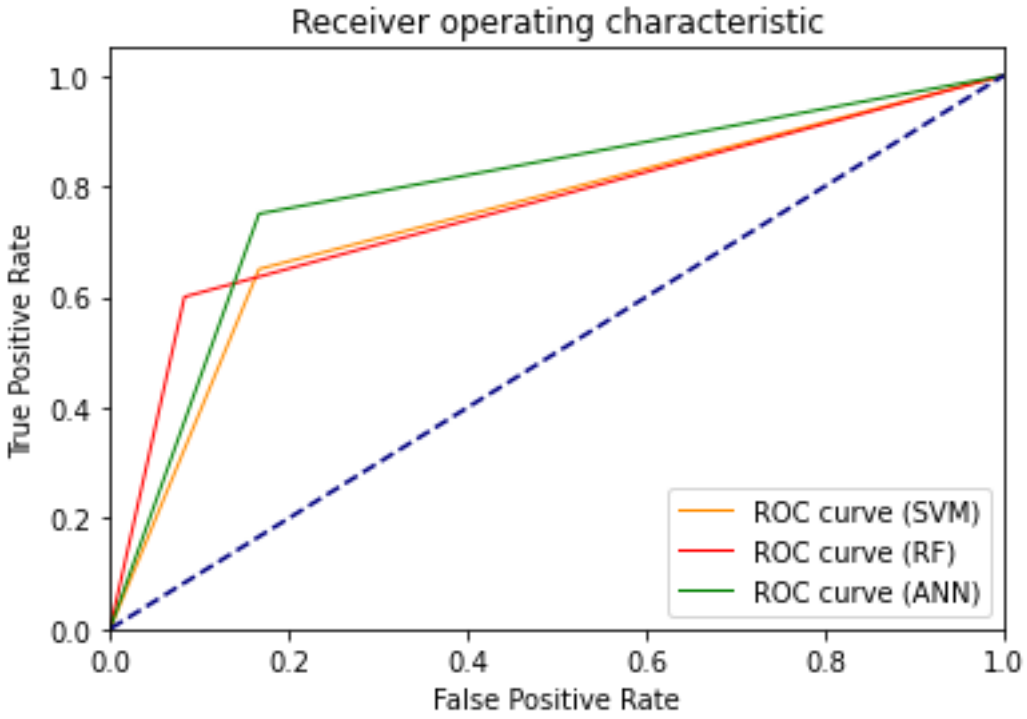

Figure 4.15: ROC Curve for DPC Descriptor

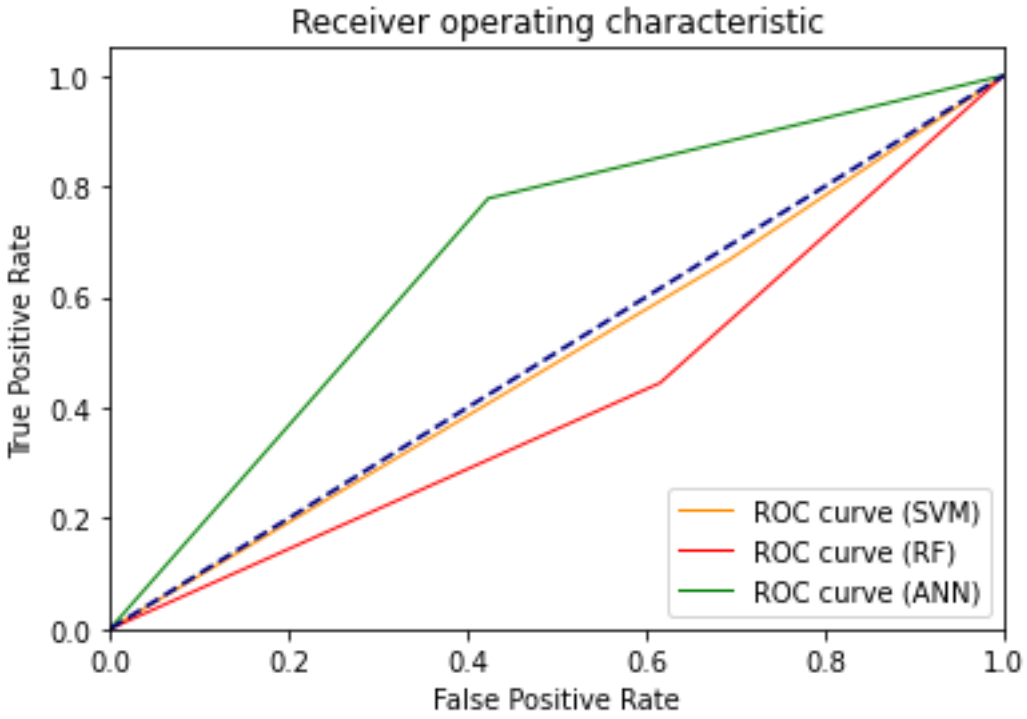

Figure 4.16: ROC Curve for GAAC Descriptor

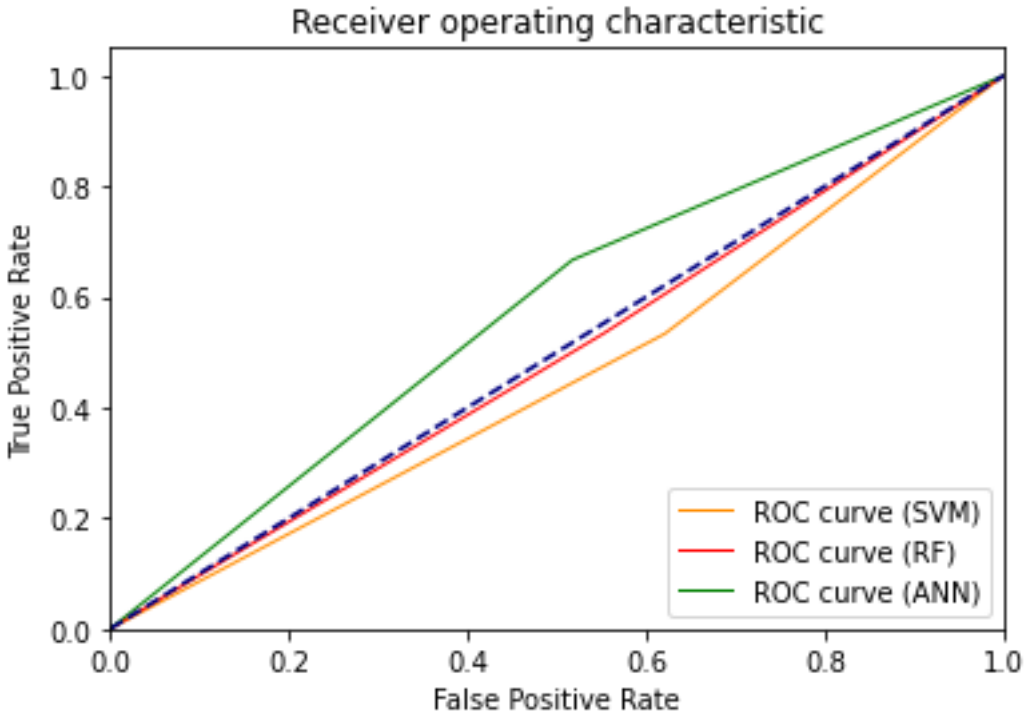

Figure 4.17: ROC Curve for GDPC Descriptor

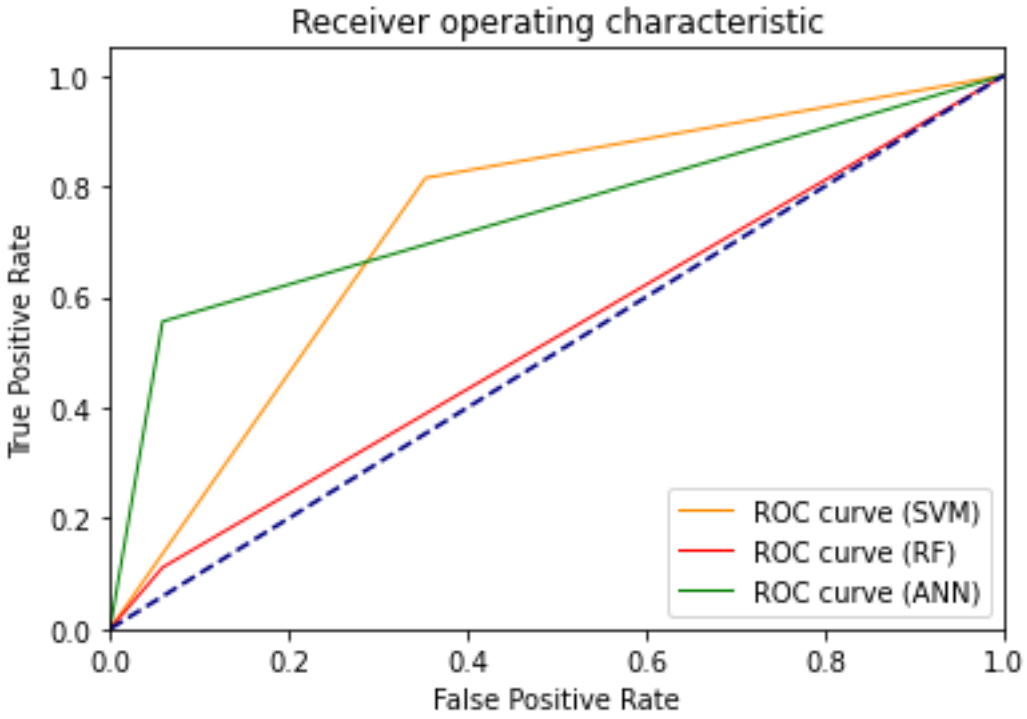

Figure 4.18: ROC Curve for Geary Descriptor

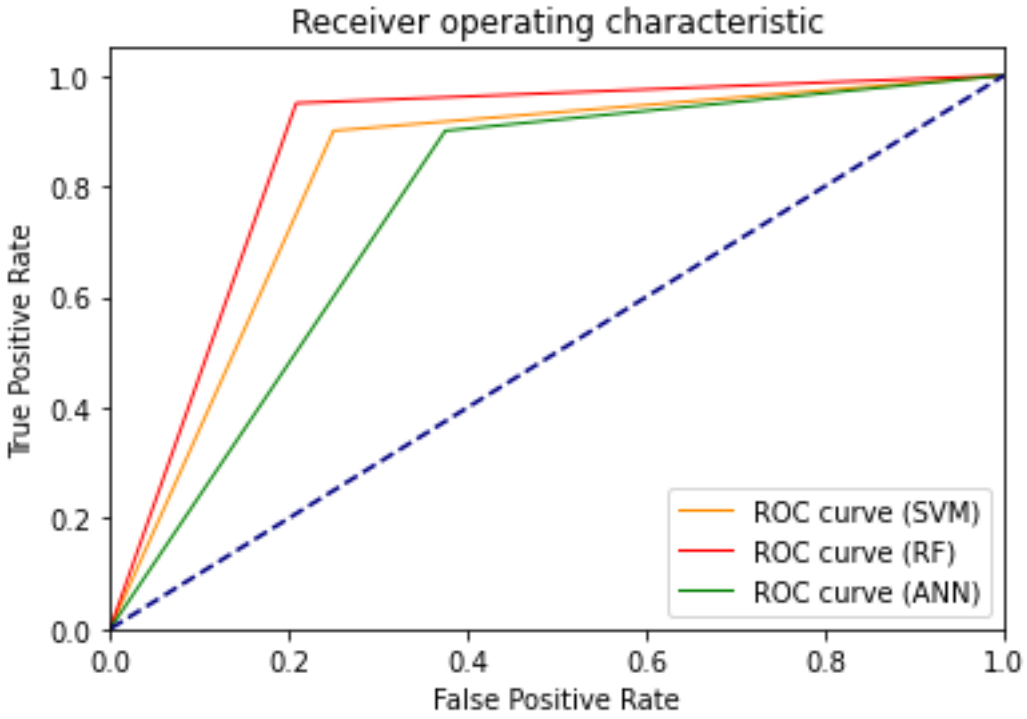

Figure 4.19: ROC Curve for Moran Descriptor

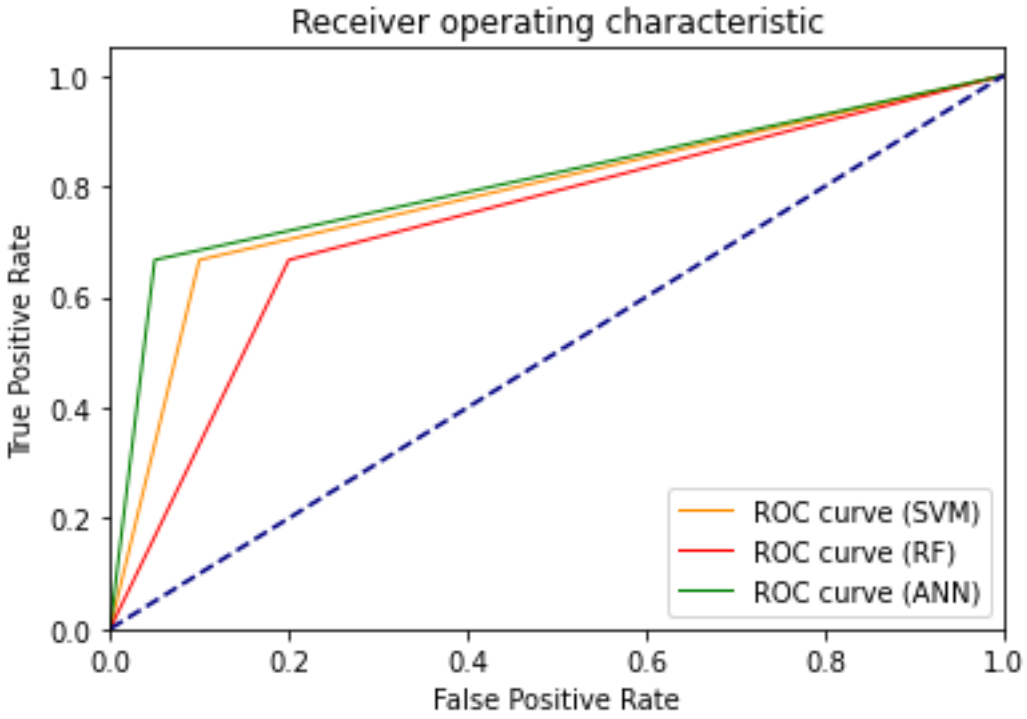

Figure 4.20: ROC Curve for NMBroto Descriptor

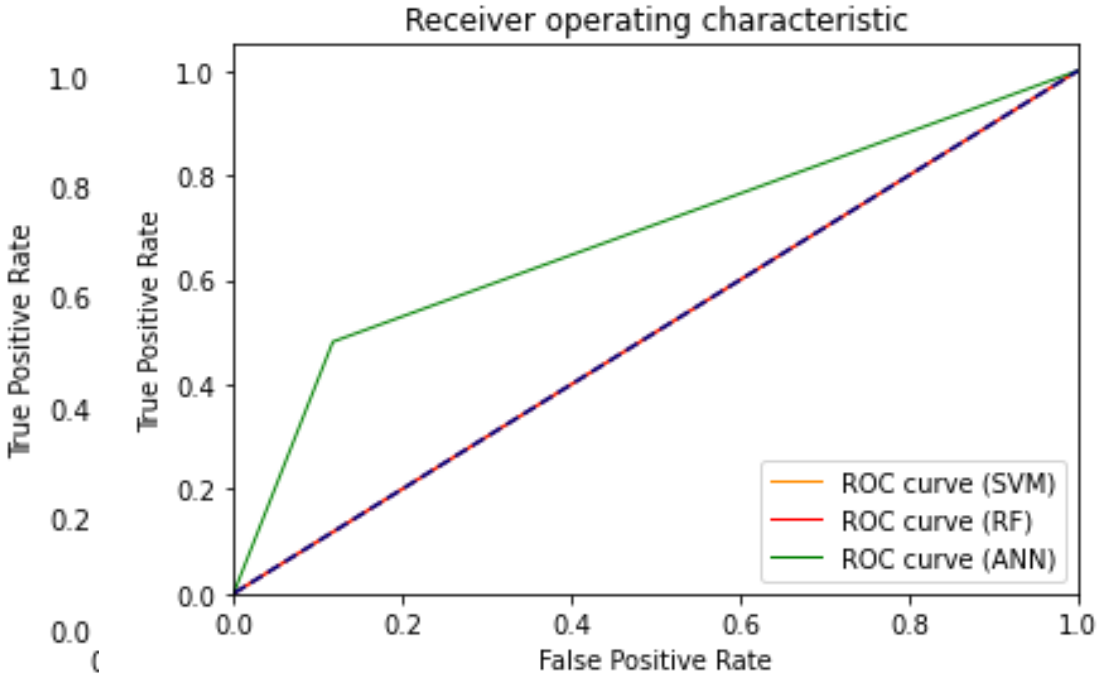

Figure 4.21: ROC Curve for SOCNumber Descriptor

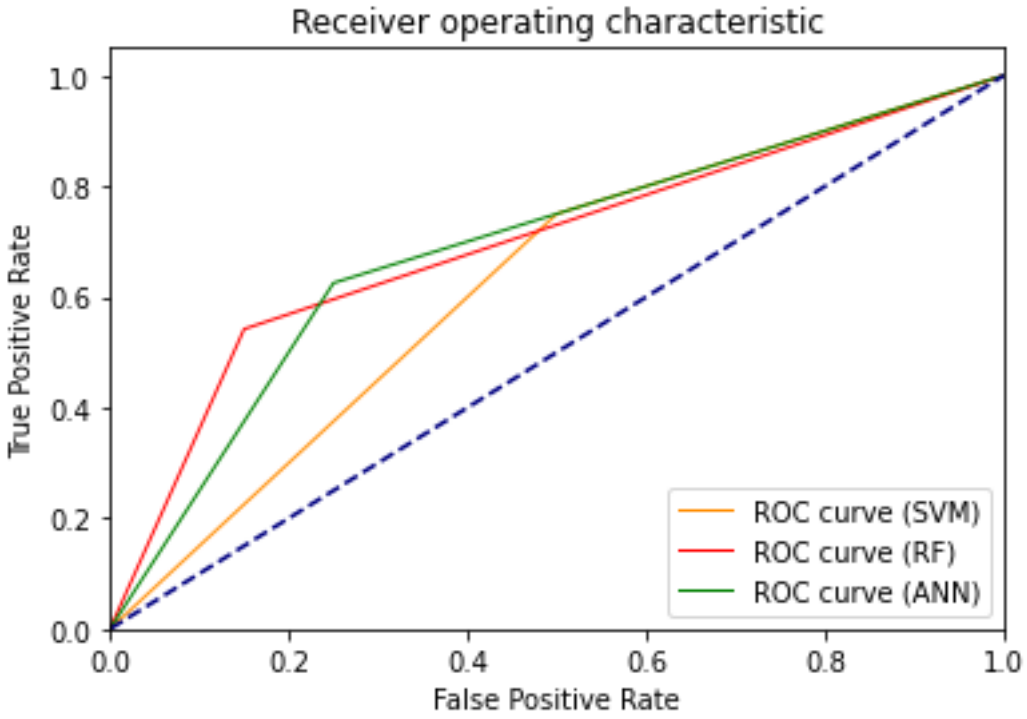

Figure 4.22: ROC Curve for QSOrder Descriptor

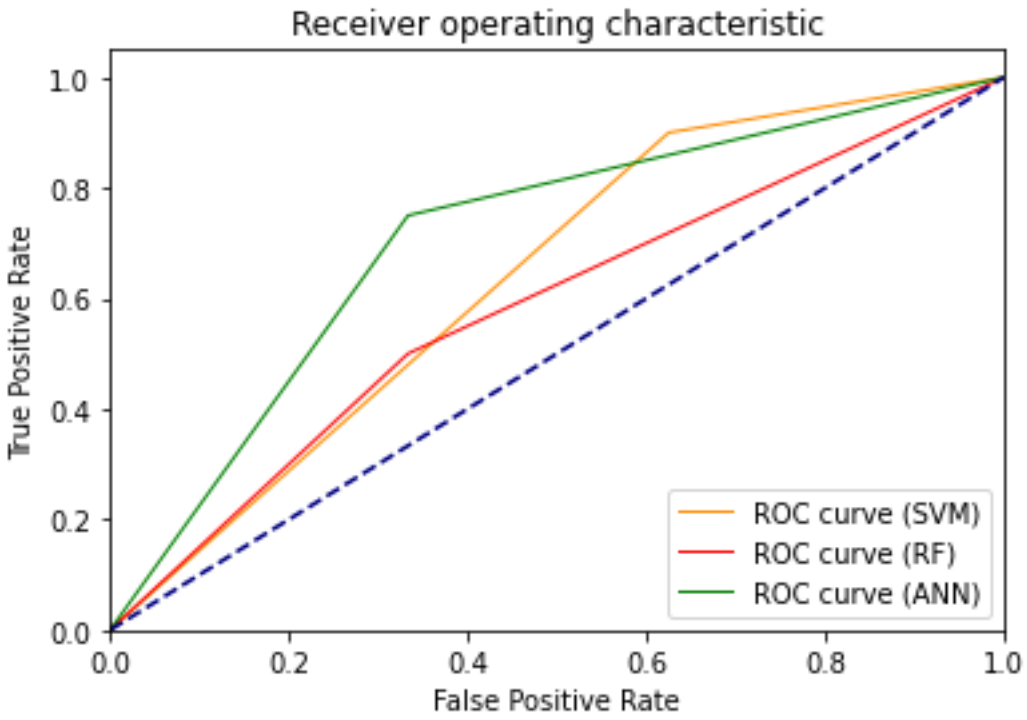

Figure 4.23: ROC Curve for PAAC Descriptor

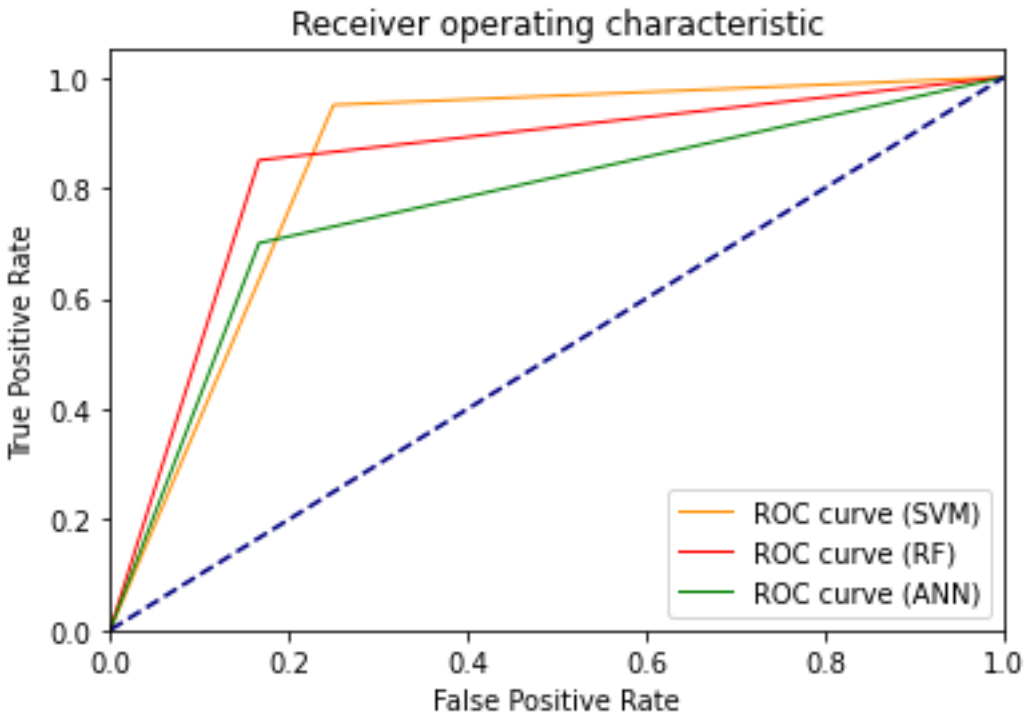

Figure 4.24: ROC Curve for TPC Descriptor

CHAPTER 5

CONCLUSIONS AND FUTURE SCOPE

5.1 General

The conclusion, based on the preceding literature analysis and practical facts, appears of to apply state-of-the-art computational approaches employed in AI to health care and biomedical concerns. Affected individuals will be more likely to benefit from present treatments if new technologies can properly anticipate results five or six years in the future. While additional data is needed, researchers hope that these algorithms will someday find their way into clinics, allowing for faster diagnoses and better care for the millions of people who suffer from this intensifying and irrevocable degenerative disease.

In a nutshell, our strategy is more sensitive than past attempts because it is multimodal i.e. Using various forms of data rather than relying solely on a single predictor type. Primary detection of AD patients may be critical for the administration of disease-modifying medications. As a result, future research should focus on predicting AD before prominent symptoms appear. Traditional slow, subjective, and data-poor evaluations will be reinstated by rapid, objective, and data-rich digital assessments in the future [17]. Moreover, the industry's transition to outcome-based services will be aided by technology. Approximately 85% of respondents believe in the use of intelligent hardware, sensors, and devices. As a result, the industry will be able to transition from selling drugs to selling outcomes.

5.2 Conclusions and Crucial Outcomes of the Study

In this research study, we proposed the analysis of the available data with ML techniques. Longitudinal data and Cross-sectional data involve demographic, clinical, and cognitive attributes, which correspond to baseline and regular features respectively. The evaluation of models demonstrates RF algorithm has the best accuracy and performance compared to other

classifiers in the case of the TADPOLE dataset. The stages of AD with performance are assessed which further aid in the detection of the Alzheimer's appropriate phase. In addition, we came up with the best feature used to predict AD progression. In the case of longitudinal data, CDRSB_bl is the most relevant feature for prediction in the early diagnosis of AD. While, in the cross-sectional data ADAS-13 holds the highest relevance. Considering other factors in feature importance, we concluded that cognitive scores are an imperative biomarker towards early detection.

In contrast, the results of sequence data help to precisely classify AD and NON-AD diseases for a given protein sequence. We conclude that TPC, CKSAAP descriptors are most relevant for the prediction of diseases. The accuracy of TPC and CKSAAP predictors is highest in the ANN classifier and their numerical accuracies are 86.36% and 89.03% respectively. We proposed all these options as potent choices as critical biomarkers for the early detection or prediction of AD.

5.3 Future Scope

The potential for assessing neurodegenerative disease may rest in the devices we each carry in our pockets where our psychological state, cognitive capability and clinical processes are pervasively and consistently tracked by the digital footprint we leave beyond. [17]

Furthermore, the future vision of our research work is to develop an AI-Based application to predict accurate results using multi-modal data which would serve as an augment in the prognosis of the disease. It is anticipated that the proposed pipeline on the provided dataset will be beneficial to the academic and scientific community to explore and manage the world of neurological disorders including AD.

.

REFERENCES

[1] A Mucke, L. (2009). Alzheimer's disease. Nature, 895-897.

[2] https://www.alz.org/alzheimers-dementia/what-is-alzheimers.

[3] H. Rhoads, “Factors that Affect Implementing the MIND diet in an Acute Care Setting to Prevent Alzheimer’s Disease,” p. 40.

[4] https://www.publichealthnotes.com/dementia-alzheimers-disease/alzheimers-brain/

[5] A. Thushara, C. UshaDevi Amma, A. John, and R. Saju, “Multimodal MRI Based Classification and Prediction of Alzheimer’s Disease Using Random Forest Ensemble,” in 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), Cochin, India, Jul. 2020, pp. 249–256. doi: 10.1109/ACCTHPA49271.2020.9213211.

[6] https://tel.archives-ouvertes.fr/tel-02425827v2/document

[7] M. Shahbaz, S. Ali, A. Guergachi, A. Niazi, and A. Umer, “Classification of Alzheimer’s Disease using Machine Learning Techniques:,” in Proceedings of the 8th International Conference on Data Science, Technology and Applications, Prague, Czech Republic, 2019, pp. 296–303. doi: 10.5220/0007949902960303.

[8] L. C. Kourtis, O. B. Regele, J. M. Wright, and G. B. Jones, “Digital biomarkers for Alzheimer’s disease: the mobile/wearable devices opportunity,” npj Digital Med, vol. 2, no. 1, p. 9, Dec. 2019, doi: 10.1038/s41746-019-0084-2.

[9] A. Khan and M. Usman, “Early Diagnosis of Alzheimer’s Disease using Machine Learning Techniques - A Review Paper:,” in Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Lisbon, Portugal, 2015, pp. 380–387. doi: 10.5220/0005615203800387.

[10] J. H. Park et al., “Machine learning prediction of incidence of Alzheimer’s disease using large-scale administrative health data,” npj Digit. Med., vol. 3, no. 1, p. 46, Dec. 2020, doi: 10.1038/s41746-020-0256-0.

[11] A. Kumar, A. Bansal, and T. R. Singh, “ABCD: Alzheimer’s disease Biomarkers Comprehensive Database,” 3 Biotech, vol. 9, no. 10, p. 351, Oct. 2019, doi: 10.1007/s13205-019-1888-0.

[12] G. Vivar, A. Zwergal, N. Navab, and S.-A. Ahmadi, “Multi-modal Disease Classification in Incomplete Datasets Using Geometric Matrix Completion,” in Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities, vol. 11044, D. Stoyanov, Z. Taylor, E. Ferrante, A. V. Dalca, A. Martel, L. Maier-Hein, S. Parisot, A. Sotiras, B. Papiez, M. R. Sabuncu, and L. Shen, Eds. Cham: Springer International Publishing, 2018, pp. 24–31. doi: 10.1007/978-3-030-00689-1_3.Thomas, H.R., Maloney, W.F., Horner, R.M.W., Smith, G.R., Handa, V.K. and Sanders, S.R., 1990. Modeling construction labor productivity. Journal of construction engineering and management, 116(4), pp.705-726.

[13] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, “Artificial Intelligence in Precision Cardiovascular Medicine,” Journal of the American College of Cardiology, vol. 69, no. 21, pp. 2657–2664, May 2017, doi: 10.1016/j.jacc.2017.03.571.

[14] A. R. Allam, K. K. Reddi, and H. Thota, “Bioinformatic Analysis of Alzheimer’s Disease Using Functional Protein Sequences,” J Proteomics Bioinform, vol. 01, no. 01, pp. 036–042, Apr. 2008, doi: 10.4172/jpb.1000007.

[15] H. T. Gorji, T. T. Khoei, and N. Kaabouch, “Biomarkers Selection Toward Early Detection of Alzheimer’s Disease,” p. 8.

[16] A. Ramesh, C. Kambhampati, J. Monson, and P. Drew, “Artificial intelligence in medicine,” Ann R Coll Surg Engl, vol. 86, no. 5, pp. 334–338, Sep. 2004, doi: 10.1308/147870804290.

[17] W. H. W. Ishak and F. Siraj, “ARTIFICIAL INTELLIGENCE IN MEDICAL APPLICATION: AN EXPLORATION,” p. 10.

[18] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, “Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine,” Database, vol. 2020, p. baaa010, Jan. 2020, doi: 10.1093/database/baaa010.

[19] Y.-K. Chan, Y.-F. Chen, T. Pham, W. Chang, and M.-Y. Hsieh, “Artificial Intelligence in Medical Applications,” Journal of Healthcare Engineering, vol. 2018, pp. 1–2, Jul. 2018, doi: 10.1155/2018/4827875.

[20] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, “Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine,” Database, vol. 2020, p. baaa010, Jan. 2020, doi: 10.1093/database/baaa010.

[21] A. Kumar and T. R. Singh, “A New Decision Tree to Solve the Puzzle of Alzheimer’s Disease Pathogenesis Through Standard Diagnosis Scoring System,” Interdiscip Sci Comput Life Sci, vol. 9, no. 1, pp. 107–115, Mar. 2017, doi: 10.1007/s12539-016-0144-0.

[22] https://tadpole.grand-challenge.org/Data/

[23] G. Martí-Juan, G. Sanroma-Guell, and G. Piella, “A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer’s disease,” Computer Methods and Programs in Biomedicine, vol. 189, p. 105348, Jun. 2020, doi: 10.1016/j.cmpb.2020.105348.

[24] M. Uddin, Y. Wang, and M. Woodbury-Smith, “Artificial intelligence for precision medicine in neurodevelopmental disorders,” npj Digit. Med., vol. 2, no. 1, p. 112, Dec. 2019, doi: 10.1038/s41746-019-0191-0.

[25] S. Naganandhini and P. Shanmugavadivu, “Effective Diagnosis of Alzheimer’s Disease using Modified Decision Tree Classifier,” Procedia Computer Science, vol. 165, pp. 548–555, 2019, doi: 10.1016/j.procs.2020.01.049.

[26] Khan and M. Usman, “Early Diagnosis of Alzheimer’s Disease using Machine Learning Techniques - A Review Paper:,” in Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Lisbon, Portugal, 2015, pp. 380–387. doi: 10.5220/0005615203800387.

[27] D. AL-Dlaeen and A. Alashqur, “Using decision tree classification to assist in the prediction of Alzheimer’s disease,” in 2014 6th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, Mar. 2014, pp. 122–126. doi: 10.1109/CSIT.2014.6805989.

[28] A. Kumar et al., “Computational and In-Vitro Validation of Natural Molecules as Potential Acetylcholinesterase Inhibitors and Neuroprotective Agents,” CAR, vol. 16, no. 2, pp. 116–127, Feb. 2019, doi: 10.2174/1567205016666181212155147.

[29] R. Shukla and T. R. Singh, “Neuroinflammation in Alzheimer’s Disease: Current Therapeutic Approaches and Recent Progress,” vol. 4, no. 5, p. 5.

[30] R. Shukla, N. S. Munjal, and T. R. Singh, “Identification of novel small molecules against GSK3β for Alzheimer’s disease using chemoinformatics approach,” Journal of Molecular Graphics and Modelling, vol. 91, pp. 91–104, Sep. 2019, doi: 10.1016/j.jmgm.2019.06.008.

[31] R. Shukla and T. R. Singh, “Identification of small molecules against cyclin dependent kinase-5 using chemoinformatics approach for Alzheimer’s disease and other tauopathies,” Journal of Biomolecular Structure and Dynamics, pp. 1–13, Nov. 2020, doi: 10.1080/07391102.2020.1844050.

[32] A. Bansal and T. R. Singh, “Epigenome-Wide DNA Methylation and Histone Modification of Alzheimer’s Disease,” in Computational Epigenetics and Diseases, Elsevier, 2019, pp. 131–148. doi: 10.1016/B978-0-12-814513-5.00009-X.

[33] P. P. Panigrahi and T. R. Singh, “Computational studies on Alzheimer’s disease associated pathways and regulatory patterns using microarray gene expression and network data: Revealed association with aging and other diseases,” Journal of Theoretical Biology, vol. 334, pp. 109–121, Oct. 2013, doi: 10.1016/j.jtbi.2013.06.013.

[34] A. Kumar and T. R. Singh, “Analysis for biological network properties of Alzheimer’s disease associated gene set by enrichment and topological examinations,” p. 9.

[35] R. Shukla and T. R. Singh, “Virtual screening, pharmacokinetics, molecular dynamics and binding free energy analysis for small natural molecules against cyclin-dependent kinase 5 for Alzheimer’s disease,” Journal of Biomolecular Structure and Dynamics, vol. 38, no. 1, pp. 248–262, Jan. 2020, doi: 10.1080/07391102.2019.1571947.

APPENDIX

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
dataset =pd.read_csv("C:/Users/Hp/Desktop/Sequence/TPC.csv")
dataset
```

Out[2]:

	AAA	AAC	AAD	AAE	AAF	AAG	AAH	AAI	AAK	AAL	...	YYN	YYP	YYQ	YYR	YYS	YYT	YYV	YYW
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000662	0.0	0.0	0.
1	0.000812	0.000406	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000406	0.000812	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.
2	0.000000	0.000000	0.000742	0.000000	0.000742	0.000742	0.000742	0.001484	0.000742	0.000000	...	0.0	0.0	0.0	0.0	0.000742	0.0	0.0	0.
3	0.000000	0.000000	0.000000	0.000000	0.001908	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.
4	0.000000	0.001473	0.000000	0.001473	0.001473	0.000000	0.000000	0.000000	0.000000	0.001473	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
215	0.002882	0.000000	0.000000	0.004323	0.000000	0.001441	0.000000	0.001441	0.000000	0.002882	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.
216	0.000000	0.000000	0.000000	0.000000	0.000000	0.001629	0.000000	0.000000	0.000000	0.001629	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.
217	0.000000	0.000000	0.000000	0.000000	0.000000	0.001903	0.000000	0.000000	0.000000	0.000951	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.
218	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.
219	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.

220 rows × 8001 columns

In [3]:
```
dataset.head()
```

Out[3]:

	AAA	AAC	AAD	AAE	AAF	AAG	AAH	AAI	AAK	AAL	...	YYN	YYP	YYQ	YYR	YYS	YYT	YYV	YYW
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000662	0.0	0.0	0.0
1	0.000812	0.000406	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000406	0.000812	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
2	0.000000	0.000000	0.000742	0.000000	0.000742	0.000742	0.000742	0.001484	0.000742	0.000000	...	0.0	0.0	0.0	0.0	0.000742	0.0	0.0	0.0
3	0.000000	0.000000	0.000000	0.000000	0.001908	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
4	0.000000	0.001473	0.000000	0.001473	0.001473	0.000000	0.000000	0.000000	0.000000	0.001473	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0

5 rows × 8001 columns

In [4]:
```
dataset.Label.value_counts()
```

Out[4]:
```
0    110
1    110
Name: Label, dtype: int64
```

In [5]:
```
Y= dataset.Label
X=dataset.drop('Label',axis=1)
```

In [6]:
```
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.metrics import accuracy_score
from sklearn.model_selection import KFold
```

In [7]:
```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2)
print('X_train:', X_train.shape, '\ty_train:', Y_train.shape)
print('X_test:', X_test.shape, '\ty_test:', Y_test.shape)
```

```
X_train: (176, 8000)    y_train: (176,)
X_test: (44, 8000)      y_test: (44,)
```

In [8]:
```
df = pd.DataFrame(data=X_train)
df
```

Out[8]:

	AAA	AAC	AAD	AAE	AAF	AAG	AAH	AAI	AAK	AAL	...	YYM	YYN	YYP	YYQ	YYR	YYS	YYT	YYV	Y
101	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	
2	0.000000	0.0	0.000742	0.000000	0.000742	0.000742	0.000742	0.001484	0.000742	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000742	0.0	0.0	
110	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.002646	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	
60	0.016529	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.008264	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	
109	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
196	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.004525	0.000000	0.004525	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	
31	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	
143	0.000000	0.0	0.000000	0.000000	0.000000	0.010989	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	
215	0.002882	0.0	0.000000	0.004323	0.000000	0.001441	0.000000	0.001441	0.000000	0.002882	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	
212	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.001185	0.001185	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	

176 rows × 8000 columns

In [9]:
```
from sklearn.preprocessing import LabelEncoder
for column in X.columns:
    le = LabelEncoder()
    X[column] = le.fit_transform(X[column])
```

In [10]:
```
from sklearn.preprocessing import MinMaxScaler as Scaler
scaler = Scaler()
scaler.fit(X_train)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

In [11]:
```
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', random_state = 0)
classifier.fit(X_train, Y_train)
```

Out[11]:
```
SVC(random_state=0)
```

In [12]:
```
Y_pred = classifier.predict(X_test)
```

In [13]:
```
k = 5
kf = KFold(n_splits=k, random_state=None)
accuracies = cross_val_score(classifier,  X_train,  Y_train, cv = kf)
```

In [14]:
```
print("Avg accuracy: {}".format(accuracies.mean()))
```

```
Avg accuracy: 0.7328571428571429
```

In [15]:
```
accuracies.std()
```

Out[15]:
```
0.05635166549518864
```

In [16]:
```
def calculate_sensitivity_specificity(Y_test, Y_pred):
    # Note: More parameters are defined than necessary.
    # This would allow return of other measures other than sensitivity and specificity

    # Get true/false for whether a breach actually occurred
    actual_pos = Y_test == 1
    actual_neg = Y_test == 0

    # Get true and false test (true test match actual, false tests differ from actual)
    true_pos = (Y_pred == 1) & (actual_pos)
    false_pos = (Y_pred == 1) & (actual_neg)
    true_neg = (Y_pred == 0) & (actual_neg)
    false_neg = (Y_pred == 0) & (actual_pos)
```

```
    # Calculate accuracy
    accuracy = np.mean(Y_pred == Y_test)

    # Calculate sensitivity and specificity
    sensitivity = np.sum(true_pos) / np.sum(actual_pos)
    specificity = np.sum(true_neg) / np.sum(actual_neg)

    return sensitivity, specificity, accuracy
```

In [17]:
```
sensitivity, specificity, accuracy = calculate_sensitivity_specificity(Y_test, Y_pred)
print ('Sensitivity:', sensitivity)
print ('Specificity:', specificity)
print ('Accuracy:', accuracy)
```

```
Sensitivity: 0.95
Specificity: 0.75
Accuracy: 0.8409090909090909
```

In [18]:
```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
print("Confusion Matrix: \n")
print(confusion_matrix(Y_test,Y_pred))
print("\n")
print("Classification Report: \n")
print(classification_report(Y_test,Y_pred))
print("\n")
print("Accuracy Score: \n")
print(accuracy_score(Y_test, Y_pred))
```

```
Confusion Matrix:

[[18  6]
 [ 1 19]]


Classification Report:
```

```
Classification Report:

              precision    recall  f1-score   support

           0       0.95      0.75      0.84        24
           1       0.76      0.95      0.84        20

    accuracy                           0.84        44
   macro avg       0.85      0.85      0.84        44
weighted avg       0.86      0.84      0.84        44


Accuracy Score:

0.8409090909090909
```

In [19]:
```
from sklearn.ensemble import RandomForestClassifier
RF=RandomForestClassifier()
```

In [20]:
```
RF = RandomForestClassifier(n_estimators=20, random_state=0)
RF.fit(X_train, Y_train)
Y_predR = RF.predict(X_test)
```

In [21]:
```
k = 5
kf = KFold(n_splits=k, random_state=None)
accuracies = cross_val_score(classifier,  X_train,  Y_train, cv = kf)
```

In [22]:
```
print("Avg accuracy: {}".format(accuracies.mean()))
```

```
Avg accuracy: 0.7328571428571429
```

In [23]:
```
accuracies.std()
```

Out[23]:
```
0.05635166549518864
```

In [24]:
```
def calculate_sensitivity_specificity(Y_test, Y_predR):
    # Note: More parameters are defined than necessary.
    # This would allow return of other measures other than sensitivity and specificity

    # Get true/false for whether a breach actually occurred
    actual_pos = Y_test == 1
    actual_neg = Y_test == 0

    # Get true and false test (true test match actual, false tests differ from actual)
    true_pos = (Y_predR == 1) & (actual_pos)
    false_pos = (Y_predR == 1) & (actual_neg)
    true_neg = (Y_predR == 0) & (actual_neg)
    false_neg = (Y_predR == 0) & (actual_pos)

    # Calculate accuracy
    accuracy = np.mean(Y_predR == Y_test)

    # Calculate sensitivity and specificity
    sensitivity = np.sum(true_pos) / np.sum(actual_pos)
    specificity = np.sum(true_neg) / np.sum(actual_neg)

    return sensitivity, specificity, accuracy
```

In [25]:
```
sensitivity, specificity, accuracy = calculate_sensitivity_specificity(Y_test, Y_predR)
print ('Sensitivity:', sensitivity)
print ('Specificity:', specificity)
print ('Accuracy:', accuracy)
```

```
Sensitivity: 0.7
Specificity: 0.8333333333333334
Accuracy: 0.7727272727272727
```

In [26]:
```
from sklearn.metrics import confusion_matrix
print("Confusion Matrix: \n")
print(confusion_matrix(Y_test,Y_predR))
print("\n")
print("Classification Report: \n")
```

```
Confusion Matrix:

[[20  4]
 [ 6 14]]


Classification Report:

              precision    recall  f1-score   support

           0       0.77      0.83      0.80        24
           1       0.78      0.70      0.74        20

    accuracy                           0.77        44
   macro avg       0.77      0.77      0.77        44
weighted avg       0.77      0.77      0.77        44



Accuracy Score:

0.7727272727272727
```

In [27]:
```
from sklearn.neural_network import MLPClassifier
mlp=MLPClassifier(hidden_layer_sizes=(90,90,90))
mlp.fit(X_train , Y_train)
Y_predM=mlp.predict(X_test)
```

In [28]:
```
k = 5
kf = KFold(n_splits=k, random_state=None)
accuracies = cross_val_score(classifier,  X_train,  Y_train, cv = kf)
```

In [29]:
```
print("Avg accuracy: {}".format(accuracies.mean()))
```

```
Avg accuracy: 0.7328571428571429
```

In [31]:
```
def calculate_sensitivity_specificity(Y_test, Y_predM):
    # Note: More parameters are defined than necessary.
    # This would allow return of other measures other than sensitivity and specificity

    # Get true/false for whether a breach actually occurred
    actual_pos = Y_test == 1
    actual_neg = Y_test == 0

    # Get true and false test (true test match actual, false tests differ from actual)
    true_pos = (Y_predM == 1) & (actual_pos)
    false_pos = (Y_predM == 1) & (actual_neg)
    true_neg = (Y_predM == 0) & (actual_neg)
    false_neg = (Y_predM == 0) & (actual_pos)

    # Calculate accuracy
    accuracy = np.mean(Y_predM == Y_test)

    # Calculate sensitivity and specificity
    sensitivity = np.sum(true_pos) / np.sum(actual_pos)
    specificity = np.sum(true_neg) / np.sum(actual_neg)

    return sensitivity, specificity, accuracy
```

In [32]:
```
sensitivity, specificity, accuracy = calculate_sensitivity_specificity(Y_test, Y_predM)
print ('Sensitivity:', sensitivity)
print ('Specificity:', specificity)
print ('Accuracy:', accuracy)
```

```
Sensitivity: 0.85
Specificity: 0.8333333333333334
Accuracy: 0.8409090909090909
```

In [33]:
```
from sklearn.metrics import confusion_matrix
print("Confusion Matrix: \n")
print(confusion_matrix(Y_test,Y_predM))
print("\n")
```

```
Confusion Matrix:

[[20  4]
 [ 3 17]]


Classification Report:

              precision    recall  f1-score   support

           0       0.87      0.83      0.85        24
           1       0.81      0.85      0.83        20

    accuracy                           0.84        44
   macro avg       0.84      0.84      0.84        44
weighted avg       0.84      0.84      0.84        44


Accuracy Score:

0.8409090909090909
```

In [34]:
```
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc
fpr, tpr, thresholds = roc_curve(Y_test,Y_pred)
roc_auc = auc(fpr, tpr)

fpr1, tpr1, thresholds = roc_curve(Y_test,Y_predM)
roc_auc1 = auc(fpr1, tpr1)

fpr2, tpr2, thresholds = roc_curve(Y_test,Y_predR)
roc_auc2 = auc(fpr2, tpr2)

plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=1, label='ROC curve (SVM)' % roc_auc)
```

```
plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=1, label='ROC curve (SVM)' % roc_auc)
plt.plot(fpr1, tpr1, color='red', lw=1, label='ROC curve (RF)' % roc_auc1)import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc
fpr, tpr, thresholds = roc_curve(Y_test,Y_pred)
roc_auc = auc(fpr, tpr)

fpr1, tpr1, thresholds = roc_curve(Y_test,Y_predM)
roc_auc1 = auc(fpr1, tpr1)

fpr2, tpr2, thresholds = roc_curve(Y_test,Y_predR)
roc_auc2 = auc(fpr2, tpr2)

plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=1, label='ROC curve (SVM)' % roc_auc)
plt.plot(fpr1, tpr1, color='red', lw=1, label='ROC curve (RF)' % roc_auc1)
plt.plot(fpr2, tpr2, color='green', lw=1, label='ROC curve (ANN)' % roc_auc2)
plt.plot([0, 1], [0, 1], color='navy', linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.show()import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc
fpr, tpr, thresholds = roc_curve(Y_test,Y_pred)
roc_auc = auc(fpr, tpr)

fpr1, tpr1, thresholds = roc_curve(Y_test,Y_predM)
roc_auc1 = auc(fpr1, tpr1)

fpr2, tpr2, thresholds = roc_curve(Y_test,Y_predR)
roc_auc2 = auc(fpr2, tpr2)

plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=1, label='ROC curve (SVM)' % roc_auc)
```

```
plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=1, label='ROC curve (SVM)' % roc_auc)
plt.plot(fpr1, tpr1, color='red', lw=1, label='ROC curve (RF)' % roc_auc1)
plt.plot(fpr2, tpr2, color='green', lw=1, label='ROC curve (ANN)' % roc_auc2)
plt.plot([0, 1], [0, 1], color='navy', linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.show()
plt.plot(fpr2, tpr2, color='green', lw=1, label='ROC curve (ANN)' % roc_auc2)
plt.plot([0, 1], [0, 1], color='navy', linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.show()
```