Title

A Comparative Study of Machine Learning Models and Feature Selection Techniques for Predicting Fragile X Tremor Ataxia Risk

Author

Poudel, Angeela

Publication Date

2024

Peer reviewed|Thesis/dissertation

A Comparative Study of Machine Learning Models and Feature Selection Techniques for Predicting Fragile X Tremor Ataxia Risk

By

ANGEELA POUDEL
THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Health Informatics
in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Prabhu Shankar, Chair

Nick Anderson

Chen-nee Chuah

Committee in Charge

2024

Contents

1 Introduction 1
1.1 Fragile X–associated tremor/ataxia syndrome (FXTAS) 2
1.2 Clinical Rating Scales 6
1.3 Motivation 8
1.4 Research Questions 11
2 Methods Description 11
3 Data description 14
3.1 Curating the Dataset 14
4 Literature Review 15
5 Proposed Method 17
5.1 Exploratory Data Analysis (EDA) 17
5.2 Description of Algorithm 21
6 Results 27
6.1 Feature selection 27
6.2 Evaluation Metrics 32
6.3 Risk Scores 34
7 Observation 36
7.1 Methodological Observation 36
7.2 Clinical Observation 39
8 Conclusion 40
1 Appendix 45

List of Figures

1 Associated risk with variation in demographic and other factors 5
2 Variables that capture information pertinent to Neuropsychological Functional Domain 9
3 Age distribution of participants, most of the participants are above 60 years of age 13
4 Ratio of participants with and without diagnosis 14
5 Distribution of CGG repeat size, in presence and absence of a FXTAS Diagnosis shows considerable overlap 16
6 Distribution of cognitive abilities, in presence and absence of a FXTAS Diagnosis [NO=GREEN,YES=RED] shows no clear linearly separable boundary 18
7 Distribution of RTI Five choice movement, BDS-Score and Global Sevrity Scores in presence and absence of a FXTAS Diagnosis shows no clear linearly separable boundary 19
8 Predictive model construction process 22
9 Comparison of number of features selected between Lasso, RFE, SFS, and RandomForest 28
10 Density plots that compare various features for individuals with and without a diagnosis of FXTAS 29
11 Correlation Matrix of Selected Features: This matrix illustrates the pairwise correlation coefficients between consistent selected features all models. The strength and direction of the correlations are indicated by the color scale, with red representing stronger positive correlations and blue representing negative correlations. 31
12 Comparison of Accuracy scores 34
13 Comparison of AUCROC between Lasso, RFE, SFS, and RandomForest 35

14 Risk Score vs 1st Trial movement and Risk Score vs New RVP:This scatter plot displays the relationship between the risk score and the new RVP (Rapid Visual Processing) A' min score from the CANTAB (Cambridge Neuropsychological Test Automated Battery) 36
15 Risk Score vs. CGG Repeat Size - Shows the correlation between CGG repeat size and scaled risk scores 37
16 Risk Score vs. RTI Five-choice movement time - Demonstrates the correlation between RTI Five-choice movement times and scaled risk scores, indicating movement time as a predictive factor for FXTAS risk 38

List of Tables

1 Balanced oversampling trained but tested on imbalanced Performance 33
2 Balanced Downsampled Trained Model tested on imbalanced test set 45

Acknowledgements

My accomplishments are the result of the tremendous support from my family, friends, the UCD MHI Graduate Studies faculty and staff, and my mentors and colleagues at UC Davis Health and the Mind Institute.
I wish to express my deep appreciation for my thesis chair, Dr. Prabhu Shankar, UC Davis School of Medicine Division of Public Health Sciences, for his support and guidance throughout the graduate program.
Additionally, I am grateful to Nick Anderson, Ph.D., Professor of Informatics, for his mentorship and offering an amazing program with an informative and relevant curriculum. I am also grateful for Professor Mark Caroll Division of Public Health Sciences for his advice and mentorship. I extend my sincere gratitude to Dr. Chen-nee Chuah PhD, for her professional mentorship and support. I would also like to acknowledge the use of AI tools, specifically ChatGPT, for assisting with grammar checking and improving the clarity of my writing. This use was in accordance with the guidelines provided by the UC Davis Center for Educational Effectiveness (2023). Finally, I want to thank my family and friends for their unwavering support and encouragement throughout my journey.

Abstract

Fragile X–associated tremor/ataxia syndrome (FXTAS) primarily affects older adults who carry the FMR1 gene premutation. This conditions include severe symptoms such as cognitive deterioration, intention tremors, neuropathy, and progressive ataxia. Despite its profound impact on individuals and their families, there is currently no dependable method for predicting the onset or progression of FXTAS. Our research aims to fill this critical gap by introducing a predictive method based on a thorough analysis of clinical, genetics and behavioral factors. We utilized a dataset comprising longitudinal records from 103 patients over three to five visits. Employing advanced feature selection techniques and Random Forest probabilistic models, we developed a highly accurate risk prediction model for FXTAS. Our study has three primary objectives: first, to find an ideal combination of Machine Learning (ML) models and feature selection techniques that perform better across different performance metrics—accuracy, recall, precision, sensitivity, specificity; second, to determine whether undersampling or oversampling provides better results across all performance metrics; and third, to quantify the risk by determining precise risk scores. Our analysis includes four feature selection methods—Random Forest, Lasso, Recursive Feature Elimination (RFE), and Statistical Feature Selection (SFS)—and four classification algorithms: Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosting (XGBoost), and Random Forest (RF). The combination of XGBoost and Recursive Feature Elimination (RFE) and the combination of Random Forest and RFE both performed exceptionally well, achieving the highest accuracy of 86.67, precision of 0.83, and recall of 0.67 compared to other models. However, Random Forest with RFE performed slightly better in measuring AUROC, indicating a superior ability to distinguish between classes. The feature selection methods results showed consistent features: Stop Signal Task (SST) Median Score and Full IQ Score which are both used to evaluate cognitive functions. Another consistent features shown were five-choice Movement Reaction Time and Purdue Pegboard Scores (right-hand and left-hand) measure which are different aspects of motor skills. These findings provide significant advancements in clinical decision-making and personalized treatment strategies for diagnosing FXTAS.

1 Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

The field of healthcare is in a constant state of evolution,driven by a strong dedication to improving patient outcomes through innovation and technology. Over the past decade, better understanding of disease mechanisms has improved the ability to prevent, diagnose, and treat common afflictions. The innovation underlying such progress continues to advance and accelerate change, with new technologies and medical interventions providing new options for care and treatment.Previous research have shown traditional approaches to healthcare, while effective in many respects, often fall short when it comes to personalized and precise interventions .

Machine learning has the potential to transform healthcare by using patient data for early prediction of diseases, which can help in preventing the need for more serious care, and to tailor treatments to individual patients. It has successfully predicted individual disease trajectories and responses to treatment by analyzing large datasets of patient information [26]. This allows healthcare providers to tailor interventions more precisely, leading to

improved outcomes and a better quality of life for patients. For example, researchers have used machine learning algorithms to predict the progression of Alzheimer's disease in individual patients based on factors such as brain imaging data, genetic markers, and cognitive assessments [5]. By identifying patterns and correlations in these data, machine learning models can provide more accurate predictions of disease progression than traditional methods. One of the key challenges in using machine learning for disease prediction is the availability of high-quality, comprehensive data. Much health data come from different sources such as interviews, surveys, and unstructured clinical notes. These data are often incomplete, biased, or noisy, which can affect the performance of machine learning models. To overcome these challenges, researchers have employed a range of machine learning algorithms, including decision trees, random forests, support vector machines, and deep learning networks [1]. These algorithms have been applied to a variety of datasets, including electronic health records, genomic data, and social media posts [19] [1]. In general, the results of these studies have shown that machine learning algorithms can accurately predict the onset of diseases, with performance comparable to or better than traditional statistical methods [19] [1].

1.1 Fragile X–associated tremor/ataxia syndrome (FXTAS)

Fragile X-associated tremor/ataxia syndrome (FXTAS) was identified just over a decade ago.It is a complex condition with difficulties in movement and thinking ability (cognition) [11].It is seen in a subgroup of older adults who are carriers of premutation alleles of the fragile X mental retardation 1 (FMR1) gene [15] [4]. This condition primarily affects males over the age of 50 with the Fragile X premutation [22]. It includes intention tremor, which is shaking during movements like reaching for an object, and issues with coordination and balance, known as ataxia. Typically, intention tremors appear first, followed by ataxia years later, but not everyone with FXTAS has both symptoms [15].Individuals with FXTAS also experience cognitive impairments, such as short-term memory loss and decreased executive function.These symptoms could affect their ability to plan, control impulses, solve problems, monitor themselves, and maintain cognitive flexibility. They might also get anxiety, depression, mood swings, or

irritability [4]. FXTAS is often misdiagnosed and goes unrecognized. Although neurologists are becoming more knowledgeable about FXTAS, primary care physicians see and treat about half of the affected individuals [4][10][14]. Diagnosis would be enhanced by better medical education for doctors and more precise, comprehensive diagnostic criteria .

Risk Factors

FXTAS is influenced by various risk factors. In selecting variables for the dataset, we specifically targeted those associated with the following risk factors, guided by expert input.

Age Age is a significant factor in FXTAS, with older carriers being more susceptible to cognitive deficits, including dementia [14]. Male carriers with FXTAS over the age of 50 have also been found to have significant reductions in general intelligence scores and marginally significant deficits in logical memory compared to their non-carrier male siblings [15].

CGG repeat size Individuals with FXTAS have a mutation in which a DNA segment, known as a CGG triplet repeat, is expanded within the FMR1 gene. Normally, this DNA segment is repeated from 5 to 40 times. In people with FXTAS, the CGG segment is repeated 55 to 200 times. This mutation is known as an FMR1 gene premutation. An expansion of more than 200 repeats, a full mutation, causes a more serious condition called **fragile X syndrome[[15][16]**. The size of the CGG repeat has been found to be inversely related to the onset age of tremor and ataxia, as well as the age of death in FXTAS patients. In a study, men with over 70 CGG repeats had six times higher risk of severe cognitive impairment compared to those with 45-70 repeats [16].Furthermore, larger CGG repeats and increased FMR1 mRNA have been related to volume loss and decreased activation in brain regions linked to working memory, respectively[16].

FXTAS stage The stages of FXTAS correspond to the length of illness and the severity of the symptoms There are six stages based on the progression of motor deficits :

1. Stage 1: Subtle or questionable tremor and/or balance problems.

2. Stage 2: Minor tremor and/or balance problems, with minimal interference in activities of daily living (ADLs).

3. Stage 3: Moderate tremor and/or balance problems with significant interference in ADLs.

4. Stage 4: Severe tremor and/or balance problems, requiring the use of a cane or walker.

5. Stage 5: Daily use of a wheelchair.

6. Stage 6: Patients are bedridden [11].

Gender While both males and females can be affected, males are generally more severely affected due to having only one X chromosome, since they don't have a second X chromosome to balance the mutation.Women likely have fewer symptoms because their second X chromosome has a normal allele.Although the symptoms of FXTAS in most women can differ from the original diagnostic criteria, they can be just as severe as those in men [12]. Previous research have also shown that in Magnetic Resonance Imaging (MRI) women typically show less brain atrophy, white matter disease, tremor, and ataxia compared to men with FXTAS [12]. Other than MRI imaging, women with and without FXTAS have more medical commodities than premutation men. A research also showed that women with FXTAS have a higher incidence of fibromyalgia and hypothyroidism compated to both women without FXTAS and men with FXTAS. It has also been shown that females with FXTAS experience peripheral neuropathy, seizures, and hypertension more frequently. It has also been been reported that some women carriers also have reeported of having multiple sclerosis and other (neurological) disorders [11] [12].

Lifetime depression Patients with FXTAS have a 65% lifetime chance of developing mood disorders, such as Major Depressive Disorder (MDD), which is much higher than in the general population of the same age.Therefore, it is important to assess the potential impact of lifetime depression on cognitive status in FXTAS [11].

Other medical illness Female carriers of the FMR1 premutation may experience a range of health issues including primary ovarian insufficiency, thyroid disorders, peripheral neuropathy, hypertension, fibromyalgia, autoimmune diseases, and migraines. Male carriers often present with type II diabetes, hypertension, sleep apnea, migraines, and cardiovascular disease.Studies have shown that around 31.4% of premutation carriers with

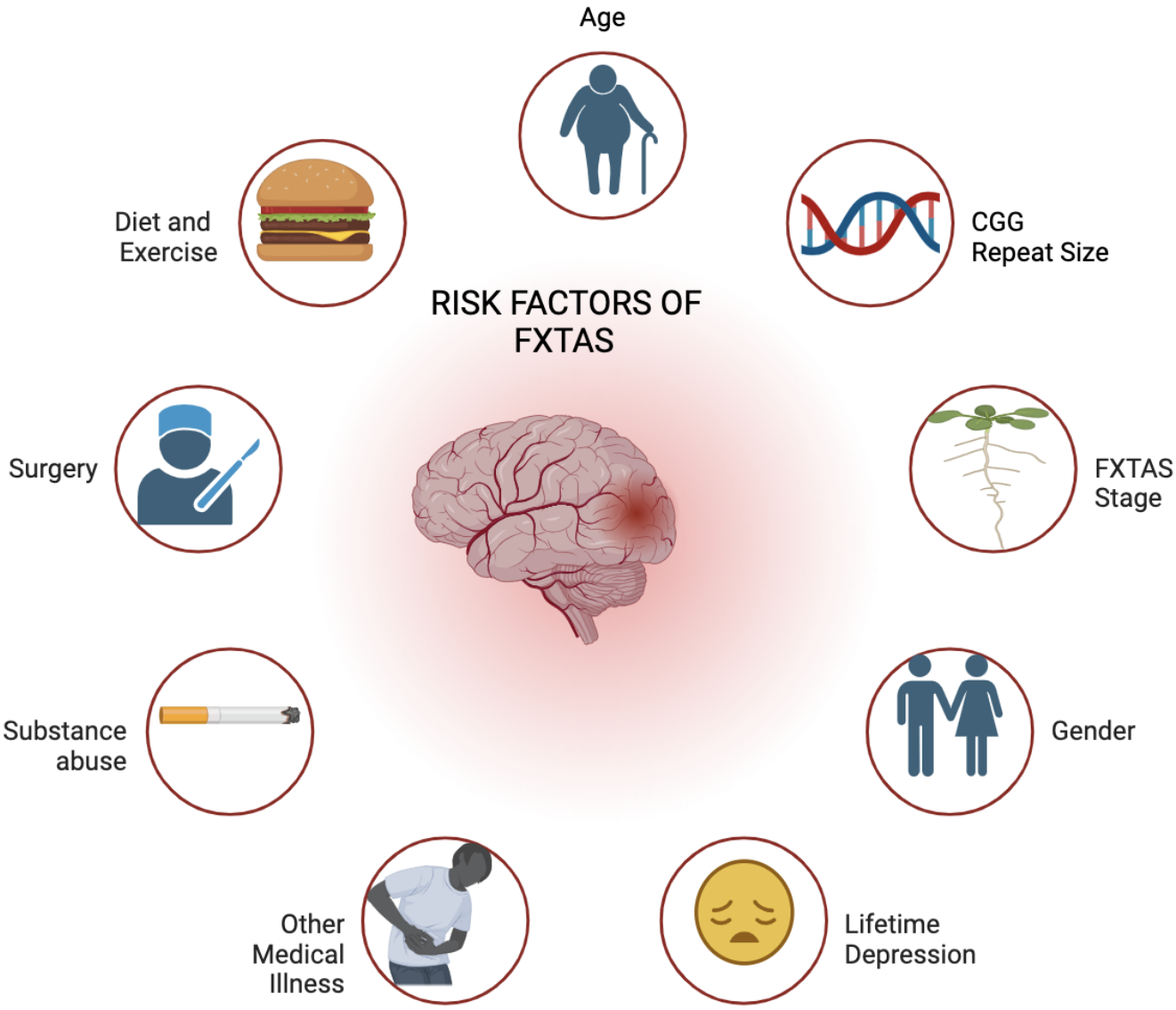

Figure 1: Associated risk with variation in demographic and other factors

FXTAS suffer from sleep apnea, which contributes to the development of metabolic syndrome and cognitive disorders [11] [9].

Substance use Alcohol and other central nervous system (CNS) depressants, such as benzodiazepines and opioids, can contribute to cognitive impairment. A review of medical records for 184 premutation carriers, both with and without FXTAS, revealed that 24% had a history of dependance on substance where alcohol was the most commonly abused substance, particularly among men [11] [9].

Surgeries with general anesthesia Pervious studies suggest that general anesthesia during surgeries might aggravate premutation-related symptoms. Individuals over 60 years old have occasionally experienced the onset of tremor or ataxia within a few weeks after undergoing surgery with general anesthesia [11].

1.2 Clinical Rating Scales

In the study, we included scales and measures that provide a comprehensive assessment of participants' memory and mental health, as outlined in the previously mentioned risk factors.

Behavior Dimensions Scale (BDS-2) score: The BDS-2 is a tool used to assess the severity of behavioral problems in individuals with intellectual disabilities. It covers various dimensions such as disruptive, self-absorbed, and asocial behaviors[Hawthorne Educational Services, Inc. (n.d.). Behavior Rating Scale.

Global severity score: This score provides an overall assessment of the severity of the psychological symptoms of the individual. It is often used in clinical settings to gauge the overall impact of a condition and assess the effect on the functionality of an individual. This could be an important factor that reflects the overall impact of the FXTAS on the individual's psychological functionality and quality of life.

Spatial working memory (SWM) error: SWM error refers to the number of errors made in a spatial working memory task. This type of task assesses an individual's ability to remember and manipulate spatial information. Errors in this test are defined as instances where participants incorrectly select boxes they have previously determined to be empty or revisit boxes already found to contain a token. This evaluation not only measures memory accuracy but also the strategy used by participants, reflecting their executive function capabilities. Such assessments are crucial

for understanding cognitive function in various neurological and psychological conditions.

Stop signal task (SST) Median: The stop signal task is a measure of response inhibition, which is the ability to stop a planned or ongoing response. The median stop signal reaction time(SSRT) is a key metric in this task.The median SSRT provides a measure of the central tendency of stopping latencies across trials or sessions, reflecting the typical response inhibition capability of a participant under the tested conditions. This median value helps in understanding the participant's inhibitory control by indicating how quickly they can typically stop their responses once a stop signal is presented.

One Touch Stockings (OTS) problem: It is a test of executive function that assesses planning and problem-solving abilities. It involves predicting the outcome of moving disks in a stockings-like configuration.

Full Scale IQ: It is a measure of overall intellectual functioning derived from standardized IQ tests. It provides an estimate of an individual's intellectual abilities compared to others of the same age.

Metacognition Index (MI) T-Score: The Metacognition Index is a measure of metacognitive abilities, which refer to the awareness and control of one's own cognitive processes. The T-score indicates how an individual's score compares to the normative population.

Behavioral Regulation Index (BRI) Score: The Behavioral Regulation Index is a measure of executive function related to self-regulation, including the ability to inhibit responses and shift between tasks.The BRI specifically measures an individual's ability to modulate and control behaviors effectively, which includes: the capacity to control impulsive responses and the ability to move freely from one situation or aspect of a problem to another in response to changing rules or demands. Patients with FXTAS often exhibit impairments in executive functions, including difficulties with task switching, problem-solving, planning, and inhibiting inappropriate actions. The BRI can help quantify these deficits, providing a clearer picture of how FXTAS is impacting an individual's cognitive control and behavioral regulation.Changes in BRI scores over time can provide valuable information about the progression of executive dysfunction.Understanding the specific areas of executive function that are most affected in an individual with FXTAS can help healthcare providers tailor cognitive rehabilitation strategies.

SCID : The Structured Clinical Interview for DSM-IV (SCID) is a

diagnostic tool used to assess various psychiatric disorders, including anxiety disorders. It provides a structured format for clinicians to evaluate symptoms and make a diagnosis.Individuals with FXTAS may also experience psychiatric symptoms or disorders. Common comorbid conditions include mood disorders, anxiety, and cognitive impairment, which could be systematically assessed using the SCID. In research studies involving individuals with FXTAS,it can be used to screen for and exclude other psychiatric disorders that might confound .It helps in delineating the psychological symptom profile that is directly related to FXTAS from other psychiatric conditions. For a comprehensive evaluation of a patient's mental health, the SCID can be part of the broader assessment toolkit used by clinicians treating patients with FXTAS, especially when psychiatric symptoms are present and may impact the patient's quality of life or the management of FXTAS.

Rapid Visual Information Processing (RVP): RVP is a test of sustained attention and vigilance. It involves detecting target sequences of digits presented rapidly and requires continuous monitoring over a period of time. FXTAS is associated with cognitive decline in some individuals, especially in areas related to executive function, memory, and processing speed. RVP can help in quantifying the extent of cognitive impairment related to these domains.Regular administration of the RVP test can provide insights into the progression of cognitive symptoms in patients with FXTAS. This is crucial for managing the disease effectively and adjusting treatment plans as needed.

1.3 Motivation

The diagnosis of FXTAS is frequently overlooked for two main reasons:

Lack of Awareness and Knowledge: FXTAS is a relatively recently identified condition, first described in 2001. It is not widely known even among medical professionals, which can lead to misdiagnosis or delayed diagnosis. The symptoms of FXTAS can vary widely in severity and presentation, and because it shares similarities with more common disorders such as Parkinson's disease or Alzheimer's disease, it can be mistaken for these other conditions [4][10][14].

Subtle Early Symptoms: The onset of FXTAS symptoms typically occurs in individuals over 50 years of age, often starting subtly and gradually worsening. Early symptoms can include minor tremors and

Functional Domain	Variable
Inhibitory Control	CANTAB SST median correct RT on Go (ms)
Planning	CANTAB OTS problems solved on first choice
Motor Function	
Manual Movement Speed	CANTAB RTI 5-choice movement time (ms)
Executive Control of Movement	BDS-2
Manual Dexterity	Purdue Pegboard (R+L+both hands)
Manual Reaction Time	CANTAB RTI 5-choice reaction time (ms)
Episodic Memory	
Verbal	WMS Logical Memory II - Recall
Visual	CANTAB PAL total errors
Visual Attention	CANTAB RVP A' signal detection
Working Memory	
Auditory	WMS Letter-Number Sequencing
Visual	CANTAB SWM between errors

Figure 2: Variables that capture information pertinent to Neuropsychological Functional Domain

balance issues, which are easy to dismiss as normal aging or misattribute to other neurological conditions [15]. As the disease progresses, symptoms become more pronounced, including significant problems with movement, memory, and mood, which might then prompt more specific investigations leading to a correct diagnosis. Because FXTAS is linked to a premutation in the FMR1 gene, it often requires specific genetic testing to confirm the diagnosis [11]. This testing is not routinely performed unless FXTAS is specifically suspected, which further complicates timely diagnosis. The identification of the FMR1 premutation can be pivotal, not only for diagnosing FXTAS but also because it has implications for family members who may also be carriers of the mutation [11] [10].

In our study, we are building machine learning models that can be used to create an efficient prediction model and determine the important features [1].

Early Detection and Diagnosis: Machine learning models can help in the early detection of FXTAS by analyzing genetic data and clinical symptoms more efficiently. Early detection can be helpful in managing the progression of FXTAS and initiating appropriate interventions sooner.It can accelerate research by identifying potential biomarkers for FXTAS and predicting the efficacy of proposed treatments based on historical data. This could lead to faster clinical trials and more effective treatments reaching patients sooner.

Personalized Treatment Plans: AI and machine learning can analyze vast amounts of data from patient histories, genetic information, and treatment outcomes to tailor personalized treatment strategies. This could be especially beneficial for FXTAS patients, as the disease manifests with varying symptoms and progression rates.

Educational Tools: This is especially useful in regions where access to specialized healthcare providers who are familiar with FXTAS is limited. The integration of machine learning models, once validated, introduces a transformative opportunity for deployment across diverse healthcare settings, especially those with limited access to specialized medical professionals. This innovation has the capacity to extend the reach of healthcare services, offering substantial benefits to individuals in underserved or remote regions. Once these models are proven to work well, we can use them in places where there aren't many specialized doctors.

1.4 Research Questions

Our study is driven by three key questions:
1. What are the most significant predictors in the diagnosis of FXTAS?
2. Which feature selection and classifier combinations leverage clinical and demographic data to accurately predict FXTAS diagnosis?
3. Generate risk scores that quantify the likelihood of developing FXTAS : It is a significant step towards personalized medicine. These scores can be derived from machine learning models that use demographic, genetic, and clinical data to assess risk levels. By integrating predictors such as age, family history, and specific genetic markers from the FMR1 gene, these models can provide healthcare providers with tools to identify high-risk individuals early.

2 Methods Description

Considering the research objectives and the characteristics of our dataset, this paper selects the following four machine learning techniques: **Random forest, Support Vector Machine (SVM), Logistic Regression,and XGBoost.**

Random Forest Health data especially the FXTAS data is intricate interactions and non-linear relationships . Studies have shown Random Forests are well-equipped to manage Random such data [1] as this method levarages from its foundational structure of decision trees, enhancing it by introducing randomness in attribute and dataset selection during the training phase. This randomness enhances its ability to capture intricate patterns, making it a powerful tool for applications such as disease prediction and gene selection. [22].TBy combining multiple decision trees, Random Forests form an ensemble that mitigates the bias typically seen in individual decision trees. This aggregation not only improves the model's accuracy but also its generalizability across different types of datas. FThis ensemble approach not only boosts model accuracy but also improves generalizability across various data types. Furthermore, Random Forests are resilient to noise and outliers, common challenges in medical datasets, and can effectively handle both categorical and numerical features without extensive preprocessing [22] [11]. In the training phase, Random Forest builds several decision trees using random subsets of the data. During the

testing phase, these trees work together to classify or predict outcomes based on input data. [5] [2].The ensemble is represented as a forest F $=\{f1,...,fn\}$, with each F being a decision tree. Predictions are made by averaging the outputs (for continuous variables) or using majority voting (for categorical variables), which reduces overfitting and enhances the model's generalization capability[2].

SVM SVM is a powerful classifier chosen for its ability to handle different feature scales and its effectiveness in high-dimensional spaces [11]. It excels in solving high-dimensional, nonlinear, small-sample pattern recognition problems, offering several unique advantages. SVM's strong theoretical foundation ensures that its extremum solution is the global optimal solution, rather than a local minimum, which contributes to its excellent generalization ability for unknown samples. These attributes make SVM a valuable tool for various applications, including regression estimation, time series prediction, and pattern recognition[1].

Logistic Regression Logistic regression is favored for binary classification tasks due to its simplicity and effectiveness as a linear model with good interpretability. This model is particularly useful in understanding the impact of each feature on the prediction outcome, making it valuable in medical studies where feature influence is crucial[11]. Its low computational requirements make it suitable for baseline modeling and quick iterations. [1] [11].Given that FXTAS diagnosis is a binary outcome (YES/NO), logistic regression is an ideal choice for this type of classification task[13].

XGBoost XGBoost [7] is a gradient-boosted decision tree designed for speed and efficiency, developed by Tianqi Chen and implemented in C++. It offers high efficiency, flexibility, and portability, utilizing the gradient boosting framework to implement machine learning algorithms [3]. XGBoost effectively handles sparse data, which is common in medical datasets due to missing values or zero entries. It is also well-known for its excellent model performance and rapid execution speed, making it suitable for large and complex datasets [1][3].

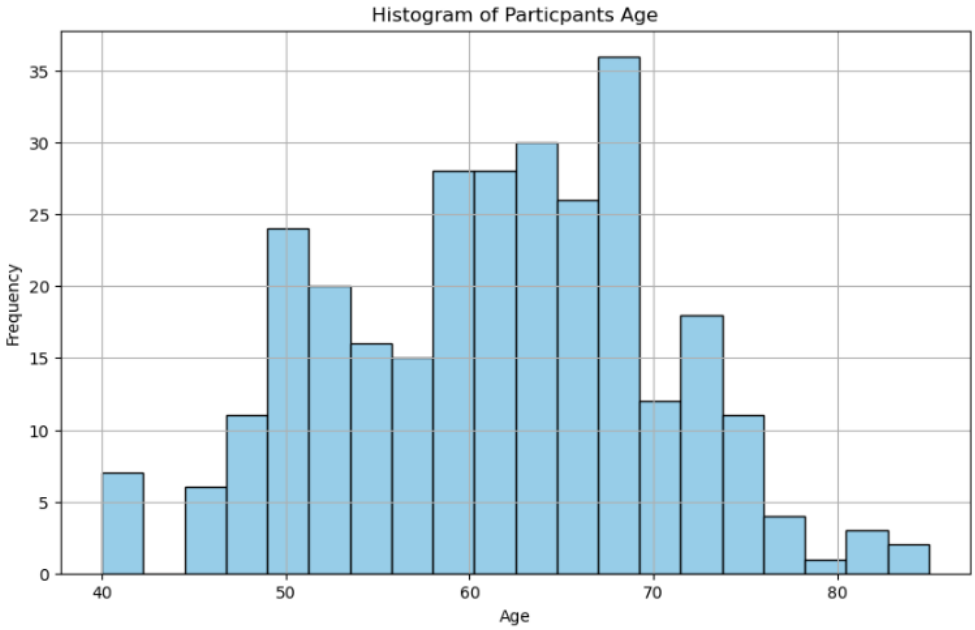

Figure 3: Age distribution of participants, most of the participants are above 60 years of age

3 Data description

3.1 Curating the Dataset

The focal dataset for our study includes data from 103 male patients, aged between 40 and above, assessing the presence of Fragile X-associated Tremor/Ataxia Syndrome (FXTAS). Of these, 31 have been diagnosed with FXTAS while 72 have not (Figure 4). This dataset encapsulates a rich array of 42 variables covering psychological well-being metrics like mood, anxiety, and depression scores; demographic details including race, ethnicity, and age; as well as behavioral, cognitive, and motor skills assessments. It provides a holistic view of the diverse symptoms and characteristics associated with FXTAS, with patients averaging three visits each.
The primary target for prediction in this dataset is the FXTAS diagnosis, encoded as 'YES' or 'NO'. The dataset is structured as $(X, y) \in \{(X_i\}_{i=1}^{N\times T\times K}, y_i)$, where X encompasses psychological, physiological, behavioral, and cognitive data. Here, N represents the number of samples, K denotes the number of features per sample, and T, ranging from 1 to 5, indicates the number of patient visits. The binary label y indicates the FXTAS status of each patient, coded as 0 or 1.

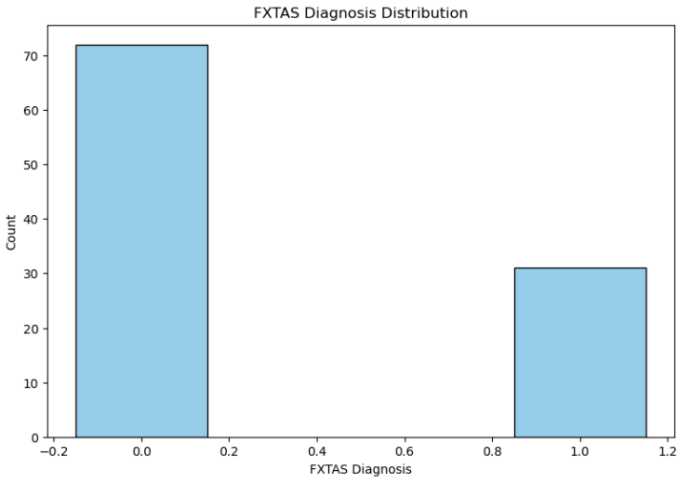

Figure 4: Ratio of participants with and without diagnosis

In addition, SCID data like Anxiety, Mood Disorder Substance Abuse are

recorded in binary format, and the dataset incorporates comprehensive demographic data and FMR1 genetic profiles. We've applied stringent data cleaning protocols to enhance the dataset's reliability, notably removing variables with over 50% missing data, such as Erectile Dysfunction, Sleep Disturbances, surgery ,Cancer and Diabetes, to prevent skewed results and maintain the accuracy of our analysis. To address discrepancies in the labeling of anxiety and mood disorders, I standardized these variables into a uniform binary format, merging them into a single column to increase data usability and consistency. This adjustment not only streamlined the dataset but also ensured a more comprehensive analysis of the interrelations between genetic markers, psychological conditions, and demographic factors, making each participant's profile as complete and informative as possible for our research objectives. I also converted survey data into a structured tabular format. For instance, questions related to diet, alcohol consumption, and coffee intake were originally in free-text format. Responses varied widely, with some participants answering "1 cup" while others specified quantities like "3-4 ounces" for their daily coffee consumption. To simplify the analysis and standardize the data, I transformed these responses into a binary format. Specifically, I changed the question format to "Do you drink coffee?" with possible answers of "yes" or "no." This adjustment allowed for more straightforward and consistent data analysis, enabling clearer comparisons and trend identification across the dataset.

4 Literature Review

Many studies have been conducted on FXTAS, and even more on using machine learning to predict various diseases. However, not many studies have specifically combined machine learning approaches to predict FXTAS, particularly a comparative study. I chose to do a comparative study instead of just selecting a single model is due to the lack of related or prior research in this area. My goal is to find the best method that is a perfect fit to analyze the complex FXTAS data. Although some researchers, like [25], have explored molecular correlations, and conducted traditional statistical analyses, machine learning has not been utilized. The results from [15], also demonstrated an 83% accuracy in predicting Alzheimer's disease using various machine learning algorithms such as voting ,SVM,Random Forest

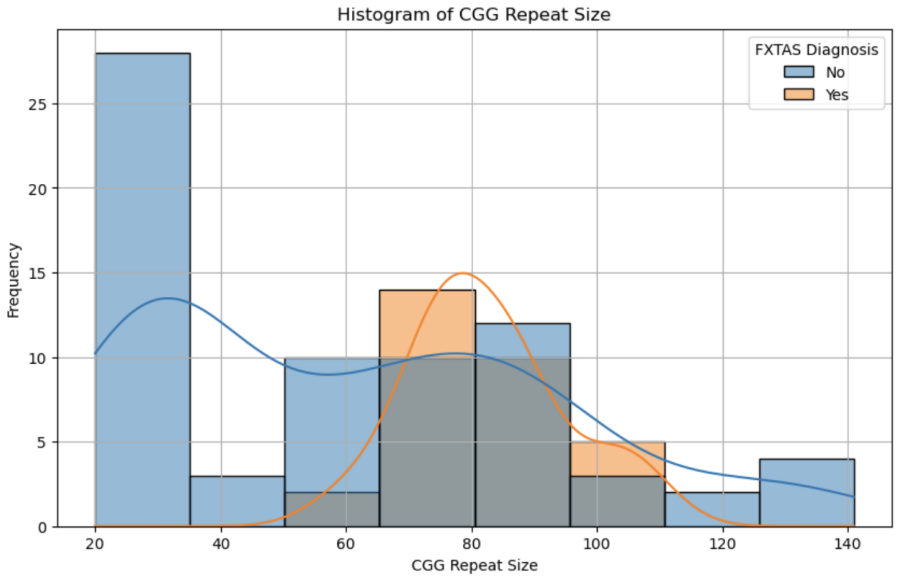

Figure 5: Distribution of CGG repeat size, in presence and absence of a FXTAS Diagnosis shows considerable overlap

ad XGBoost, inspired us to include a range of models. The inspiration behind us incorporating logistic regression is the study by [17] that showcased the outstanding performance of Gradient Boosting Tree (GBT) models, achieving AUC scores as high as 0.939. The combined insights from these studies have been very helpful in developing our approach. They emphasize the significance of using multiple models, tackling data-specific issues like class imbalance, and highlight the importance of feature selection and model interoperability. Study such as [20] highlighted the difficulties in obtaining sufficiently large and accurate labels for supervised learning, emphasizing the need for a robust dataset for our study. In summary, [20] study provided us with critical insights into the nuances of applying machine learning in neurodegenerative disease research. It guided us in dataset preparation, algorithm selection, and the overall design of our machine learning approach, emphasizing the importance of accurate data labeling, the potential of different learning paradigms, and the careful selection of algorithms based on the nature of our data. Study by [24] showed how tree based model such as Random Forest is used in prediction of neurodegenerative conditions such as Alzheimer's Disease (AD). The

result showed high accuracy of 93%. Thee study also motivated us to implement comparable techniques in our research, particularly using a random forest classifier and balanced sampling approach to address class bias. The study [2] focused on the importance of using explainable machine learning methods to analyze AD. This led us to utilize Grid Search for hyperparameter tuning of our Random Forest and XGBoost models, ensuring that our model's results are both interpretable and clinically valuable. The paper also focused on the importance of the proper selection of machine learning algorithms based on data type and volume aligns with our methodology. We meticulously evaluated these aspects to choose the most suitable algorithms for our FXTAS dataset. The insightful literature review and careful consideration were crucial due to the heterogeneous and complex nature of data commonly found in neurodegenerative disease research.

5 Proposed Method

5.1 Exploratory Data Analysis (EDA)

Prior to developing the model, an exploratory data analysis (EDA) was carried out. The first step in data cleaning addressed inconsistencies, managed outliers, and standardized data formats. To assist with subsequent modeling decisions, a thorough analysis of variable characteristics was performed, focusing on identifying binary and linear attributes.

A key aspect of the EDA was the evaluation of missing values. By computing the number of missing entries per subject per feature, we determined the percentage of patients lacking data for each feature at any point during their visits. The usability of each feature was then assessed based on the percentage of missing values. Features with a high percentage of missing values per column per patient were excluded from the analysis due to insufficient information. Conversely, features with a low proportion of missing values were deemed useful. Medical professionals at the UCD Mind Institute, with their expertise and insights, identified clinically recognized normal values for FXTAS patients, which were then used to impute the missing data. To preserve the integrity of critical data for subsequent analyses, it was decided to use these expert-informed normal values.

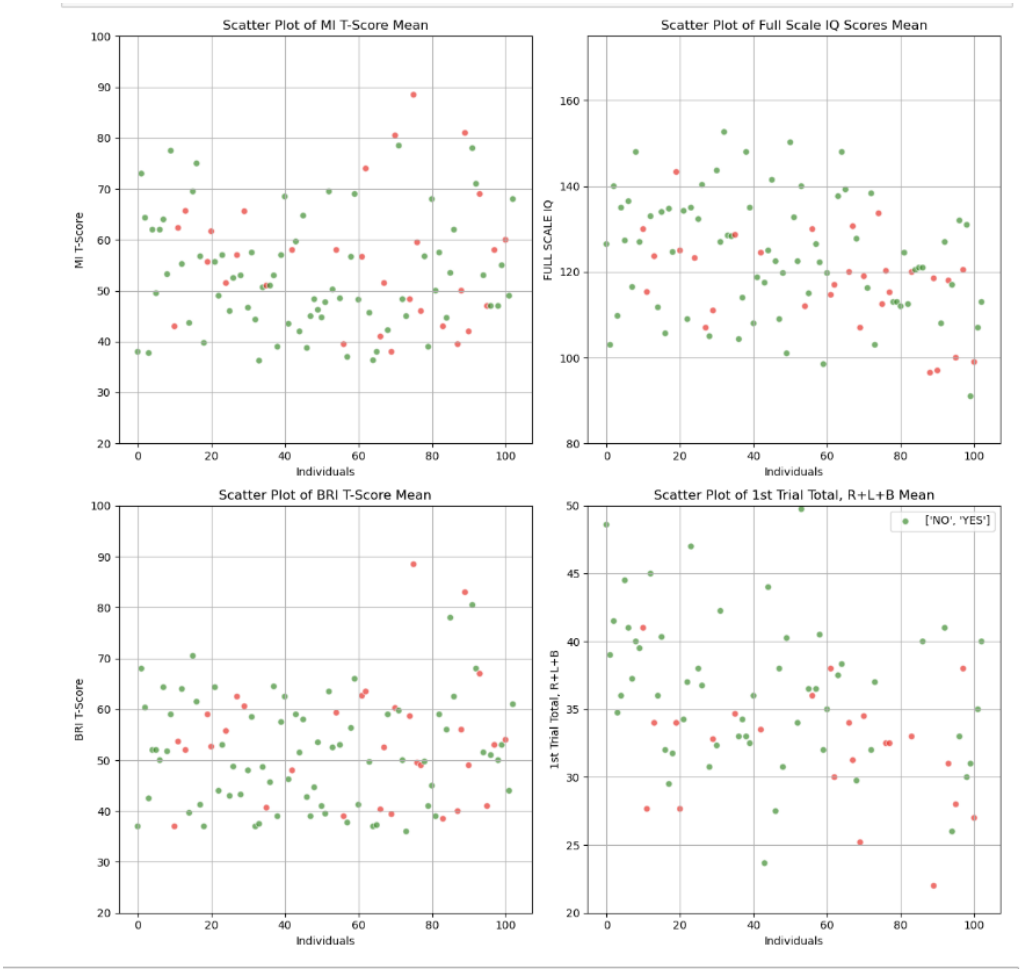

Figure 6: Distribution of cognitive abilities, in presence and absence of a FX-TAS Diagnosis [NO=GREEN,YES=RED] shows no clear linearly separable boundary

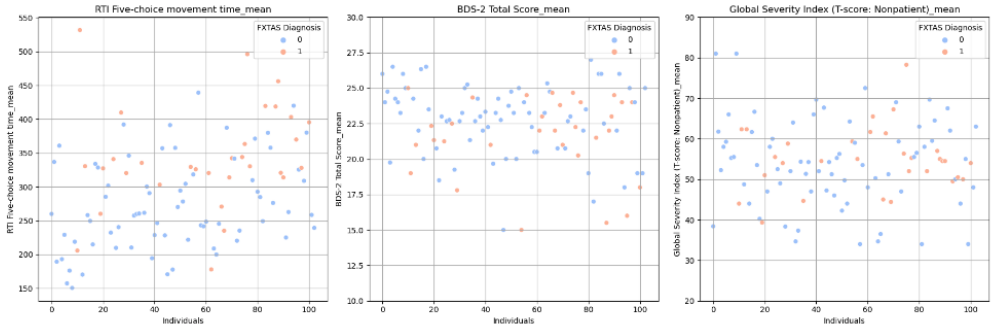

Figure 7: Distribution of RTI Five choice movement, BDS-Score and Global Sevrity Scores in presence and absence of a FXTAS Diagnosis shows no clear linearly separable boundary

In the process of diagnosing FXTAS, we are analyzing the temporal dynamics of relevant variables. To achieve this, we calculated the mean, median, minimum, maximum, standard deviation, and slope for each feature across all time points for each patient. This analysis is essential to observe temporal changes. The resulting dataset is structured with one row per time series, including columns for each of the calculated statistics. Feature columns with missing values for the slope or standard deviation were specifically imputed with 0, as this indicated that the patient had only one recorded visit for those features. Similarly, the mean, minimum, median, and maximum values were imputed with clinically recognized normal values to maintain consistency across the dataset. The comprehensive EDA not only prepared the dataset for subsequent modeling but also uncovered valuable patterns and trends, providing a deeper understanding of the predictive factors associated with FXTAS. The strategic decisions regarding missing values, including the specific imputation criteria, enhanced the dataset's robustness for further analyses. The histogram in Figure 5 illustrates the distribution of CGG repeat sizes among individuals tested for FXTAS, categorized by their diagnostic status. The data is represented in two distinct groups: individuals diagnosed with FXTAS (red) and those without a diagnosis (blue). CGG repeat sizes are plotted along the x-axis, ranging from approximately 20 to over 140, with frequency displayed on the y-axis.
The kernel density estimate (KDE) for the non-diagnosed group peaks at

around 40 CGG repeats, indicating a higher frequency of lower CGG repeat sizes in this group. In contrast, the diagnosed group's KDE shows a broader distribution with a peak at around 80 CGG repeats, suggesting a significant prevalence of higher repeat sizes among individuals diagnosed with FXTAS. This distribution pattern highlights a critical trend: higher CGG repeat sizes are markedly more common in individuals with FXTAS, underlining the potential genetic underpinnings of the syndrome. This visualization supports the analysis of CGG repeat size correlation with FXTAS diagnosis. In Figure 6, the Scatter Plot Analysis of Behavioral and Cognitive Metrics shows :

RTI Five-choice Movement Time: This plot displays a wide distribution of reaction times across individuals, with no clear distinction between those diagnosed with FXTAS (red) and those without (blue). It uggests a varied impact of FXTAS on motor response times.

BDS-2 Total Score: The scores, indicative of cognitive function related to decision speed and problem-solving, show overlap between the two groups. Individuals with FXTAS do not consistently exhibit lower scores, indicating that cognitive decline specific to these abilities may not be pronounced in all FXTAS cases.

Global Severity Index (T-score, Nonpatient): Global seeverity score also shows a similar pattern, with considerable overlap between diagnosed and non-diagnosed individuals. This can suggest that FXTAS's impact on general psychological health may vary widely among patients.

MI T-Score Mean and Full-Scale IQ Scores : These plots further evaluate cognitive capacities, with T-scores spanning from normal to below-average ranges across both groups. The Full-Scale IQ scores are similarly distributed, highlighting that while some individuals with FXTAS show lower cognitive performance, it is not universally characteristic of all diagnosed individuals.

BI T-Score Mean and 1st Trial Total, R+L+B : Both plots explore different cognitive and memory retention metrics. The scores reflect a broad spectrum of cognitive abilities, with no definitive pattern segregating individuals based on their FXTAS diagnosis status.

This plot in figure 3 reveals that age distribution among the participants does not show a clear correlation with FXTAS diagnosis, indicating that age alone is not a predictor of the syndrome in this sample. The correlation coefficient calculated for the relationship between Substance Abuse Disorder and Anxiety Mood Disorder means is approximately 0.021. This

value suggests that there is a negligible positive correlation. The changes in the Substance Abuse Disorder are very weakly associated with changes in the Anxiety and Mood Disorders. The low magnitude of this correlation coefficient indicates that, within this dataset, as Substance Abuse Disorder scores increase, there is only a slight and statistically insignificant increase in the Anxiety and Mood Disorder scores.T
he correlation coefficient between is approximately age and the Global Severity Index T-scores -0.241, which indicates a weak negative correlation between age and the Global Severity Index T-scores. This suggests that as age increases, there is a slight tendency for the Global Severity Index scores to decrease. However, the correlation is not strong.

5.2 Description of Algorithm

Our study employs a standard plug-and-play deep-learning pipeline Figure 8 .Each component of this pipeline is examined through extensive experimentation and is refined based on the results obtained. The pipeline consists of the following sub-components.

1. Derivation of the Feature Set: The dataset contains physiological, behavioral, and cognitive feature values for patients. Clinically, it has been noted that changes in these values are often more crucial for diagnosing FXTAS than the absolute values themselves. Therefore, learning these derived features can improve a model's ability to distinguish between patients who have experienced significant declines and those with preexisting conditions. This is similar to correctly identifying a patient who has recently experienced a cognitive decline versus misidentifying a patient who already has lower cognitive abilities. We calculated the mean, median, minimum, maximum, standard deviation, and slope for each feature across all time points for each patient using **[Algorithm 1]** in order to incorporate the knowledge of these trends in our learning problem.

2. Undersampling, Oversampling and Dataset prevalence ratios: We first randomly split the entire data into training and testing sets with the following distribution: 80% for training and 20% for testing. The dataset originally had only 30% of the samples as positive. [27] showed that imbalanced ratios in small data settings train very unstabilized models, therefore we chose to balance our training datasets by - 1. Undersampling and 2. Oversampling. We undersampled the dataset to balance the class labels. For oversampling, SMOTE [6] was used to make the class ratios

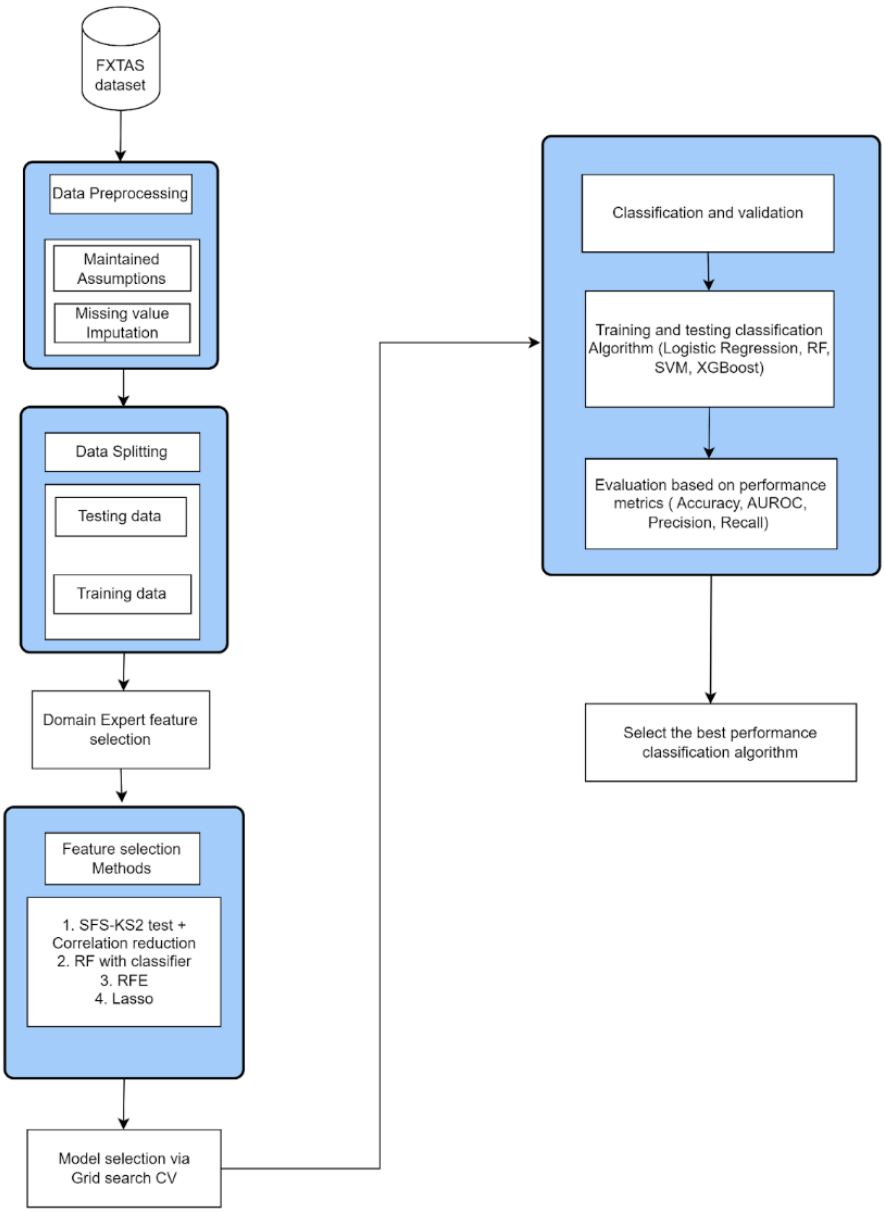

Figure 8: Predictive model construction process

Algorithm 1 Summarization of Feature Values

Input: $(X, y) \in \{(X_i\}_{i=1}^{N \times T \times K}, y_i)$

For Each i^{th} **of the N samples**

Step 1: For ${X_i}^T \times K$ Group data from all the T visit observations.

Step 2 Make $X_{i,j,T}$ for j in K features

Step 2.1 Calculate Summary $X_{i,j} =$
$[\mu(X_{i,j,T}), min((X_{i,j,T})), max((X_{i,j,T})), stddev((X_{i,j,T})), slope((X_{i,j,T}))]$

Step 2.2 Replace $X_{i,j} = SummaryX_{i,j} \in K'$

Output: $(X, y) \in \{(X_i\}_{i=1}^{N \times K'}, y_i)$

balanced. We train the models on both of these balanced settings and then test them on the original imbalanced test set to check for usability in real world. Our findings found the Oversampling models outperformed the undersampled models. Hence, in the main section of the paper we will discuss only the oversampled model results; leaving the undersampled results in Appendix 1.

3. Feature Selection: This dataset is characterized by high dimensionality relative to the sample size. Additionally, redundant information and noise can obscure meaningful data interpretation[21]. Feeding such a large number of features into the model could lead to overfitting. To address these challenges, we adopted a hybrid feature selection methodology, combining domain expert knowledge and algorithmic techniques [23]. Initially, experts manually selected features based on their comprehensive knowledge of FXTAS, ensuring clinical relevance. Following this, we applied algorithmic feature selection to identify the most effective subset of features for the learning problem, thus validating our hypothesis. **[Algorithm 2]**

To arrive at the most optimal feature set we do a comparative study between four feature selection methods.

A. Statistical Feature Selection (SFS) [Algorithm 3] We utilized Sequential Feature Selection (SFS) combined with the Kolmogorov-Smirnov 2 (KS2) test and correlation reduction techniques. This strategy enabled us to gradually include features based on their statistical relevance while minimizing redundancy. This method was chosen for its ability to harness

Algorithm 2 Feature Selection method description

Input: FXTAS data $(X, y) \in \{(X_i, y_i)\}_{i=1}^{N\times(K+M)}$
Available methods: FeatureSelectionMethods = SFS, RFE, RandomForest, Lasso
For FeatSelection in FeatureSelectionMethods :
Selected_Features = FeatSelection(All_features $\in K + M$)
Output : Selected_Features $\in K$

the statistical discriminative properties present in the data.

Algorithm 3 SFS

Input: FXTAS data $(X, y) \in \{(X_i, y_i)\}_{i=1}^{N\times(K+M)}$
SFS
Step 1: Initialize Accepted_features, Selected_Features
Step 2: For each Feature i in K:
Step 2.1 : Divide $X_{N,i}$ to $X_{N,i,0}$ and $X_{N,i,0}$ as per labels 0 and 1
Step 2.2 : ACCEPT = two-sample Kolmogorov-Smirnov test ($X_{N,i,0}$, $X_{N,i,0}$)
Step 2.2.1 IF ACCEPT is TRUE : ADD i to Accepted_features
Step 2.2.2 ELSE : Discard i
Step 2.3 Selected_Features = Correlation_reduction(Accepted_features)
Output : Selected_Features $\in K$

B. Recursive Feature Elimination (RFE) [Algorithm 4] This algorithm identifies the most important features by repeatedly training the model with a subset of selected features. The most irrelevant features are gradually eliminated, allowing the most relevant features to be ranked at the top. This method effectively accounts for the non-linear relationships between features.

C. Random Forest Feature Selection [Algorithm 5] We rely on the random forest model's method to decide which features are important by selecting the top 25 % of features. This method is effective at taking the non-linear relationships between the features into account.

D. Lasso Regression Feature Selection [Algorithm 6]
This method is used for sparsity-based feature selection, forcing features to have either 0 or non-zero weights. It essentially fits a linear model with an

Algorithm 4 RFE

Input: FXTAS data $(X, y) \in \{(X_i, y_i)\}_{i=1}^{N\times(K+M)}$
RFE
Step 1: Initialize Accepted_features, Selected_Features
Step 2: Repeat until LEN(Accepted_features) is 10:
Step 2.1 : Train RF1 - RandomForestModel((X, y))
Step 2.2 : Remove Features having low RF1.feature$_i$*importancescores*
Step 2.3 : Add remaining features to Accepted_features
Step 3 : Selected_Features =Accepted_features
Output : Selected_Features $\in K$

Algorithm 5 Random Forest Feature Selection

Input: FXTAS data $(X, y) \in \{(X_i, y_i)\}_{i=1}^{N\times(K+M)}$
Random Forest Feature Selection
Step 1: Initialize Selected_Features
Step 2: Train RF1 - RandomForestModel((X, y))
Step 3: Sort Features according to RF1.feature$_i$*importancescores*
Step 4: Add First 25 % to Selected_Features
Output : Selected_Features $\in K$

L-1 penalty in its loss function. Features with non-zero weights are deemed important by the algorithm.

Each of these feature selection methods gave an optimal subset of features that it considered the most relevant. Machine Learning models are then trained using these respective sets and all features. The performance of each combination of feature set and Model was then recorded to arrive at the most optimal feature selection method and subset.

3. Model Training, Testing, and Hyperparameter Optimization

By using a thorough Grid Search process, we fine-tuned the parameters of each model to find the best one. This careful evaluation ensured we selected the most effective model for our dataset.

We also performed extensive hyperparameter tuning using Grid Search Cross-Validation. By testing a wide range of possible settings, we identified the best hyperparameters that maximized model performance, measured by the highest Cross-Validation AUC-ROC. This approach also helped prevent overfitting during training. Detailed steps of this process are described in

Algorithm 6 Lasso Feature Selection

Input: FXTAS data $(X, y) \in \{(X_i, y_i)\}_{i=1}^{N \times (K+M)}$

Lasso Selection

Step 1: Initialize Selected_Features

Step 2 : Train LR1 - LinearModel$((X, y))$ with L1 penalty

Step 3 : Access LR1.coefficients of each feature

Step 4 : For each feature i:

Step 5 : If LR1.coefficient for i is 0: Drop i

Step 6 : Else Add i to Selected_Features

Output : Selected_Features $\in K$

[Algorithm 7].

Algorithm 7 Grid Search Cross Validation Hyperparameter finetuning

Input: FXTAS data $\{(X_i, y_i)\}_{i=1}^{N\times(K+M)}$

Initialization: Data set $(X, y) = \{(X_i, y_i)\}_{i=1}^{N\times K}$ Number of Selected K Features ; Z Classifiers {SVM,LR,RF,XGBoost} having H hyper parameter combination each

Repeat for other 'Z' ML Algorithms classifiers:

Step 1: Divide (X_s, y_s) into 5 equal random folds

Step 2: Define Grid - 'H' models

Step 3: For each test folds (5 combinations)

Step 3.1 Train the model on 4 folds and test on one fold.

Step 3.2 Record Accuracy, AUROC, Precision, Recall

Step 4: Calculate average AUROC – KCV AUROC

Step 5: Select and Save the Model with the highest average KCV AUROC.

Inference:

1. Retrieve the selected model with maximum KCV AUROC.
2. Do inference and testing on holdout test set.

Output: Inference labels $y_t = \{\hat{y}_j\}_{j=1}^{M}$

6 Results

6.1 Feature selection

Figure 9 illustrates that the feature selection methods proposed in this paper significantly reduce the number of redundant features, improving computational efficiency and predictive accuracy. Specifically, SFS found 7 features, RFE identified 10, Lasso found 26, and Random Forest discovered 44 features .

Consistent features across all methods include:

1. CGG Repeat size (The number of times a particular DNA sequence (cytosine-guanine-guanine) is repeated in the FMR1).
2. Memory and reaction time movement-based features
   - 'RTI Five-choice movement time'
   - '1st Trial Total, ,R+L+B median'
3. Global Severity Index (T-Score)

Additionally, these methods emphasized the importance of slopes in predicting outcomes. Notable slope-related selected features are:

- Stop signal task median score
- 'Calculated Age for Current Visit
- Purdue pegboard right-hand, left-hand score
- Five-choice movement reaction time
- 'Body Mass Index
- Full IQ score

Figure 9: Comparison of number of features selected between Lasso, RFE, SFS, and RandomForest

In all plots in figure 10 individuals without FXTAS (green curve) show less variability and are more centered around zero for the selected features (change in BMI, Full Scale IQ,Stop signal task and 1st trial total slope) analyzed, indicating less change over time. In contrast, individuals with FXTAS (blue curve) show greater variability and a broader spread, this means there is more significant changes in these features over time. The results suggest that FXTAS is associated with greater variability and changes in BMI over time compared to individuals without FXTAS. This might suggest that FXTAS may be associated with metabolic or lifestyle factors leading to greater fluctuations in body weight. The slopes of Full

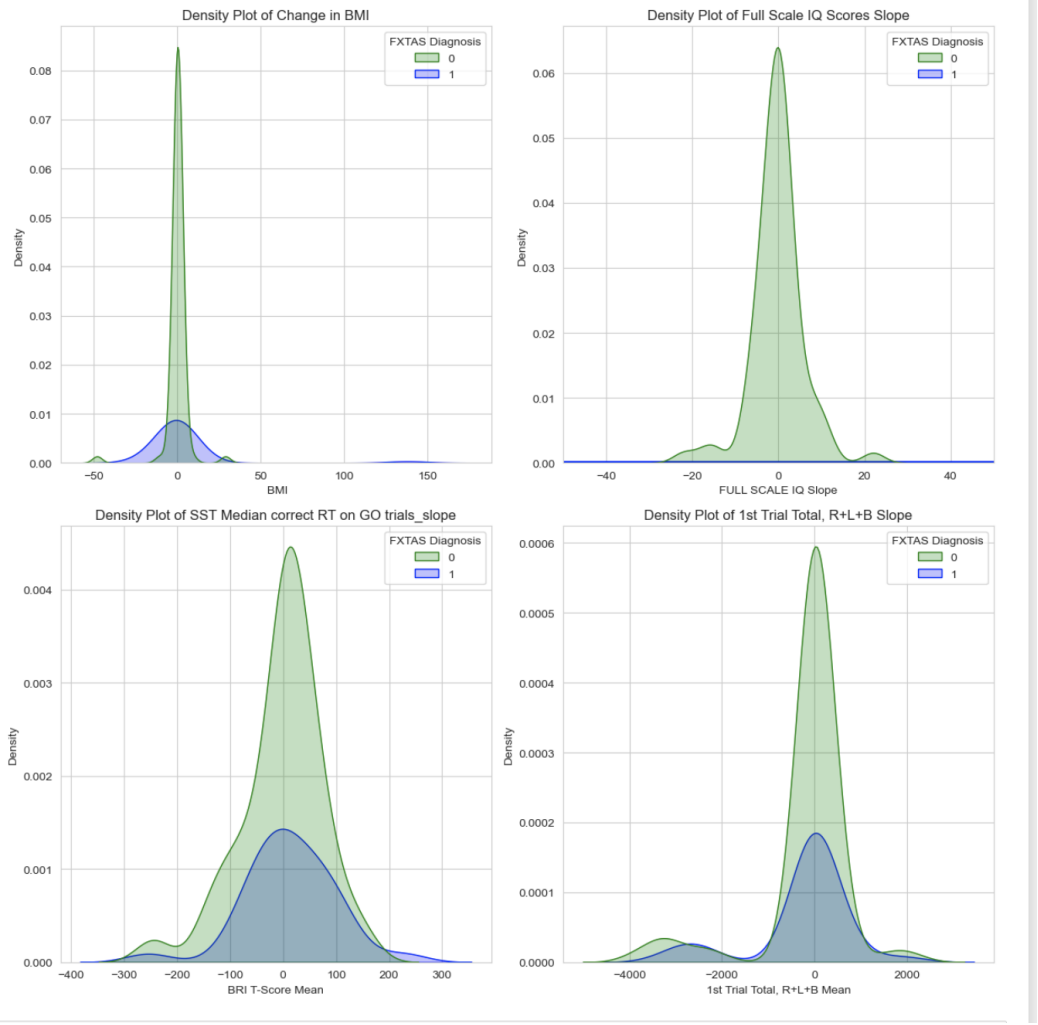

Figure 10: Density plots that compare various features for individuals with and without a diagnosis of FXTAS

Scale IQ scores with FXTAS are more variable compared to those without FXTAS indicating that cognitive decline in FXTAS patients can be highly variable, with some patients experiencing significant changes in their cognitive abilities over time. The reaction times for GO trials in the stop signal task (SST) also shows greater variability for individuals with FXTAS. FXTAS may impair motor response times and cognitive processing speed. The variability in motor task performance (combined scores for right, left, and both hands) also suggests that FXTAS affects fine motor skills and coordination, leading to more pronounced changes over time. To further visualize how these selected features are correlated to each other, we generated a correlation matrix of selected features in figure 11, which displays the correlation coefficients between pairs of variables. The strongest correlation observed is between CGG Repeat Size and RTI Five-choice movement time (0.36). This suggests a moderate relationship between these two variables, meaning that as the CGG Repeat Size increases, the RTI Five-choice movement time tends to increase as well. This correlation may indicate that larger CGG repeat sizes could be associated with slower movement times in the RTI five-choice task, which might reflect an impact on motor function or cognitive processing speed. The remaining correlations are very weak, close to zero, which suggests little to no linear relationship between those pairs of variables.

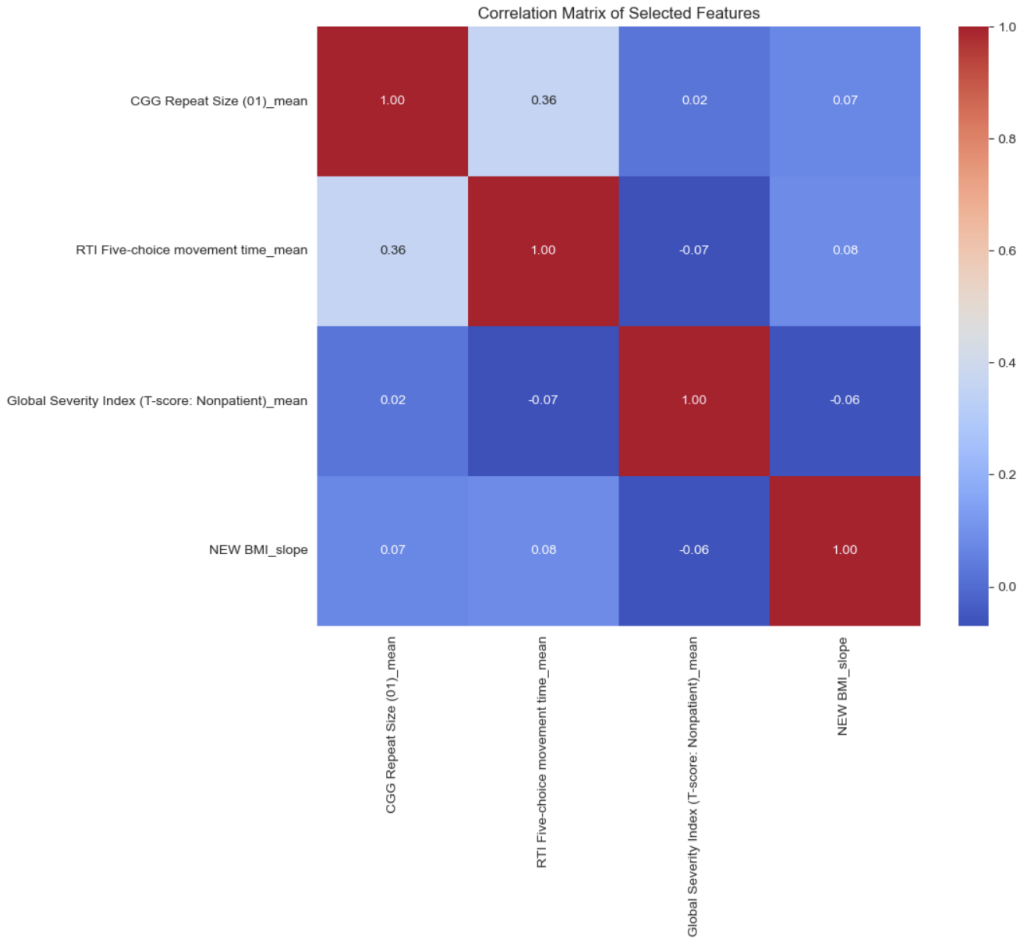

Figure 11: Correlation Matrix of Selected Features: This matrix illustrates the pairwise correlation coefficients between consistent selected features all models. The strength and direction of the correlations are indicated by the color scale, with red representing stronger positive correlations and blue representing negative correlations.

6.2 Evaluation Metrics

Comparative Analysis of Predictive Models for FXTAS In this study, we evaluated several predictive models across different metrics to identify the most effective approach for diagnosing Fragile X-associated Tremor/Ataxia Syndrome (FXTAS). The performance metrics in Table 1 are based on accuracy, precision, recall, sensitivity, specificity, Positive Predictive Value , Negative Predictive Value and AUROC values. The combination of XGBoost and Recursive Feature Elimination (RFE) and the combination of Random Forest and RFE both performed exceptionally well, achieving the highest accuracy of 86.67% , precision 0.86 and AUROC 0.90. It is very important to chose a model with good AUCROC particularly in medical diagnostics, where the cost of misclassification can be high [8]. Additionally, the use of RFE appears to be a common factor in enhancing model performance, as it consistently improved precision and recall across both XGBoost and Random Forest models. This combination is ideal for medical diagnostics situations where both false positives and false negatives need to be minimized such as in patient screening where both false negatives and false positives carry significant consequences.
Feature Selection Impact: The impact of feature selection techniques on model performance was notably significant. Using **all features** without selection generally result in poorer performance. This result shows the importance of proper feature selection in optimizing model accuracy and predictive power.

Table 1: Balanced oversampling trained but tested on imbalanced Performance

Model	Feature Selection	Accuracy	AUROC	Precision	Recall	Sensitivity	Specificity	AUPRC	PPV	NPV
Model (SVM)	SFS	76.67	0.73	0.62	0.56	0.56	0.86	0.55	0.62	0.820
Model (SVM)	Lasso	66.67	0.78	0.45	0.56	0.56	0.71	0.47	0.45	0.79
Model (SVM)	RFE	83.33	0.74	0.75	0.67	0.67	0.90	0.71	0.75	0.86
Model (SVM)	RF	66.67	0.61	0.33	0.11	0.11	0.90	0.33	0.33	0.70
Model (SVM)	All features	63.33	0.76	0.43	0.67	0.67	0.62	0.58	0.43	0.81
Model (Random Forest)	SFS	83.67	0.9	0.78	0.78	0.78	0.90	0.82	0.78	0.90
Model (Random Forest)	Lasso	76.67	0.87	0.60	0.67	0.67	0.81	0.77	0.60	0.85
Model (Random Forest)	RFE	86.67	0.90	0.86	0.67	0.67	0.95	0.85	0.86	0.87
Model (Random Forest)	RF	76.67	0.87	0.60	0.67	0.67	0.81	0.77	0.60	0.85
Model (Random Forest)	All features	76.67	0.88	0.62	0.56	0.56	0.86	0.74	0.62	0.82
Model (XGB)	SFS	83.33	0.9	0.75	0.67	0.67	0.90	0.78	0.75	0.86
Model (XGB)	Lasso	83.67	0.9	0.86	0.67	0.67	0.95	0.84	0.86	0.87
Model (XGB)	RFE	86.67	0.88	0.86	0.67	0.67	0.95	0.80	0.86	0.87
Model (XGB)	RF	76.67	0.88	0.62	0.56	0.56	0.86	0.76	0.62	0.82
Model (XGB)	All features	83.33	0.88	0.83	0.56	0.56	0.95	0.78	0.83	0.83
Model (Logistic Regression)	SFS	66.67	0.78	0.46	0.67	0.67	0.67	0.62	0.46	0.82
Model (Logistic Regression)	Lasso	56.67	0.48	0.17	0.11	0.11	0.76	0.27	0.17	0.66
Model (Logistic Regression)	RFE	70.00	0.77	0.50	0.44	0.44	0.81	0.56	0.50	0.77
Model (Logistic Regression)	RF	66.67	0.70	0.43	0.33	0.33	0.81	0.46	0.43	0.73
Model (Logistic Regression)	All features	60	0.61	0	0	0	0.86	0.35	N/A	.667

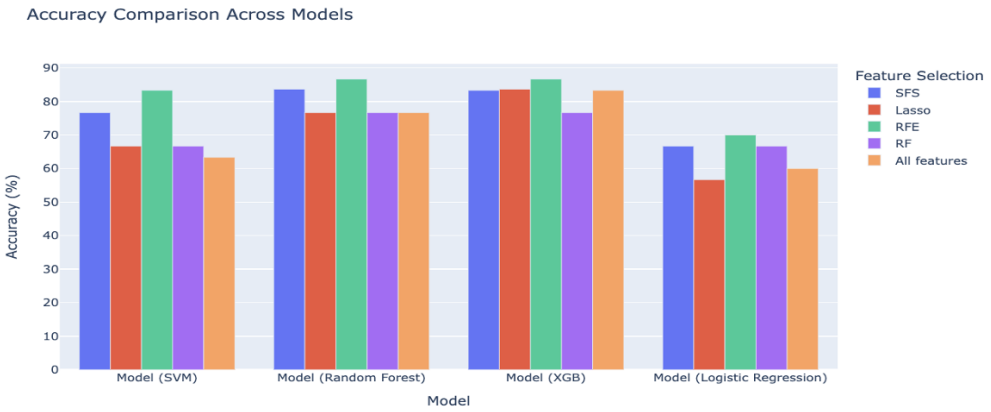

Figure 12: Comparison of Accuracy scores

6.3 Risk Scores

The risk scores represents a composite measure or index of an individual's risk for developing FXTAS. These scores are scaled to a standard range, making them comparable across individuals. In figure 16 demonstrates the correlation between RTI Five-choice movement times and scaled risk scores,. Since, RTI Five-choice movement time is a consistent feature across all model, we wanted too further visualize and see the correlation between the features and scaled risk scores. The positive correlation suggests that individuals with higher RTI Five-choice movement times are at a greater risk of developing FXTAS. This relationship highlights the predictive value of RTI Five-choice movement time as a biomarker for FXTAS risk. Figure 15 illustrates that there is a positive correlation between CGG repeat size and the scaled risk scores for FXTAS. Higher CGG repeat sizes appear to be associated with higher risk scores. This trend supports the hypothesis that larger CGG repeat sizes could be indicative of increased severity or likelihood of developing FXTAS. The scatter in the lower range of CGG repeat sizes suggests variable risk, potentially influenced by other genetic or environmental factors not captured solely by CGG repeat size. In figure 14, the red trend line indicates a negative correlation between risk socres and the Rapid Visual Processing (RVP) score and the Purdue Pegboard trial (PPT) (Right and left hand) which means as the RVP and PPT decreases, the risk score increases. This implies that better

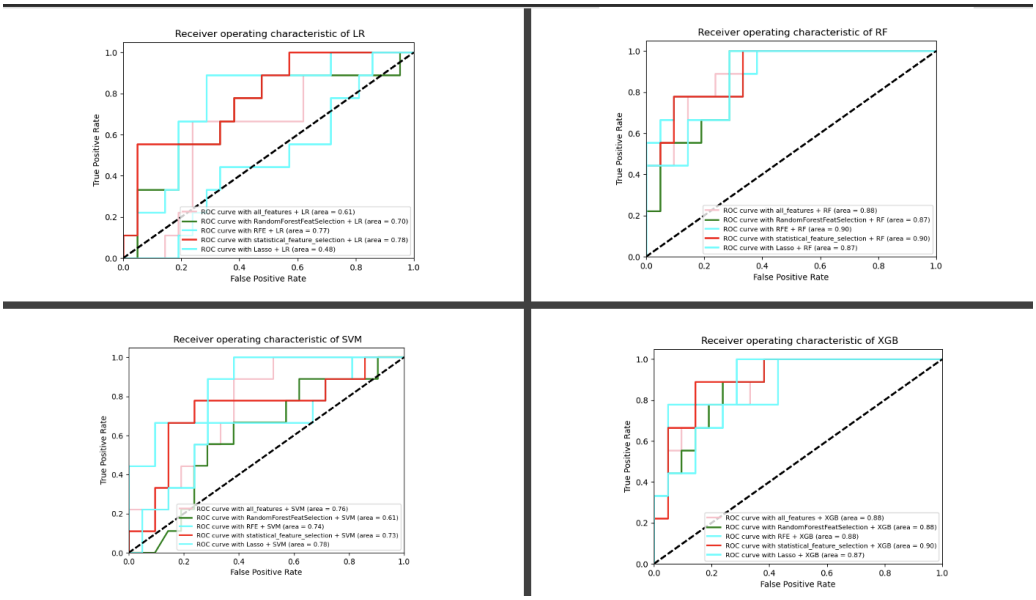

Figure 13: Comparison of AUCROC between Lasso, RFE, SFS, and RandomForest

performance on the RVP test (higher scores, indicating better visual attention and signal detection) is associated with lower risk scores for FXTAS. Similarly, as the 1st trial total increases, the risk score decreases. This suggests that higher scores in the 1st trial total (indicating better manual dexterity and coordination) are associated with lower risk scores. In the context of FXTAS, better manual dexterity might be linked to a lower severity or risk of the condition.

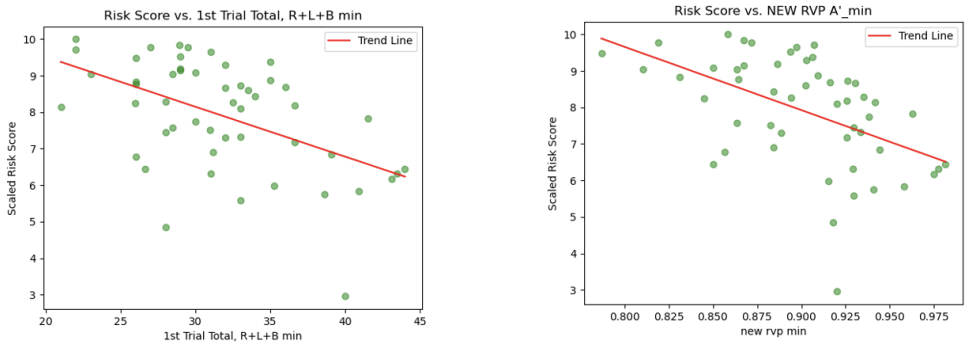

Figure 14: Risk Score vs 1st Trial movement and Risk Score vs New RVP:This scatter plot displays the relationship between the risk score and the new RVP (Rapid Visual Processing) A' min score from the CANTAB (Cambridge Neuropsychological Test Automated Battery)

7 Observation

7.1 Methodological Observation

One of the keys observation from this study is the importance of using the feature selection model. In evaluating diverse machine learning algorithms using complex health medical data and limited samples, the choice of feature selection methods significantly influenced model performance. When we ran the algorithms without the feature selection, it gave us lower accuracy compared the models with feature selection method underscoring the necessity of feature selection and proving our hypothesis. Out of all four feature selections , RFE consistently demonstrated competitive or superior results across algorithms. RFE iteratively trains a tree based model and looks at its feature importance scores to decide which features are relevant. It then repeats the cycle with this new feature set and trains the model again to arrive at a new feature set. This helps it to arrive at an optimal feature set that reduces most multicolinearity and removes redundant features. The optimal feature set helps the model to focus only on the important features. Studies have shown that RFE often outperforms other methods in terms of accuracy, especially in fields like bioinformatics, where selecting the most relevant genes or biomarkers is crucial for disease

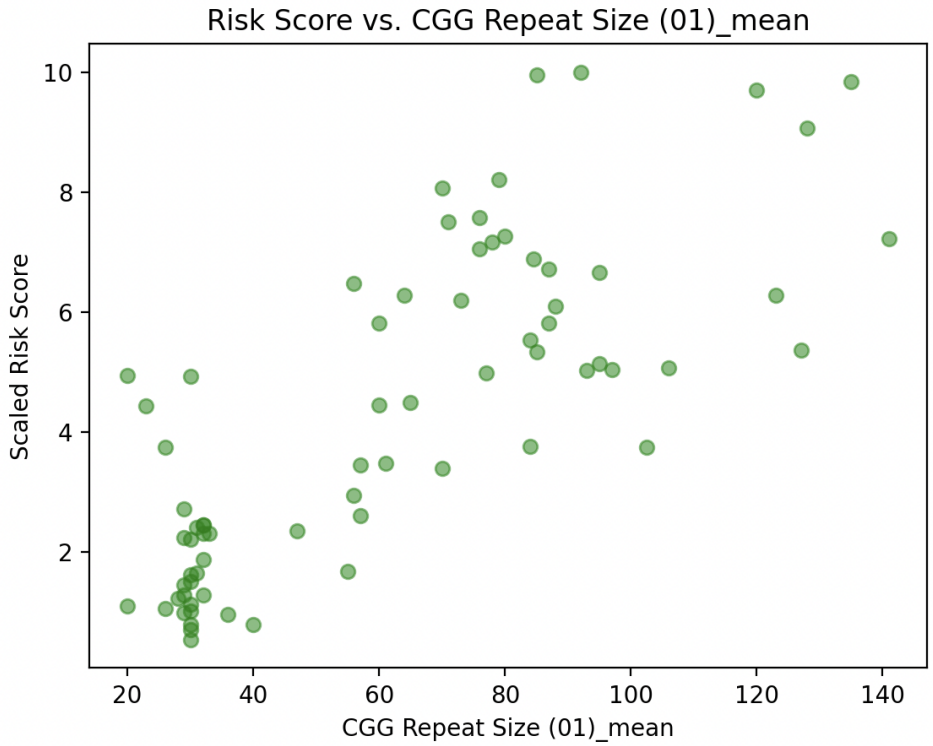

Figure 15: Risk Score vs. CGG Repeat Size - Shows the correlation between CGG repeat size and scaled risk scores

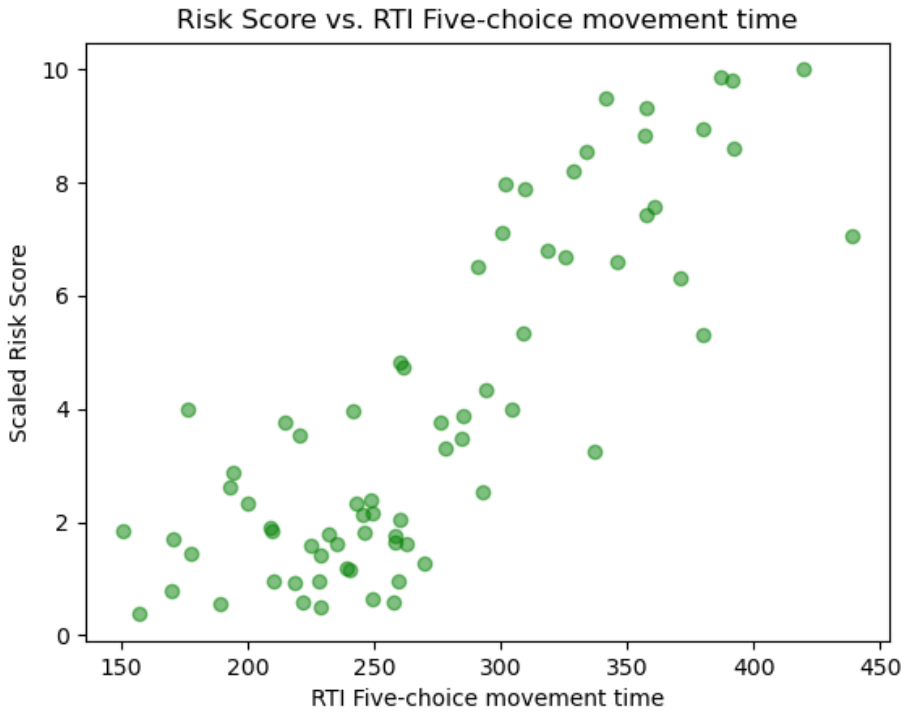

Figure 16: Risk Score vs. RTI Five-choice movement time - Demonstrates the correlation between RTI Five-choice movement times and scaled risk scores, indicating movement time as a predictive factor for FXTAS risk

diagnosis and prognosis [18]. Our analysis revealed that both XGBoost and Random Forest, especially when coupled with RFE, stands out as the most effective model for predicting FXTAS diagnosis within our complex yet limited dataset. XGBoost is robust to noise in the dataset due to its ensemble nature and it utilizes the limited available data very efficiently through its iterative process of boosting, where each subsequent model attempts to correct the errors of the previous one. This iterative approach maximizes the information gained from a limited dataset. Infact, the iterative nature of both XGBoost and RFE could have contributed to their effectiveness. XGBoost iteratively refines its predictions by correcting errors from previous iterations, while RFE iteratively removes the least important features to enhance model performance.This combined iterative process could ensure that the most relevant features are selected and that the model is robust and well-tuned for prediction. XGBoost's iterative boosting process maximizes the information gained from limited datasets by focusing on areas of error correction, while RFE systematically eliminates less important features, streamlining the model and reducing noise.
A key observation here is both XGBoost and Random Forest are tree-based algorithms that build models using decision trees as their base learners. They are also both emsemble methods as Random Forest uses a bagging (Bootstrap Aggregating) approach to create multiple decision trees independently on random subsets of the data and features, and then aggregates their predictions where as XGboost uses a boosting approach to sequentially build trees, where each new tree attempts to correct the errors made by the previous trees.

7.2 Clinical Observation

The feature selection methods results showed consistent features: Stop Signal Task (SST) Median Score and Full IQ Score which are both used to evaluate cognitive functions.The SST assesses response inhibition, a critical aspect of executive function, while the Full IQ Score provides a comprehensive measure of overall intellectual ability. The feature selection result and figure 10 both shows the impact of FXTAS on the cognitive function. Another consistent feature shown was the five-choice Movement Reaction Time and Purdue Pegboard Scores (right-hand and left-hand) measure, which are different aspects of motor skills. The reaction time test assesses the speed and accuracy of motor responses, while the Purdue

Pegboard Test evaluates fine motor dexterity and coordination. Both of these assessments are used for quantifying the motor impairments characteristic of FXTAS. The results also suggest the impact of FXTAS on motor skills, especially manual movement speed and dexterity. These results highlight the significant cognitive and motor impairments in FXTAS patients. The SST median score, reflecting response inhibition, suggests that FXTAS patients have difficulties controlling impulsive actions, which is consistent with observed deficits in executive functioning. Similarly, the prolonged reaction times in the five-choice movement test indicate slower motor responses, aligning with the tremor and ataxia symptoms typical of FXTAS. Another key observation is the importance of changes in age and BMI as important features.

8 Conclusion

Our research made numerous pioneering contributions through extensive data analysis and modeling. It highlighted that genetic factors, age, obesity, motor function, episodic memory, and working memory are significant and sufficient predictors of FXTAS. Comparative studies revealed that iterative filtering of features combined with tree-based ensemble modeling methods can predict FXTAS with high accuracy, providing valuable insights into methodologies for determining the risk of developing FXTAS. The XGBoost model and Random Forest models both consistently outperformed expectations in both the downsampled and oversampled settings. The success of this combination was evinced by the high accuracy scores of 86.67 % in the oversampling setting. An exceptional precision score of 0.86 also highlights the capability of the best-performing model to identify FXTAS correctly. Clinical Application of the model was also demonstrated by identifying key trends in risk scores concerning important clinical variables. This is a helpful indicator for clinicians to know which patients are more susceptible to developing the syndrome. Using these insights and highly accurate predictive models, patients being investigated for other syndromes can also be simultaneously screened and flagged for FXTAS in areas without clinical expertise. Moreover, the number of assessments can be reduced, saving time and resources for doctors, patients, and the healthcare system. If the predictive model identifies certain factors as significant, clinicians can prioritize these factors

in their evaluations. Additionally, patient education and clinical decision support systems can be significantly enhanced based on these findings.

References

[1] Md Manjurul Ahsan, Shahana Akter Luna, and Zahed Siddique. "Machine-learning-based disease diagnosis: A comprehensive review". In: *Healthcare*. Vol. 10. 3. MDPI. 2022, p. 541.

[2] Bojan Bogdanovic, Tome Eftimov, and Monika Simjanoska. "In-depth insights into Alzheimer's disease by using explainable machine learning approach". In: *Scientific Reports* 12.1 (2022), p. 6508.

[3] JR Brouwer, Ralph Willemsen, and BA Oostra. "The FMR1 gene and fragile X-associated tremor/ataxia syndrome". In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 150.6 (2009), pp. 782–798.

[4] Haewon Byeon. "Is the random forest algorithm suitable for predicting parkinson's disease with mild cognitive impairment out of parkinson's disease with normal cognition?" In: *International journal of environmental research and public health* 17.7 (2020), p. 2594.

[5] Wenbing Chang et al. "A machine-learning-based prediction method for hypertension outcomes based on medical data". In: *Diagnostics* 9.4 (2019), p. 178.

[6] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[7] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[8] Tianqi Chen, T He, and M Benesty. "Xgboost documentation". In: *dmlc XGBoost* (2018).

[9] X Fragile. "Foundation". In: *Consensus of the Fragile X Clinical & Research Consortium on Clinical Practices. Medication for Individuals with Fragile X Syndrome. Washington DC, USA. Updated version October* (2012).

[10] CM Greco et al. "Neuropathology of fragile X-associated tremor/ataxia syndrome (FXTAS)". In: *Brain* 129.1 (2006), pp. 243–255.

[11] Randi J Hagerman et al. "Treatment of fragile X-associated tremor ataxia syndrome (FXTAS) and related neurological problems". In: *Clinical interventions in aging* 3.2 (2008), pp. 251–262.

[12] Randi Jenssen Hagerman et al. "Intention tremor, parkinsonism, and generalized brain atrophy in male carriers of fragile X". In: *Neurology* 57.1 (2001), pp. 127–130.

[13] S Jacquemont et al. "Aging in individuals with the FMR1 mutation". In: *American journal on mental retardation* 109.2 (2004), pp. 154–164.

[14] Piers Johnson et al. "Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease". In: *BMC bioinformatics* 15 (2014), pp. 1–14.

[15] C Kavitha et al. "Early-stage Alzheimer's disease prediction using machine learning models". In: *Frontiers in public health* 10 (2022), p. 853294.

[16] Maureen A Leehey et al. "The fragile X premutation presenting as essential tremor". In: *Archives of Neurology* 60.1 (2003), pp. 117–121.

[17] Qian Li et al. "Early prediction of Alzheimer's disease and related dementias using real-world electronic health records". In: *Alzheimer's & Dementia* 19.8 (2023), pp. 3506–3518.

[18] Zifa Li, Weibo Xie, and Tao Liu. "Efficient feature selection and classification for microarray data". In: *PLOS ONE* 13.8 (Aug. 2018), pp. 1–21. DOI: 10.1371/journal.pone.0202167. URL: https://doi.org/10.1371/journal.pone.0202167.

[19] Etienne Minvielle et al. "Current developments in delivering customized care: a scoping review". In: *BMC Health Services Research* 21 (2021), pp. 1–29.

[20] Monika A Myszczynska et al. "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases". In: *Nature Reviews Neurology* 16.8 (2020), pp. 440–456.

[21] I. Ruiz. *Curse of Dimensionality*. `https://cruizbran.medium.com/curse-of-dimensionality-7dbd183aadfd`. Apr. 2021.

[22] Maria Jimena Salcedo-Arellano et al. "Fragile X syndrome and associated disorders: Clinical aspects and pathology". In: *Neurobiology of disease* 136 (2020), p. 104740.

[23] Anurag Tiwari and Amrita Chaturvedi. "A hybrid feature selection approach based on information theory and dynamic butterfly optimization algorithm for data classification". In: *Expert Systems with Applications* 196 (2022), p. 116621.

[24] Matthew Velazquez, Yugyung Lee, and Alzheimer's Disease Neuroimaging Initiative. "Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects". In: *Plos one* 16.4 (2021), e0244773.

[25] Jun Yi Wang et al. "Clinical and molecular correlates of abnormal changes in the cerebellum and globus pallidus in fragile X premutation". In: *Frontiers in Neurology* 13 (2022), p. 797649.

[26] Zukhruf Zain et al. "Leveraging Artificial Intelligence and Machine Learning to Optimize Enhanced Recovery After Surgery (ERAS) Protocols". In: *Cureus* 16.3 (2024).

[27] Wanwan Zheng and Mingzhe Jin. "The effects of class imbalance and training data size on classifier learning: an empirical study". In: *SN Computer Science* 1 (2020), pp. 1–13.

1 Appendix

Table 2 contains details about the performance of different models when they were trained on the Undersampled/Downsampled version of the training data. The training data consisted of 22 positive and 22 negative samples in a class-balanced ratio. And the models were trained according to Algorithm 7 described before. After training, the model were evaluated on test sets that are in an imbalanced ratio of 0.3 for positive to negative samples.

Table 2: Balanced Downsampled Trained Model tested on imbalanced test set

Model	Feature Selection	Accuracy	AUROC	Precision	Recall	Sensitivity	Specificity	AUPRC	PPV	NPV
Model (SVM)	SFS	76.67	0.78	0.60	0.67	0.67	0.81	0.64	0.60	0.71
Model (SVM)	Lasso	30.00	0.50	0.30	1.00	1.00	0.00	0.65	0.30	N/A
Model (SVM)	RFE	80.00	0.82	0.64	0.78	0.78	0.81	0.61	0.64	0.78
Model (SVM)	RF	30.00	0.50	0.30	1.00	1.00	0.00	0.65	0.30	N/A
Model (SVM)	All features	30.00	0.50	0.30	1.00	1.00	0.00	0.65	0.30	N/A
Model (Random Forest)	SFS	73.33	0.89	0.54	0.78	0.78	0.71	0.80	0.54	0.76
Model (Random Forest)	Lasso	76.67	0.86	0.62	0.67	0.67	0.76	0.77	0.62	0.70
Model (Random Forest)	RF	77.67	0.89	0.58	0.78	0.78	0.76	0.8	0.58	0.89
Model (Random Forest)	RFE	80.00	0.92	0.64	0.78	0.78	0.81	0.85	0.64	0.89
Model (Random Forest)	All features	76.67	0.83	0.58	0.78	0.78	0.76	0.70	0.58	0.89
Model (XGB)	SFS	83.33	0.91	0.70	0.78	0.78	0.86	0.79	0.70	0.79
Model (XGB)	Lasso	73.33	0.83	0.56	0.56	0.56	0.81	0.56	0.56	0.65
Model (XGB)	RF	73.33	0.76	0.57	0.44	0.44	0.86	0.67	0.57	0.61
Model (XGB)	RFE	83.33	0.87	0.67	0.85	0.85	0.81	0.68	0.67	0.88
Model (XGB)	All features	73.33	0.84	0.54	0.78	0.78	0.71	0.70	0.54	0.76
Model (Logistic Regression)	SFS	73.33	0.77	0.56	0.56	0.56	0.81	0.57	0.56	0.65
Model (Logistic Regression)	Lasso	66.67	0.63	0.45	0.56	0.56	0.71	0.34	0.45	0.62
Model (Logistic Regression)	RF	66.67	0.57	0.46	0.67	0.67	0.67	0.32	0.46	0.67
Model (Logistic Regression)	RFE	73.33	0.74	0.55	0.67	0.67	0.76	0.46	0.55	0.70
Model (Logistic Regression)	All features	66.67	0.58	0.46	0.67	0.67	0.67	0.32	0.46	0.67