Predicting Rett Syndrome: By leveraging Genetic variation and Phenotypic characteristics

Sulabh Bartaula

Master's Thesis

MEng in Big Data Analytics
2024

Master's Thesis

Sulabh Bartaula

Predicting Rett Syndrome: By leveraging Genetic variation and Phenotypic characteristics

Arcada University of Applied Sciences: MEng in Big Data Analytics, 2024

Identification number

N/A

Commissioned by:

N/A

Abstract:

This thesis is motivated by personal experience with the author's own child, who is affected by Rett syndrome. Rett syndrome is a rare and complex neurological genetic disorder. It is typically observed in girls from birth and tends to manifest by the age of two, occurring in approximately 1 in every 10,000 to 15,000 live female births. The condition is primarily caused by mutations in the MECP2 gene, which is located on the X chromosome. Through this research, we aim to leverage genetic variation and phenotypic characteristics to predict Rett syndrome early and accurately. We will identify and analyze genetic variations, including nucleotide changes, amino acid changes, types of sequence changes, mutations/polymorphisms, and more, to understand how they contribute to the risk of developing Rett syndrome. These predictive models can assist researchers, doctors, and clinicians in identifying the risk of the condition and enhancing the quality of life for individuals with Rett syndrome.

Keywords: Rett Syndrome, Cause, MECP2, Symptoms, Genetic Variation, Nucleotide change, Amino acid, Mutation, Treatment, therapy, Machine learning, Random Forest, XGB, SVM, Neural Networks, Hyperparameter, Model.

FOREWORD

I am motivated to research this topic due to my personal experience with my own child, who has Rett syndrome.

Firstly, I would like to extend my heartfelt thanks to the Rett Syndrome Research Trust, particularly to the Chief Executive Officer, Monica Coenraad, and the Chief Scientific Officer (Head of Clinical Development), Jana Von Hehn, PhD, for their unwavering support in data acquisition, motivation, and providing detailed information about rare and neurogenetic disorders as Rett Syndrome

I would also like to express my sincere gratitude to the Big Data Analytics Team at Arcada University of Applied Sciences, including my study colleagues and teachers, among whom I'd like to specifically mention Dr. Magnus Westerlund, Dr. Leonardo Espinosa Leal, and Dr. Amin Majd. Without their invaluable contributions, this journey would not have been possible. Special appreciation goes to my supervisor, Dr. Leonardo Espinosa Leal, for his scientific expertise and insightful guidance throughout the project.

Lastly, I extend my heartfelt gratitude to my family, especially my parent's Father Ghan Shyam Bartaula, Mother Dal Chini Bartaula, Father-in-law Ananda Nath Paudyal and Mother-in-law Sumitra Kumari Humgai for their unwavering support, and to my friends for their spiritual and mental encouragement throughout this journey. Without their support, I would never have accomplished this feat.

Helsinki, April 2024
Sulabh Bartaula

To my beloved Stella

My daughter is my mentor, my daughter is my strength.
My daughter is the center of my world every day.
With her by my side, I stand tall and proud of her.
She teaches me how to laugh when tunes get tough.
She teaches me to let go of little woes and find happiness.
Even when life throws Challenges.
She's my heart's treasure, my guiding light.
Grateful for her, every day and night.

Abbreviations

RSRT	Rett Syndrome Research Trust
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
MECP2	Methyl CpG Binding Protein 2
MBD	Methyl-CpG Binding Domain
TRD	Transcriptional Repression Domain
NLS	Nuclear Localization Signal
FOXG1	Forkhead-box G1
CDKL5	Cyclin-dependent kinase-like 5
ACGT	Adenine (A), Cytosine (C), Guanine (G), and Thymine (T)
SVM	Support Vector Machine
LR	Logistics Regression
RF	Random Forest
TN	True Negative
FP	False Positive
TP	True Positive
FN	False Negative

Table

Figures

Table of Content

1 Introduction

1.1 Motivation

This thesis is motivated by personal experience with my own loved one daughter, who is affected by Rett syndrome. Witnessing the challenges faced by individuals and families affected by Rett syndrome has ignited a passionate determination within me to make a meaningful impact. Through this research, we aim to leverage genetic variation and phenotypic characteristics to predict Rett syndrome early and accurately. We will identify and analyze genetic variations, including nucleotide changes, amino acid changes, types of sequence changes, mutations/polymorphisms, and more, to understand how they contribute to the risk of developing Rett syndrome. By integrating advanced computational techniques, we seek to develop predictive models that may aid in the early diagnosis or prediction of Rett syndrome. These predictive models can assist researchers, doctors, and clinicians in identifying the risk of the condition and enhancing the quality of life for individuals with Rett syndrome and their families.

1.2 Research Objectives

The main research objective is to develop a predictive model to identify individuals at risk of developing Rett syndrome, a rare neurogenetic disorder.

The Specific Research Objectives of the project are:

- Conduct a comprehensive critique review of existing literature on Rett syndrome, focusing on rare neurogenetic aspects.
- Identify and analyze genetic variations associated with Rett syndrome, including nucleotide changes, amino acid changes, types of sequence changes, mutations, and polymorphisms.
- Investigate phenotypic features linked with Rett syndrome, encompassing physical traits, developmental milestones, neurological symptoms, and behavioral patterns observed in affected individuals.
- Implement and evaluate machine learning models, such as Logistic Regression, Support Vector Machine, and Random Forest, to predict individuals at risk of developing Rett syndrome based on genetic and phenotypic data.

By addressing these specific objectives, the project aims to contribute to the advancement of personalized care and intervention strategies, enhancing the quality of life for those affected by this disorder, including my own beloved child.

2 Background

2.1 Rett Syndrome

Rett syndrome is a complex and rare neurological genetic disorder. It is typically observed in girls from birth and tends to manifest by the age of two, occurring in approximately 1 in every 10,000 to 15,000 live female births ("Rettin oireyhtymä," Rett Syndrome Finland). It is usually caused by mutation in the gene MECP2, which is located on the X-chromosome (Hunter, K. 2007). It is characterized by a period of normal development followed by a loss of acquired skills, like communication skills, including speech and social interaction and the development of severe physical and cognitive impairments. The first signs of Rett syndrome include loss of acquired speech and loss of purposeful hand use for activities such as eating or playing. Individuals may also develop abnormal walking or delay in walking, repetitive hand movements, such as clapping or wringing, and abnormal breathing while awake, sleeping disturbances, teeth chewing and so on. Other behaviors like breathing abnormalities, seizures, and autistic-like behaviors have been observed in Rett children ("About Rett Syndrome," Reverse Rett & ("Description," European Rett Syndrome Association)).

In1966, Dr. Andreas Rett, An Austrian pediatrician initially identified Rett syndrome, which is characterized by same repetitive hand-washing motions and loss of acquired skills the Rett syndrome Handbook Hunter, K. (2007). In 1982, Dr. Bengt Hagberg found similar symptoms and mentioned in his research article details of the disease and Baylor College of Medicine (Houston, TX) named Ruthie Amir discovered MECP2 in 1999, the gene MECP2 when mutated, causes Rett syndrome ("Rettin oireyhtymä," Rett Syndrome Finland).

2.1.1 Symptoms

Symptoms commonly observed in Rett syndrome include:

- Loss of speech
- Loss of purposeful use of hands
- Involuntary or repetitive hand movements handwashing like
- Loss of mobility or gait disturbances
- Loss of muscle tone
- Seizures or Rett “episodes”
- Poor blood circulation
- Breathing issues
- Sleep disturbances
- Slowed rate of growth for head, feet and hands
- Teeth grinding

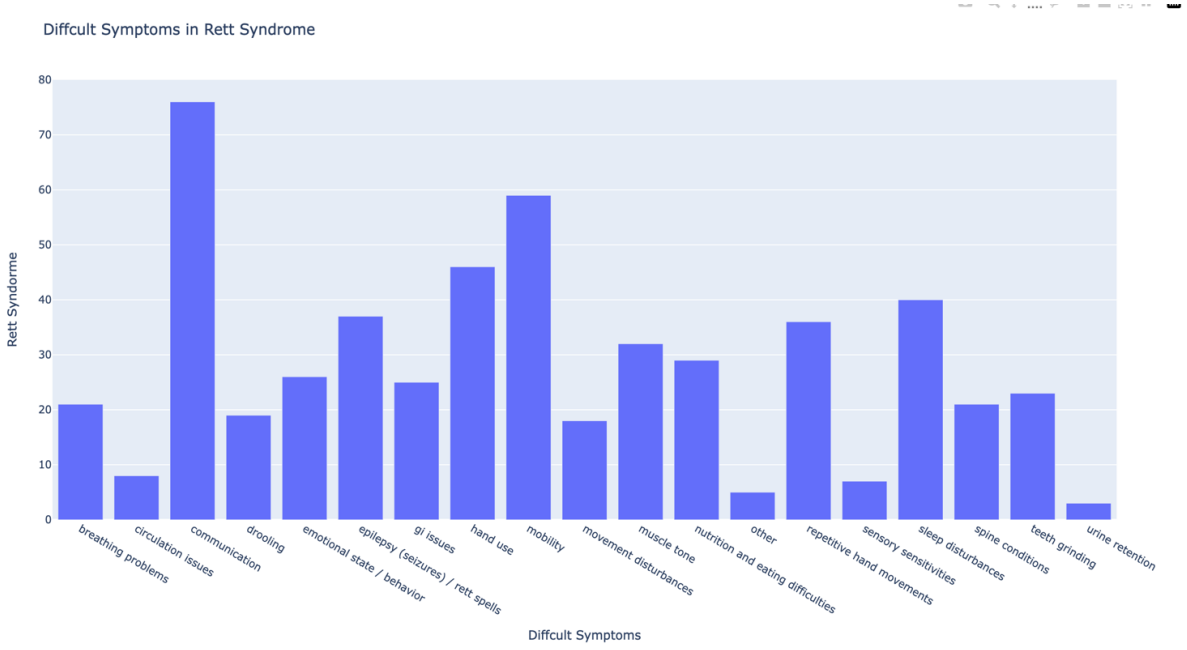


Figure 1. Difficult symptoms observed in individuals Rett patients

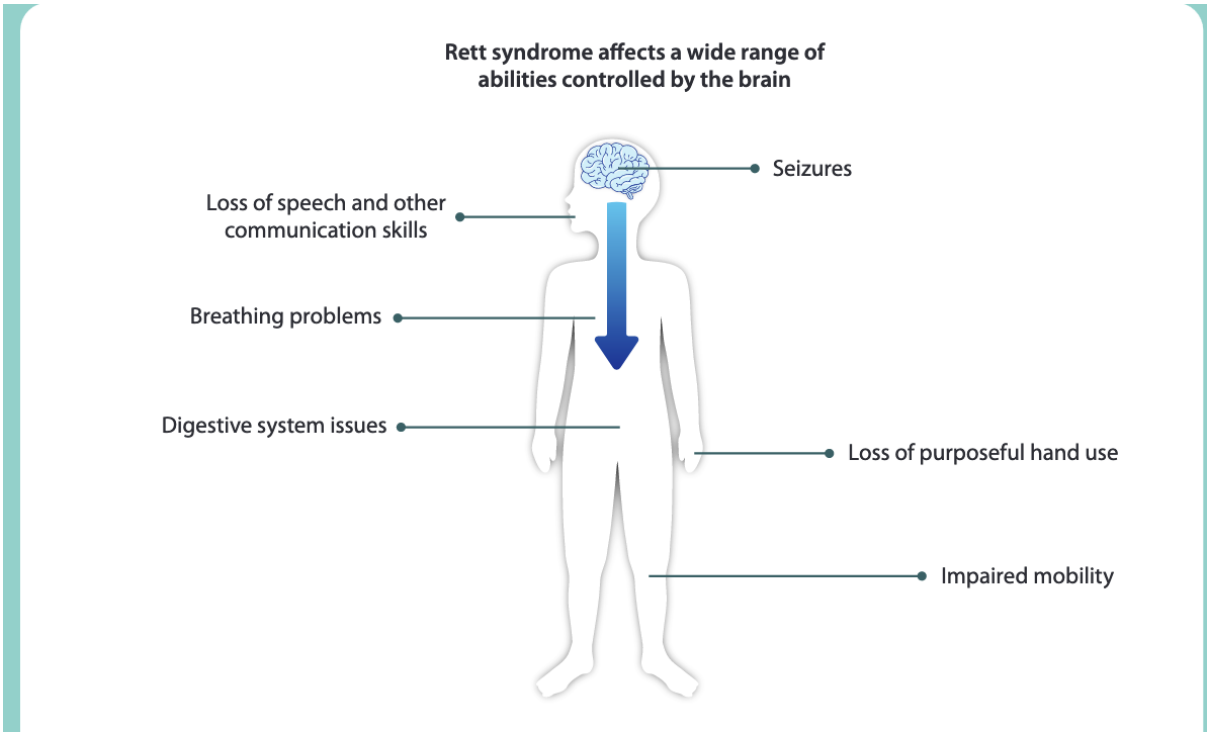


Figure 2. Rett syndrome affects a wide range of abilities controlled by the brain (Neul, Percy, Benke and co-authors)

Individuals with Rett syndrome can have a wide range of symptoms, and not all individuals will experience every symptom listed. Additionally, the severity of symptoms can vary widely among affected individuals ("About Rett Syndrome," Reverse Rett).

The life expectancy of a woman with Rett syndrome was seen around 50 years but with male was lacking. However, there are some unexpected and unexplained death in younger ages due to difficult causes like breathing, seizure, neuro and so on.

As the father of a child with Rett syndrome, I have found that physical activities (such as swimming, walking, interacting with pet animals, and interacting with other children) and recreational activities (like listening to and watching songs, dancing, and communicating) be considered as treatment with Rett syndrome. Each activity is blessed, and it should be done according to the abilities and interests of the individual with Rett. You can always try and push the limits, be surprised and enjoy your day with Rett children and see what they can do. They may need physical assistance, aid or adapted equipment to do physical activities and recreation. This might help them regain or improve their loss abilities like hand function, walking, speaking, breathing, sleep problems, eye- contact and so on.

Activities for children with Rett syndrome include adapted versions of watching TV, listening to music, and interacting with pets. Sensory-friendly outings like farm visits (interaction with pets-household animals), interaction with other children, and swimming cater to their needs. Simple activities such as blowing bubbles or gentle dancing can also be enjoyable (Frequently Asked Questions, European Rett Syndrome Association). Reading with tactile books or audiobooks supports learning. Therapy sessions, like music therapy or hydrotherapy, provide holistic benefits. Careful adaptation and collaboration with therapists ensure meaningful engagement and enhance the quality of life for children with Rett syndrome.

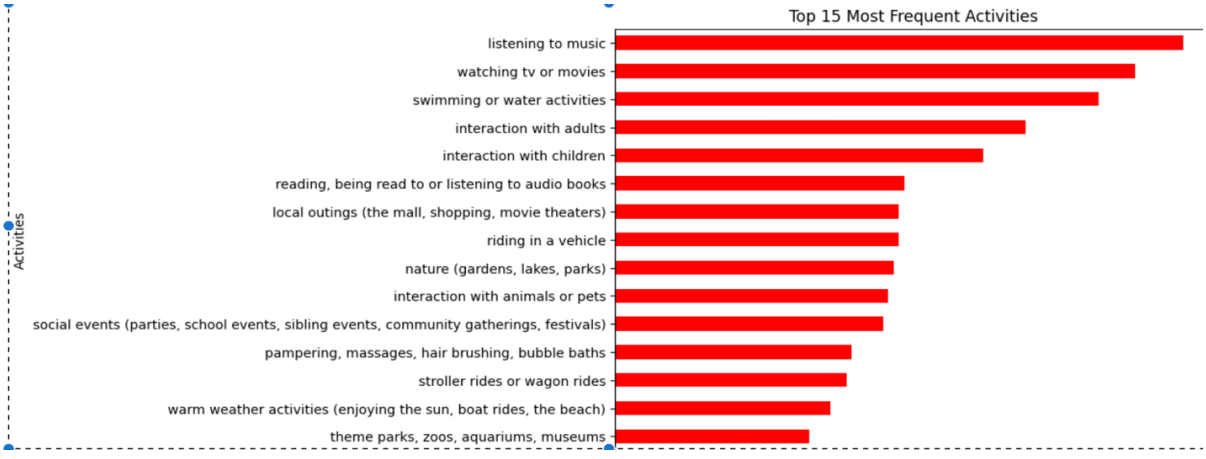


Figure 3. The Most Beneficial Activities for Children with Rett Syndrome

Education for children with Rett syndrome is highly individualized, tailored to their unique abilities, strengths, and challenges. It involves a collaborative effort among educators, therapists, and parents to enrich learning environments and improve quality of life.

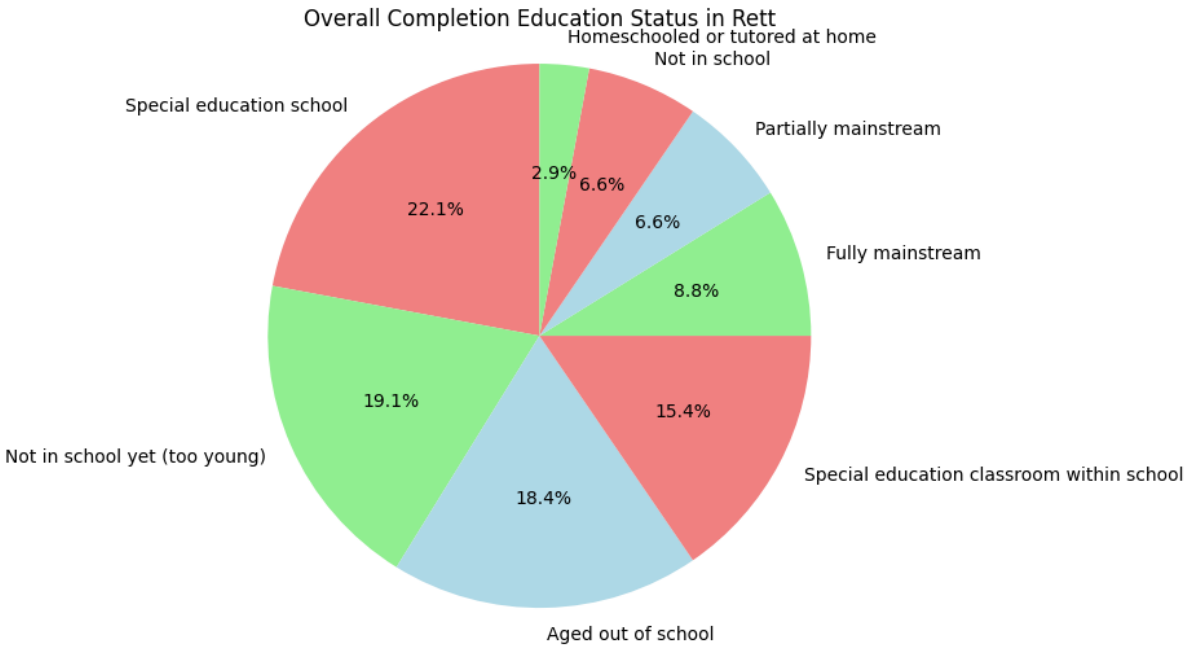


Figure 4. Overall Education status of children with Rett syndrome

2.2.2 Cause

The main cause of Rett syndrome is due to mutation in MECp2 Gene. Rett syndrome is a neuro genetic disorder disease. Every living thing contains DNA, which continues the genetic instruction for the growth, functioning, development, and reproduction of organisms. It is composed of nucleotide bases: adenine (A), cytosine (C), guanine (G), and thymine (T). These base pairs encode the genetic information that determines an organism's traits and characteristics ("Genetics Primer," Reverse Rett).

Figure 5. Double helix DNA structure (RSRT)

From the above figure, we can observe a single strand of DNA is made of letters: **A T G C T C G A A T A A A T G T G A A T T T G A.** Every 3 letters make up a word, also known as amino acids. Many words together make up a gene. The MECP2 gene is made up of almost 500 amino acids.

ATG CTC GAA TAA ATG TGA ATT TGA

DNA resides within the chromosomes and structures in genes. Genes are the basic units of heredity and contain the instructions to assemble single amino acids for producing specific proteins. These proteins make up our entire body, from our muscles to our skin and from our joints to our brain. Each gene consists of coding regions (exons) and non-coding regions (introns), with the coding regions specifying the amino acid sequence of the protein. In human DNA, there are approximately 30,000 genes. Mutations in genes can arise spontaneously or be inherited from parents and can result in alterations to protein structure or function. Mutations in the MECP2 gene represent the primary genetic cause of the Rett Syndrome ("Genetics Primer," Reverse Rett).

As we know MECP2 gene, which is located on the X chromosome, encodes the methyl-CpG-binding protein 2 (MECP2). MECP2 is a multifunctional protein that plays a crucial role during brain development (Rodney C. Samaco and Jeffrey L. Neul. Complexities of Rett Syndrome and MeCP2.) It is highly expressed in neurons and is involved in various neuronal processes, including synaptic plasticity, dendritic maturation, and neuronal survival. It acts as a "reader" of DNA methylation patterns, binding to methylated CpG dinucleotides and modulating chromatin structure and transcriptional activity (Alvaro Hermida-Ameijeiras J. Clin. Med. 2023).

Mutations in the MECP2 gene disrupt the normal function of the MECP2 protein, leading to dysfunction in various neuronal processes (Hunter, K. 2007). These mutations can include missense mutations (altering a single amino acid), nonsense mutations (resulting in premature protein truncation), frameshift mutations (insertion or deletion of nucleotides, leading to a change in the reading frame), or large deletions affecting multiple exons. The specific type and location of the mutation within the MECP2 gene can influence the severity and clinical presentation of Rett syndrome. Dysfunction of the MECP2 protein disrupts normal brain development and function, leading to the characteristic features of Rett syndrome (Alvaro Hermida-Ameijeiras J. Clin. Med. 2023).

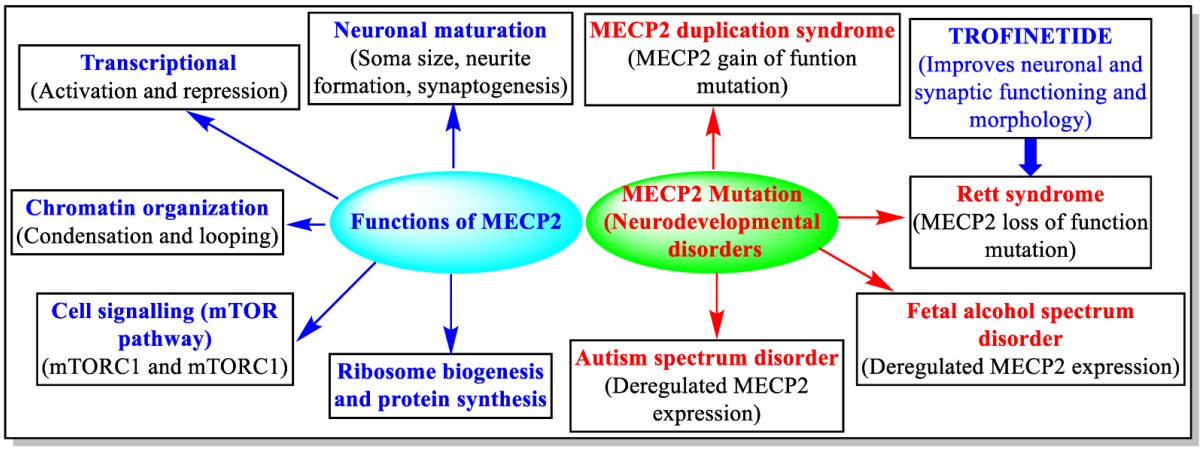


Figure 6. Functions and disorders associated with MECP2 mutations. (Alvaro Hermida-Ameijeiras J. Clin. Med. 2023)

2.3.1 Types and phases of Rett syndrome

There are two types of Rett Syndrome: "a classic" or "atypical variant" Rett syndrome. Both differ by their symptoms or by the specific gene mutation ("Types of Rett Syndrome," National Institute of Child Health and Human Development).

A classic Rett Syndrome

The majority of Rett syndrome patients have the classic form, which typically can be seen at the age between 6-18 months of Rett child. It can be characterized by symptoms of hand-movement, language development, eye-contact, breathing and so on. The four phases of a classic Rett syndrome are as follows:

Early stage: Early stage begins between 6 and 18 months of age. During this stage, a child with Rett syndrome may start to show signs of delayed development, slowly decreased eye-contact, sleeping problems, delayed motor skills and no interest in playing with toys.

Rapid Destructive Phase or Regression stage: This phase usually occurs between ages 1 to 4. It is identified by loss of skills acquired earlier like speech (words like Dad, Mom, Di, Moi, Aama, Tulle, Isi, ääiti, Syö and so on), walk or ability to crawl. Children may also develop repetitive hand movements, breathing irregularities, and social withdrawal during this stage.

Plateau Phase: In this stage, which typically begins between ages 2 and 10, the symptoms of Rett syndrome stabilize, some children may have improvement in some area like language, eye-contact, motor skills and so on. Seizures and movement problems are common at this stage.

Late Motor Stage: This phase usually begins in adolescence, around age 10 or older. Individuals with Rett syndrome may experience a gradual decline in motor function, leading to increased mobility difficulties and the development of problems such as muscle stiffness and spasticity are seen at this stage ("Types of Rett Syndrome," National Institute of Child Health and Human Development).

atypical Rett syndrome

There are currently five known variants of atypical Rett syndrome, defined by characteristic symptoms, age at which the symptoms present, or genetic.

Around 15-20% of Rett syndrome patients manifest a non-classic or atypical, if their symptoms do not meet all the diagnostic criteria for the disease, it is considered atypical. Atypical forms of the disease may be milder or more severe than the classic type. Several subtypes of atypical Rett syndrome have been defined; these are summarized below.

Preserved Speech Variant: In this variant, individuals with Rett syndrome maintain the ability to speak or communicate verbally to some extent or couple of words. While they may still experience other symptoms of Rett syndrome, affecting motor skills or hand stereotypies, breathing and so on.

Congenital Rett Syndrome: In this variant, symptoms of Rett syndrome are seen from birth to infant which often presents developmental delays and may have similar clinical features compared to classic Rett syndrome.

Zappella Variant: This variant is named after Italian physician Alessandro Zappella, It can be seen on early stage before 6 months. A higher prevalence of epilepsy can be seen.

Hanefeld Variant: This variant, also known as the early seizure variant, is characterized by severe developmental delays and complication of frequent seizures. This variant may have symptoms like classic Rett syndrome.

Other Rare Variants: There are other rare variants of Rett syndrome that may present with atypical features or clinical manifestations. These variants may involve mutations in genes other than MECP2 or have unique combinations of symptoms (Hunter, K. 2007).

2.4.1 Diagnostics

Diagnosing Rett syndrome typically involves a combination of clinical evaluation, medical history assessment, and genetic testing.

Clinical Evaluation

Clinical Evaluation in Rett syndrome is based on a period of normal development followed by a loss of purposeful hand skills and the development of repetitive hand movements, along with other characteristic features. Healthcare professionals observe the individual's behavior and physical symptoms, such as repetitive hand movements, motor abnormalities, breathing irregularities, and loss of acquired skills (Rett Syndrome: Diagnosis & Treatment, Mayo Clinic).

Medical History

Medical History A detailed developmental history is obtained, focusing on milestones and any regression or loss of skills. From family history information, particularly regarding developmental disorders or neurological conditions, may be relevant (RSRT).

Genetic Testing

The main cause of Rett Syndrome is associated with mutation in the MECP2 gene, in the X chromosome. Individuals with Rett-like symptoms may have mutations in genes other than MECP2, such as CDKL5 and FOXG1. Genetic testing, such as DNA sequencing, is performed to identify mutations in this gene (Genetics Primer, Reverse Rett).

It is always important for the diagnostic process to be thorough and collaborating with teams like healthcare professionals, including neurologists, geneticists, and developmental

specialists. Early and accurate diagnosis is important for implementing appropriate interventions and support for individuals with Rett syndrome and their families.

Treatment

The treatment of Rett syndrome is to be planned based on the symptoms seen in each Rett child. Treatment may require a team coordinate of specialists Pediatricians, pediatric neurologists, gastroenterologists, speech therapists, psychiatrists, nutritionists, and other healthcare professionals. Various therapeutic modalities, including symptomatic treatment, gene therapies, and Daybue medicine (Alvaro Hermida-Ameijeiras J. Clin. Med. 2023), are utilized to manage symptoms and promote well-being. Importantly, psychosocial support for the entire family is integral to the treatment process. It might help by providing emotional support, education, and resources to the families navigate the challenges of living with Rett syndrome and promote resilience and well-being.

Symptomatic Treatment

Symptomatic treatments remain important for managing specific symptoms and improving quality of life. These may include medications for seizures, gastrointestinal issues, sleep disturbances, and behavioral problems commonly associated with RTT. Drugs may be used to treat a variety of symptoms associated with Rett syndrome including seizures, anxiety, sleep disturbances, breathing problems, stereotypic hand movements, and certain gastrointestinal abnormalities and so on (Rett Syndrome: Diagnosis & Treatment, Mayo Clinic).

2.4.2 Gene therapy and small molecule therapies

Gene therapy and small molecule approaches, other innovative strategies such as antisense oligonucleotide therapy, RNA-targeted therapies, and neurotrophic factor- are used for treatments for Rett syndrome (Gene Therapy, Reverse Rett). These approaches aim to restore MeCP2 function at the molecular level, mitigate disease progression, and improve neurological outcomes in individuals with Rett syndrome by delivering a functional copy of the MECP2 to affected cells or tissues (Hunter, K. 2007).

2.4.3 Medicine

In 2023, **Trofinetide** (Daybue) was approved by the U.S. Food and Drug Administration (FDA) as the first treatment for Rett syndrome in adults and children 2 years of age and older. Trofinetide is a novel synthetic analog of [glypromate], also known as glycine-proline-glutamate (GPE), a naturally occurring protein in the brain and the N-terminal tripeptide of insulin-like growth factor 1 (IGF-1). Patients take Daybue either orally or via gastrostomy tube. Daybue is taken twice a day, in the morning and in the evening, with or without food and dose is based on patient weight (Alvaro Hermida-Ameijeiras J. Clin. Med. 2023).

Trofinetide is designated chemically as (2S)-2-{[(2S)-1-(2-aminoacetyl)-2-methylpyrrolidine-2carbonyl] amino}pentanedioic acid (IUPAC). Its empirical formula is C13H21N3O6, and its molecular weight is 315.33 g/mol. The chemical structure is:

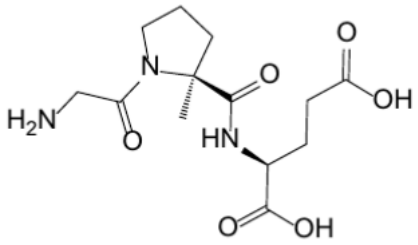


Figure 7. Organic structure of Trofinetide (Daybue)

2.4.4 Treatment therapy

Most Rett syndrome children will benefit, as improvement on quality of life-form occupational, physical, speech, rehabilitative and behavioral therapy. Lastly, psychosocial support for the entire family is essential too (Treatments for Rett Syndrome," National Institute of Child Health and Human Development).

Physical Therapy: Physical therapy can help individuals maintain mobility, improve muscle tone, and manage motor difficulties.
Occupational Therapy: Occupational therapy aims to enhance daily living skills and independence, focusing on activities like self-care, communication, and fine motor skills.
Speech Therapy: Speech therapy can be beneficial for those with communication difficulties. Alternative forms of communication, such as assistive devices like reading with tactile books or audiobooks supports on communication.
Behavioral Interventions: Behavioral strategies may be employed to manage challenging behaviors and improve the individual's quality of life.
Supportive Care: A supportive and multidisciplinary approach involving a team of healthcare professionals, including neurologists, developmental specialists, and others, can help address the complex needs of individuals with Rett syndrome.
Family Support: Families of individuals with Rett syndrome often benefit from educational and emotional support. Support groups and counseling services can provide valuable resources and assistance.

2.2 Machine Learning in Rett Syndrome

In healthcare, advanced tools and technologies have continuously evolved. Today, machine learning (ML) plays a crucial role in improving patient care by organizing data and identifying trends. In Rett syndrome, machine learning (ML) utilizes algorithms and techniques to analyze clinical and genetic data, revolutionizing our approach to diagnostics and treatment (Machine Learning Enhances Care for Rett Syndrome, Emory University, 2023). ML algorithms run through extensive big data sets, identifying patterns and associations crucial for understanding and managing the condition effectively. ML algorithms are used in various areas related to Rett Syndrome, including:

Diagnostics: ML algorithms assist in accurately identifying Rett syndrome by analyzing clinical and genetic data, aiding in early diagnosis and intervention.

Genomic Analysis: ML algorithms analyze large genomic datasets to identify genetic variations like Nucleotide changes, Amino acid changes, types of sequence changes, mutation, domain location changes, associated with Rett syndrome, facilitating personalized treatment strategies based on individual genetic profiles (EURORDIS – Rare Diseases Europe Introduction to genetics and genomics).

Drug Discovery: ML accelerates the drug discovery process by analyzing biological data to identify potential therapeutic compounds and predict their effectiveness in targeting specific molecular pathways involved in Rett syndrome.

Biomarker Discovery: ML analyzes diverse biological data to identify biomarkers indicative of Rett syndrome's status and progression, aiding in early diagnosis, disease monitoring, and treatment evaluation.

Clinical Decision Support: ML-powered clinical decision support systems empower healthcare providers to make informed decisions by analyzing patient data and medical literature. ML algorithms deliver evidence-based recommendations for diagnosis, treatment planning, and patient management, enhancing the quality of care for individuals with Rett syndrome.

In conclusion, the integration of machine learning (ML) into the management of Rett syndrome marks a significant advancement in healthcare. By leveraging ML algorithms to analyze clinical and genetic data, healthcare professionals can improve diagnostic accuracy, tailor treatment strategies, and enhance patient care. ML's ability to identify patterns and genetic variations crucial for understanding the condition has paved the way for personalized medicine approaches in Rett syndrome.

3 Research Methodology

3.1 Data Collection

The genetic variations and phenotype co-relations data sourced from the Rett Syndrome Research Trust (RSRT). The RSRT is a non-profit organization dedicated to accelerating research to cure Rett syndrome and related MECP2 disorders. The data collection process

involved gathering information on genetic variations observed in individuals with Rett syndrome. The source of DNA used for genetic analysis in RTT research and clinical testing includes blood, brain tissue, fibroblasts, and other biological samples. The techniques or methods used to detect genetic variations or mutations MECP2 are (qPCR, array-CGH, QF-PCR, MLPA, DHPLC, Unknown, Direct for exon 1, MLPA for exons 1-4, Direct, SSCP, RT-PCR, DGGE, ECMA, CSGE, PCR/diagnostic restriction, Direct for CDKL5, direct and MLPA for MECP2, qRT-PCR and so on). The key aspects on gene variation in Rett syndrome are as nucleotide changes, amino acid alterations, type of sequence changes, mutation/polymorphism classification, domain change locations, phenotypic manifestations, inheritance patterns, extent of coding region screened and so on. Here's a breakdown of the data collection process based on the columns:

3.1.1 Nucleotide Change

Nucleotide indicates the change at the nucleotide level. Nucleotides are the building blocks of DNA, consisting of a phosphate group, a sugar molecule (deoxyribose in DNA), and one of four nitrogenous bases: adenine (A), thymine (T), cytosine (C), and guanine (G). Nucleotide change provides detailed information about the change in DNA sequence like substitutions, deletions, insertions or duplications of nucleotides, which can have implication for genetic variation. Some values from Nucleotide changes columns are as follow:

- **c.-395G>T** indicates a substitution of nucleotide G with nucleotide T at position 395 in the DNA sequence.
- **c.-206_-205delGC** represents a deletion of nucleotides GC starting from position 206 to 205 in the DNA sequence.
- **c.-168-?_*?dup** indicates a duplication of an unknown number of nucleotides starting from position 168 in the DNA sequence.
- **c.-168-?_26+?del** (deletion exons 1 and 2): Deletion of exons 1 and 2, spanning from an unspecified position upstream to position 26 downstream.
- **c.-167_-99del**: Deletion of a sequence from position 167 to position 99 upstream of the coding sequence.
- **c.-160A>T**: Substitution of nucleotide A with T at position 160 upstream of the coding sequence.

3.1.2 Amino Acid changes

Amino acid describes the resulting change at the amino acid level. Amino acids are the building blocks of proteins, and the sequence of amino acids determines the structure and function of a protein. When a nucleotide change occurs in the DNA sequence, it can lead to a change in the corresponding amino acid in the protein sequence during translation. This change can be classified into several types, such as missense mutations, nonsense mutations, frameshift mutations, or silent mutations, depending on the specific alteration and its impact on the resulting protein (Genetics-primer, Reverse Rett). Studies have shown that missense mutations, particularly those affecting critical functional domains of MeCP2 in Rett Syndrome. For example, Amino Acid changes in Rett patients are as follow:

- **MeCP2_e1: p.G19AfsX28:** This indicates a frameshift mutation in the MeCP2 protein at position 19, resulting in a change from glycine (G) to alanine (A), followed by a premature stop codon (fsX28).
- **MeCP2_e1: p.E17K**: This indicates a missense mutation in the MeCP2 protein at position 17, where glutamic acid (E) is replaced by lysine (K).
- **Intronic variation:** These entries indicate variations occurring within intronic regions of the DNA, which typically do not result in changes to the amino acid sequence but may still have functional consequences.
- **p.M1?**: This denotes uncertainty about the amino acid change at position 1 of the protein, often indicating that the mutation occurs at the start codon (methionine, M) but the specific alteration is not known.
- **5'UTR variation**: These describe variations in the untranslated regions (UTRs) at the 5' end of mRNA transcripts, which can affect gene expression but do not alter the amino acid sequence directly.
- **p.(N126+S486)fs**: This indicates a frameshift mutation leading to the insertion of asparagine (N) at position 126 followed by serine (S) at position 486, resulting in a frameshift and potential premature stop codon.

3.1.3 Type of Sequence Changes

Type of sequence specifies the type of mutation in the genetic sequence. The type of sequence like 5'UTR variation, In-frame insertion or deletion, Exon deletions / Exonic deletions, Missense, Splicing, Intronic variation, Silent, non-sense, complex rearrangement, Exonic duplication, nonstop are observed in DNA sequence.

- Missense mutation: A single nucleotide that changes one amino acid into a different amino acid.
- Nonsense mutation: A single nucleotide changing an amino acid into a stop codon, leading to early termination of protein synthesis.
- Frameshift mutation: An insertion or deletion that changes all of the amino acids downstream of the mutation.
- Silent mutation: A nucleotide change that does not result in any change to the amino acid sequence due to redundancy in the genetic code.
- Miscellaneous mutation: A mutation that does not fit into the above categories, or a compound mutation.
- Intronic variation: Variations occurring within introns, which are non-coding regions of genes

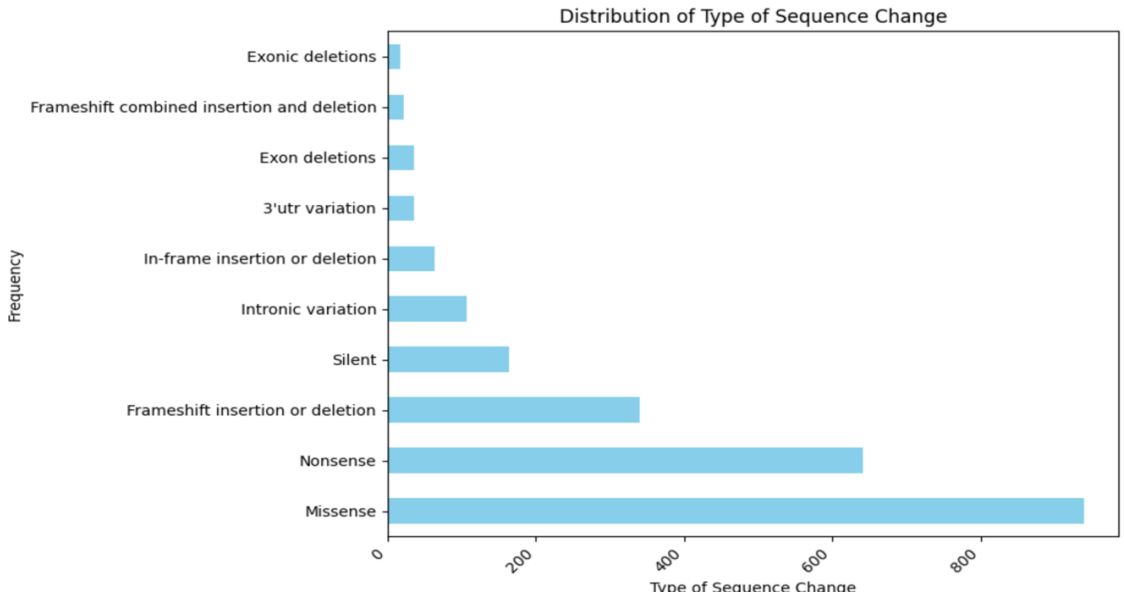


Figure 8. Distribution of type of sequence change

The figure shows the distribution of mutation type seen in data source. Missense, Nonsense, Frameshift, Intronic, and Exonic mutations in the MECP2 gene can cause Rett syndrome (Lobo, R. A., Vieira, I. C., Medeiros, A. M., Brito, J. S., Souza, B. S., & Araújo, C. G., Insights into Rett Syndrome,2021). These mutations alter the gene's instructions, leading to the characteristic symptoms of the condition. Larger insertions, duplications, or deletions of genetic material can also contribute to Rett syndrome. However, silent mutations, which don't change the resulting protein, typically do not cause Rett syndrome.

3.1.4 Mutation/Polymorphism

Genetics changes into different types of sequence, each indicating whether the change is associated with disease (mutation) or not (polymorphism). Based on dataset, Examples might include missense mutations, frameshift mutations, or nonsense mutations that disrupt the normal functioning of a protein cause disease and silent mutation not causing disease Hunter, K. 2007).

3.1.5 Domain Change Locations

Domain change location describes the location of a mutation or variation within protein to distinct structural or functional units. Domain change locations within the MECP2 protein, such as the methyl-CpG-binding domain (MBD) and transcriptional repression domain (TRD), are critical determinants of MeCP2 function. Other domain change locations are 5'UTR, N-term, Inter-domain region, 3'UTR, Unknown, C-term, Intronic, NLS and so on which can help in understanding protein function, structure and disease pathology (Hunter, K. 2007).

MBD (Methyl-CpG Binding Domain): The mutation occurs within the MBD, a domain found in proteins that bind to methylated DNA.

TRD (Transcriptional Repression Domain): The mutation occurs within the TRD, a domain involved in transcriptional regulation by repressing gene expression.

NLS (Nuclear Localization Signal): The mutation occurs within the NLS, a sequence that directs proteins to the cell nucleus.

N-term/MBD/interdomain/TRD/NLS/C-term: Indicates mutations occurring across multiple domains, including the N-terminal domain, MBD, inter-domain regions, TRD, NLS, and C-terminal domain.

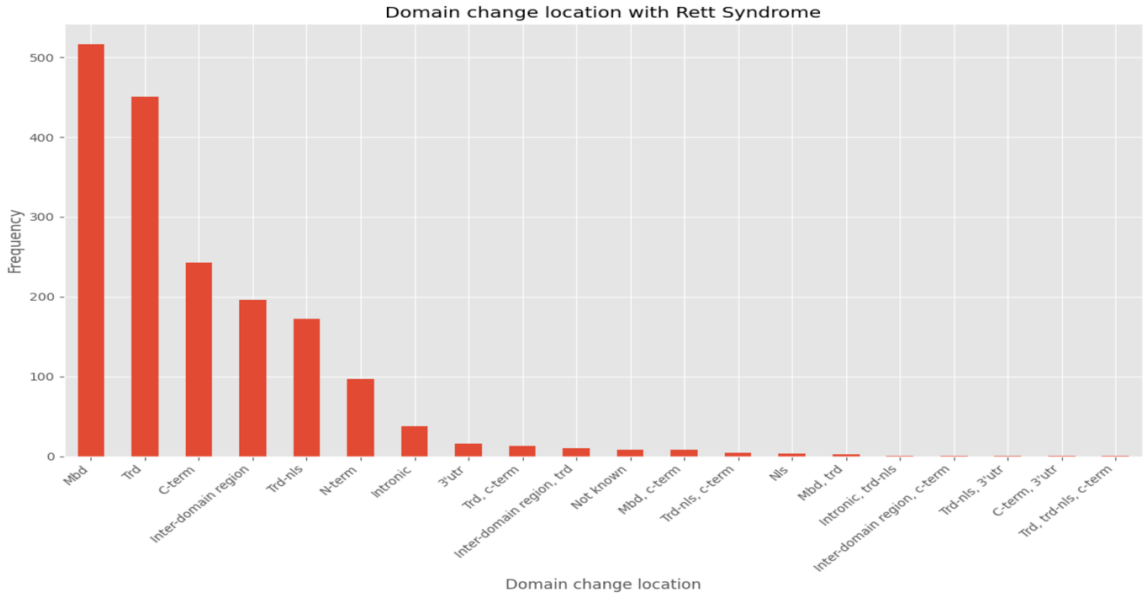


Figure 9. Domain change location in Rett Syndrome

3.1.6 Phenotype

Phenotype describes different phenotypic presentations associated with Rett syndrome, a neurodevelopmental disorder. It may vary widely depending on the nature and location of MECP2 mutation like Rett syndrome – classical, Atypical congenital, Male variant, forme fruste, Preserved speech, late regression, Congenital onset, Rett syndrome – unknown.

3.2 Data Preprocessing

Data preprocessing is a crucial step in the machine learning pipeline that involves preparing and cleaning raw data to make it suitable for analysis and modeling. We used data cleaning steps like checking and removing duplicates, handling and missing values, dropping irrelevant columns, identifying and correcting errors or inconsistencies in a dataset. This process ensures the dataset is clean and improves the quality, reliability and integrity of the data (Michael Zats, Feature extraction).

Data extraction is a process of extracting, manipulating and transforming useful features from raw data. We have used label encoding for converting categorical variables into numerical representations. We have used domain knowledge to extract features (characteristics, attributes, properties) from raw data.

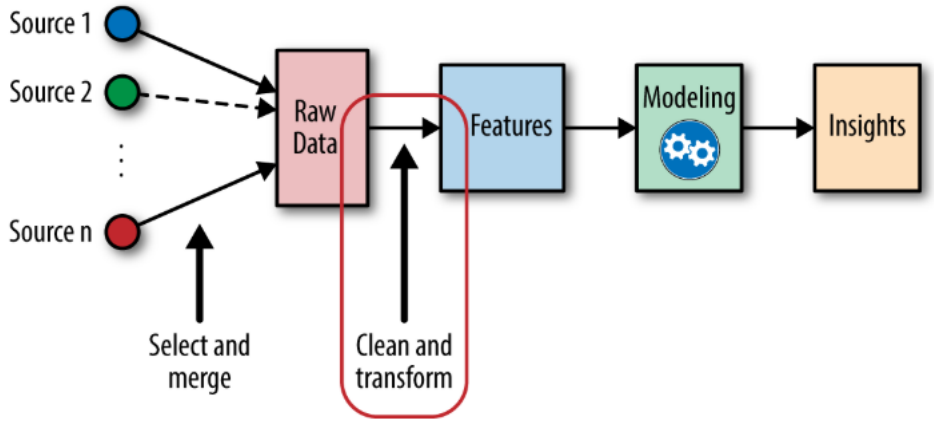


Figure 10. Process of feature extraction (Michael Zats, Feature extraction)

We have used **input field selection** as **Nucleotide change, Amino acid change, Type of sequence change, Mutation/polymorphism, Sex, Patient ID and domain change location.**

And target field as new column **Rett Syndrome** extracted from the phenotype.

3.3 Data Exploration and Visualization

Demographics of Rett syndrome: It is estimated that one in every 10 000 to 15 000 girls are born with Rett syndrome. This means around 360,000 girls and women in the world have the Rett syndrome disorder. The incidence of Rett syndrome in males is currently unknown. The most cases of Rett syndrome are sporadic (91%) and are not inherited from parents. However, Rett syndrome is Familia (9%) can be inherited in an X-linked dominant pattern, with affected fathers passing the mutated gene to their daughters.

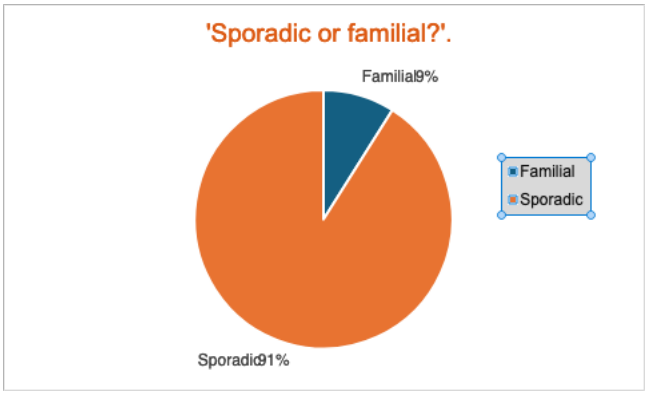


Figure 11. Sporadic or familial in Rett syndrome

Amino Acid changes (top 30): Variations leading to changes in the amino acid sequence of the MeCP2 protein in Rett syndrome are observed in the below figure. like P.T158M means missense mutation occurs in Mecp2 at position 168 due to protein change from threonine to methionine, **P.R168X** Indicates a nonsense mutation resulting in a premature stop codon at position 168. **p.M1**: Indicates a change affecting the methionine residue at position 1. **p.A2V**: Indicates an alanine to valine substitution at position 2. **p.A6_A8dup**: Indicates a duplication of the alanine residues at positions 6 to 8.

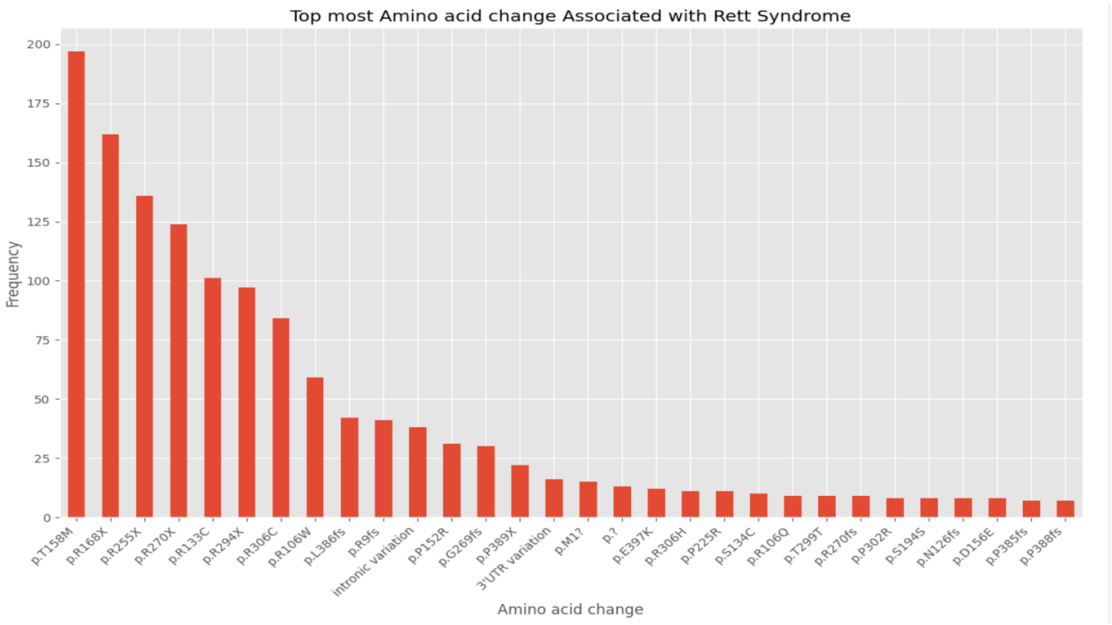


Figure 12. Top 30 Distribution of amino acid change in Rett patients

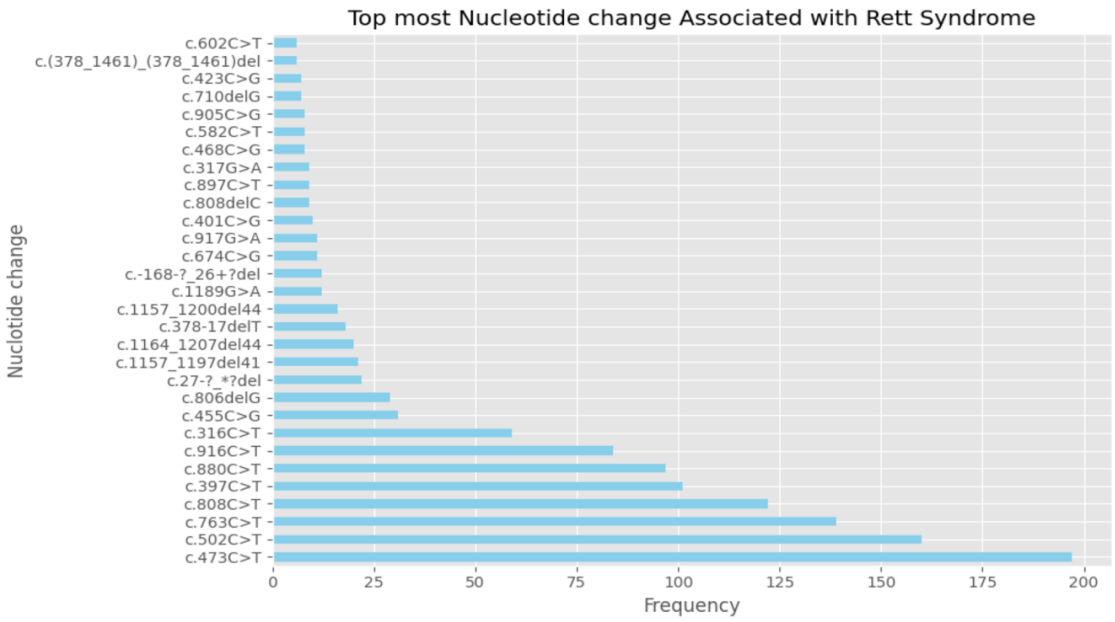


Figure 13. Distribution of topmost 30 Nucleotides changes in Rett patients

Nucleotide changes (Topmost 30): The nucleotide changes listed appear to represent substitutions, duplications, insertion, Silent and complex rearrangements mutations observed in individuals with Rett syndrome. Like as c.573C >T, c.160A>T, c.113C>T, C-455C >G , c.[*8500C>G];[*8503delC] and c.[1157_1197del41; 1232_1240del9], c.[1125_1137del13; 1138_1263inv; 1158_1201del; 1263_1264insGGA], c.-206_-205delGC, c.-146_-138dup9, c.-146_-138dup9, c.-138_-134dupCGCCG and so on.

Mutation//polymorphism: The mutation/polymorphism summarizes variations in genetic sequences as observed from data source. Mutations associated with disease cause pathological conditions, while polymorphisms do not cause disease to have neutral effects. Silent polymorphisms don't alter protein function. Unknown variants await further research for their impact on health.

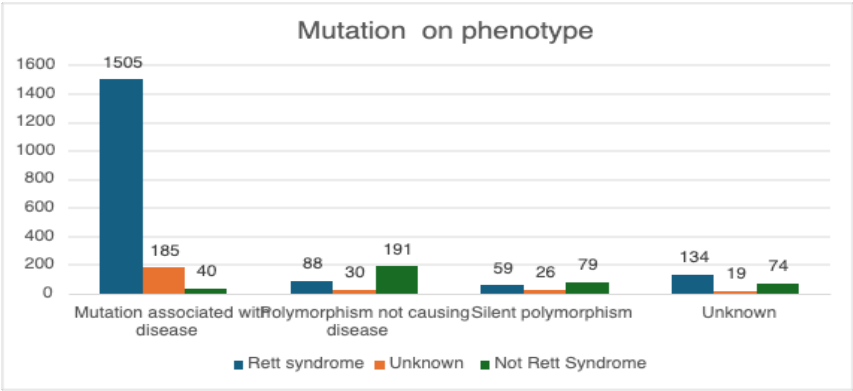


Figure 14. Mutation or polymorphism associated with Phenotypic

3.3 Model Deployment

Various machine learning algorithms were deployed to develop predictive models for Rett syndrome susceptibility. Logistic regression, Random Forest, XGBoost, Support Vector Machines (SVM) and Neutral Network algorithms were utilized. These models were trained using a comprehensive dataset containing genetic variations, clinical features, and demographic information of both RTT patients and healthy controls.

Model	Accuracy
Logistic Regression	76.35871
Random Forest	91.30432
SVM	76.49463
XGBoost	92.79894
Neural Network	79.2120

Table 1. Accuracy of predicated models

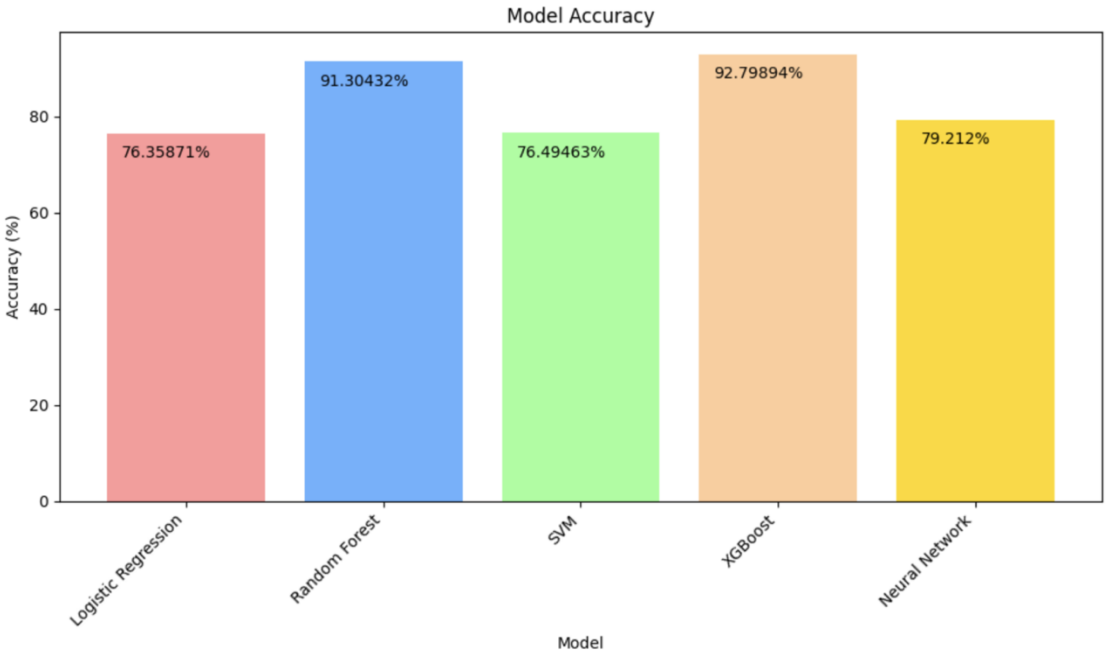


Figure 15. Models by Accuracy

Logistic Regression achieved 76.35% accuracy, Random Forest 91.30%, SVM 76.49%, XGBoost 92.79% and Neural Network 79.21%. The Random Forest and XGBoost models showed robustness, indicating potential for clinical application in early intervention.

These accuracy scores reflect the performance of each model in discriminating between RTT patients and healthy individuals based on the input features. The high accuracy achieved by Random Forest and XGBoost models suggests their effectiveness in capturing the complex relationships between genetic variations, phenotypic traits, and disease susceptibility. Further optimization and validation of these models could enhance their clinical utility in predicting Rett syndrome and facilitating early intervention strategies.

4. Results

When the classification was performed on the testing data, this indicated that the Random Forest and XGBoost performed better in this Rett syndrome classification compared to other models Logistics Regression and Support Vector machine.

These results are from a classification task where different machine learning models were used to predict the presence or absence of Rett Syndrome based on certain features.

```
Logistic Regression Classification Report:
                  precision    recall  f1-score   support

No Rett Syndrome       0.59      0.24      0.34       107
   Rett Syndrome       0.77      0.97      0.86       551
         Unknown       0.00      0.00      0.00        76

        accuracy                           0.76       734
       macro avg       0.45      0.40      0.40       734
    weighted avg       0.67      0.76      0.70       734

Random Forest Classification Report:
                  precision    recall  f1-score   support

No Rett Syndrome       0.81      0.69      0.75       107
   Rett Syndrome       0.93      0.97      0.95       551
         Unknown       0.92      0.86      0.88        76

        accuracy                           0.92       734
       macro avg       0.89      0.84      0.86       734
    weighted avg       0.91      0.92      0.91       734

SVM Classification Report:
                  precision    recall  f1-score   support

No Rett Syndrome       0.00      0.00      0.00       107
   Rett Syndrome       0.75      1.00      0.86       551
         Unknown       0.00      0.00      0.00        76

        accuracy                           0.75       734
       macro avg       0.25      0.33      0.29       734
    weighted avg       0.56      0.75      0.64       734

XGBoost Classification Report:
                  precision    recall  f1-score   support

No Rett Syndrome       0.80      0.69      0.74       107
   Rett Syndrome       0.93      0.97      0.95       551
         Unknown       0.90      0.83      0.86        76
```

```
              accuracy                           0.91       734
             macro avg       0.88      0.83      0.85       734
          weighted avg       0.91      0.91      0.91       734

Neural Network Classification Report:
                        precision    recall  f1-score   support

No Rett Syndrome            0.46      0.54      0.50       107
   Rett Syndrome            0.82      0.90      0.86       551
         Unknown            0.00      0.00      0.00        76

              accuracy                           0.75       734
             macro avg       0.43      0.48      0.45       734
          weighted avg       0.68      0.75      0.72       734
```

Figure 16. Model evaluation

The Random Forest model achieved an accuracy of approximately 0.9046 on the test set, indicating that it correctly classified about 90.46% of the instances.

Here's a breakdown of the classification report: The confusion matrix provides a detailed breakdown of the model's predictions, showing the number of true positives, false positives, and false negatives for each class.

Precision: Precision represents the proportion of true positive predictions out of all positive predictions made by the model. For class "Not Rett Syndrome", precision is 0.78, meaning that out of all instances predicted as "Not Rett Syndrome", 78% were actually "Not Rett Syndrome". For class "Rett Syndrome", precision is 0.92, indicating that 92% of instances predicted as "Rett Syndrome" were actually "Rett Syndrome". For class "Unknown", precision is 0.92, indicating that 92% of instances predicted as "Unknown" were actually "Unknown".

Recall: Recall represents the proportion of true positive predictions out of all actual positive instances in the dataset. Recall measures how well the model identifies instances of each class. For "Not Rett Syndrome," it correctly identified 66% of all actual instances. For "Rett Syndrome," it correctly identified 97% of all actual instances. For "Unknown," it correctly identified 80% of all actual instances.

F1-score: F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. For class "Not Rett Syndrome", the F1-score is 0.72. For class "Rett Syndrome", the F1-score is 0.94. For class "Unknown", the F1-score is 0.86.

Support: Support represents the number of actual occurrences of each class in the test set. The macro average and weighted average F1-scores are 0.84 and 0.90, respectively.

Overall, the Random Forest model demonstrates strong performance across all metrics, with high precision, recall, and F1-scores for each class, as well as a high overall accuracy.

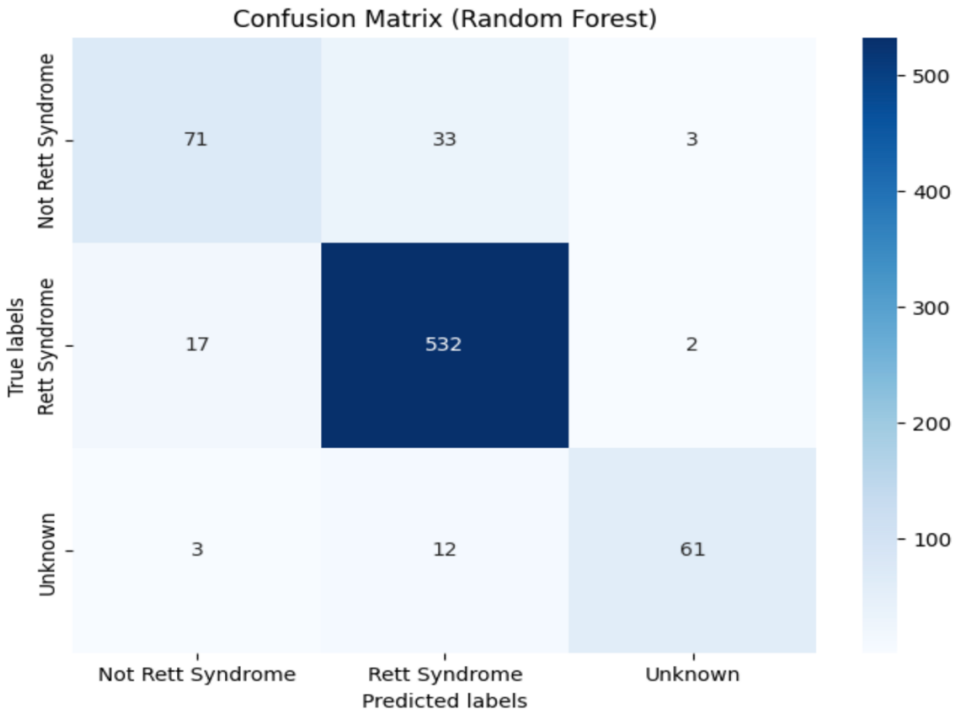


Figure 17. Confusion matrix for the Random Forest model's prediction of Rett Syndrome

In the confusion matrix for the Random Forest model's prediction of "Rett Syndrome":

True Negative (TN): 71 instances correctly predicted as not being "Rett Syndrome."

False Positive (FP): 33 instances incorrectly predicted as "Rett Syndrome" when they are not.

False Negative (FN): 17 instances of "Rett Syndrome" incorrectly predicted as another class.

True Positive (TP): 532 instances correctly predicted as "Rett Syndrome."

The Random Forest model demonstrates strong performance in identifying instances of "Rett Syndrome," as evidenced by the high number of true positives. However, there are still some misclassifications (both false positives and false negatives), indicating areas where the model could potentially improve its predictions.

4.1 Hyperparameter Tuning

Hyperparameter tuning is an essential part of model development to get the model function optimally. In the case of Rett syndrome datasets, we use GridSearchCV to help us find the best way to make our models perform their best, so the models can give us the most accurate predictions possible.

Hyperparameters such as C, kernel, max_depth, min_samples_leaf, min_samples_split, n_estimators, learning rate, regularization strength, number of layers, activation, and neurons per layer can significantly impact the model's performance. By systematically evaluating different combinations of hyperparameters, grid search helps identify the configuration that maximizes predictive accuracy or performance metrics specific to Rett syndrome datasets.

Logistic Regression: The grid search for Logistic Regression found that the best parameters are **{'C': 0.1, 'solver': 'liblinear'}**, and the best cross-validation score achieved with these parameters is approximately 78.9%.

```
Logistic Regression Best Parameters: {'C': 0.1, 'solver': 'liblinear'}
Logistic Regression Best Score: 0.7887611074989164
```

Random Forest: Those above parameters suggest that the Random Forest model performed best with a maximum depth of 20, no constraints on the minimum number of samples required to be at a leaf node (**min_samples_leaf=1**), a minimum number of samples required to split an internal node of 2 (**min_samples_split=2**), and 200 trees (**n_estimators=200**) and the best corss –validation score is 92.82%.

```
Random Forest Best Parameters: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
Random Forest Best Score: 0.9281995015171217
```

Support Vector Machine: The SVM hyperparameter tuning was conducted using GridSearchCV. The dataset consists of four classes with varying numbers of instances, 2-fold cross-validation were applied. After evaluating different combinations, the best parameters for the SVM model were found to be 'C': 0.1 and 'kernel': 'linear' and the best mean cross-validated score achieved with these parameters is approximately 78.6%

```
Class Distribution:
{0: 625, 1: 2874, 2: 343, 3: 2}
Number of splits: 2
Initializing GridSearchCV...
Fitting GridSearchCV...
Fitting 2 folds for each of 6 candidates, totalling 12 fits
[CV] END ................................C=0.1, kernel=linear; total time= 2.7min
[CV] END ................................C=0.1, kernel=linear; total time= 1.8min
[CV] END ..................................C=0.1, kernel=rbf; total time=   0.2s
[CV] END ..................................C=0.1, kernel=rbf; total time=   0.2s
[CV] END ...................................C=1, kernel=linear; total time= 3.2min
[CV] END ...................................C=1, kernel=linear; total time= 4.1min
[CV] END ....................................C=1, kernel=rbf; total time=   0.2s
[CV] END ....................................C=1, kernel=rbf; total time=   0.2s
[CV] END ...............................C=10, kernel=linear; total time= 2.9min
[CV] END ...............................C=10, kernel=linear; total time= 3.0min
[CV] END ....................................C=10, kernel=rbf; total time=   0.2s
[CV] END ....................................C=10, kernel=rbf; total time=   0.2s
SVM Best Parameters: {'C': 0.1, 'kernel': 'linear'}
SVM Best Score: 0.785639958376691
```

XGBoost: The optimal hyperparameters for XGBoost, identified through GridSearchCV, are a learning rate of 0.1, a maximum tree depth of 7, and 100 trees (or estimators). With these settings, the model achieved a score of 93.39%, indicating strong predictive performance.

```
XGBoost Best Parameters: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 100}
XGBoost Best Score: 0.9339401820546163
```

Neural Network: The best settings for the neural network model are using the logistic activation function, a regularization parameter of 0.001, and one hidden layer with 100 neurons. With these settings, the model achieved a score of 81.12%, showing its effectiveness in predictive tasks.

```
Neural Network Best Parameters: {'activation': 'logistic', 'alpha': 0.001, 'hidden_layer_sizes': (100,)}
Neural Network Best Score: 0.8111833550065018
```

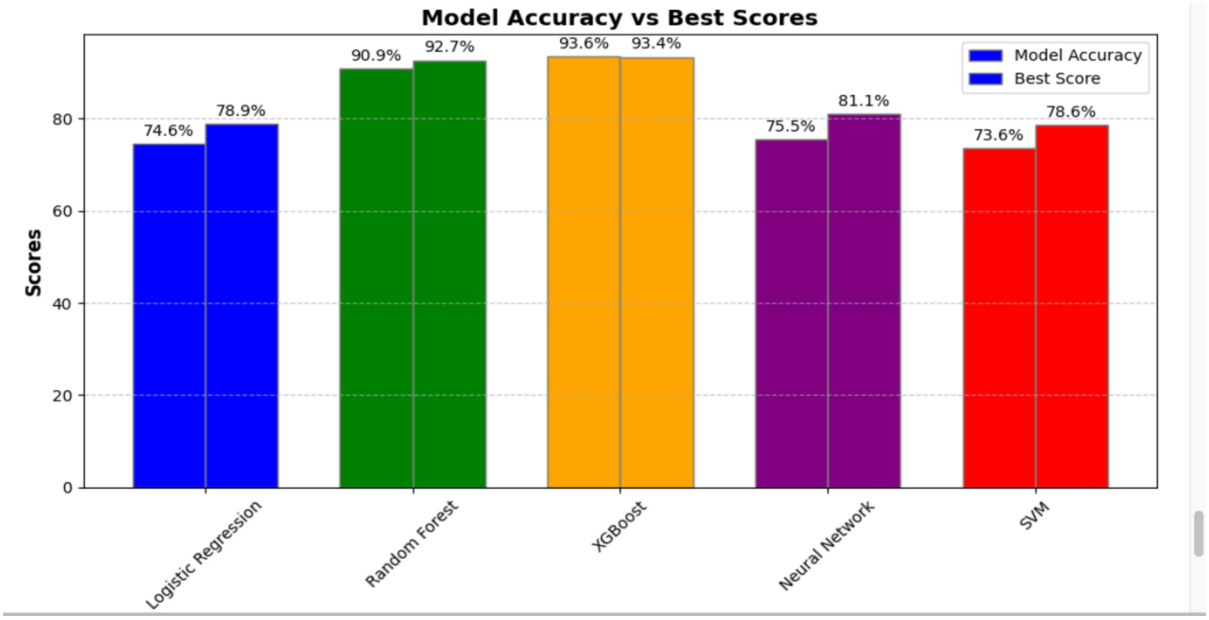


Figure 18. Comparison of models Accuracy Vs Best Score after Parameter tuning

In summary, the accuracy of each model after hyperparameter tuning varies, XGBoost and Random Forest showed strong predictive ability, Neural Network performed well, while Logistic Regression and SVM offered decent accuracy. The performance of each model depends on various factors such as the complexity of the data, the choice of hyperparameters, and the algorithm's inherent strengths and limitations.

5 Conclusion

In this thesis, we aimed to leverage genetic variation and phenotypic characteristics to predict Rett syndrome early and accurately. We identified and analyzed genetic variations associated with Rett syndrome, including nucleotide changes, amino acid changes, types of sequence changes, and mutations/polymorphisms, to understand their contributions to the condition's risk. Additionally, we thoroughly characterized phenotypic features commonly observed in affected individuals, including physical traits, developmental milestones, neurological symptoms, and behavioral patterns.

By integrating ML algorithms to analyze clinical and genetic data, healthcare professionals can improve diagnostic accuracy, tailor treatment strategies, and enhance patient care. ML's ability to identify patterns and genetic variations crucial for understanding and analyzing the condition has paved the way for personalized medicine approaches in Rett syndrome. These models aim to assist researchers, doctors, and clinicians in identifying the risk of the condition and enhancing the quality of life for individuals with Rett syndrome and their families.

Moving forward, our research lays the foundation for several future directions and opportunities in the prediction of Rett syndrome. Firstly, we will focus on refining predictive models by integrating additional data sources and improving machine learning computational algorithms to enhance accuracy and reliability. Ethical considerations regarding genetic testing and patient privacy will continue to be carefully addressed throughout the research process.

Furthermore, we will get more insights into the genes to figure out exactly how changes in our DNA lead to RTT. This isn't just about finding the main gene that causes RTT but also understanding all the other genetic factors that can influence how the disease shows up in different people.

6 References

Hunter, K. (2007). Rett Syndrome Handbook (2nd ed.). International Rett Syndrome Association.

Rosa Angela Fabio, Tindara Caprì, Gabriella (2020). Understanding Rett Syndrome: A guide to symptoms, management and treatment

Neul, J. L., Kaufmann, W. E., Glaze, D. G., Christodoulou, J., Clarke, A. J., Bahi-Buisson, Motil, K. J. (2010). Rett syndrome: revised diagnostic criteria and nomenclature.

Barbro Lindberg (2006). Understanding Rett Syndrome. A practical guide for Parents, Teachers and Therapists.

Hagberg, B. (2002). Clinical manifestations and stages of Rett syndrome. Mental Retardation and Developmental Disabilities Research Reviews.

Rett Syndrome types and phases. https://www.nichd.nih.gov/health/topics/rett/conditioninfo/types/ (Accessed 04.01.2024)

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., & Zoghbi, H.Y. (1999). "Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2." Nature Genetics, 23(2), 185–188.

Rodney C. Samaco and Jeffrey L. Neul. " Complexities of Rett Syndrome and MeCP2. https://www.jneurosci.org/content/31/22/7951 (Accessed 14.01.2024)

J. Clin. Med. 2023, Rett Syndrome medicine Trofinetide 12(15), 5114; https://doi.org/10.3390/jcm12155114 (Accessed 21.01.2024)

Rett Syndrome Research Trust, Genetics Variations in Rett Syndrome. https://reverserett.org/genetics-primer/ (Accessed 09.10.2023)

Machine Learning Enhances Care for Rett Syndrome, Emory University, 2023 https://news.emory.edu/stories/2023/03/hs_machine_learning_rett_syndrome_01-03-2023/story.html/ (Accessed 05.05.2024)

Gene Therapy, Reverse Rett https://reverserett.org/cures/gene-therapy/ (Accessed 03.02.2024)

Treatments for Rett Syndrome,National Institute of Child Health and Human Development https://www.nichd.nih.gov/health/topics/rett/conditioninfo/treatments/ (Accessed 02.02.2024)

Rett Syndrome: Diagnosis & Treatment, Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/rett-syndrome/diagnosis-treatment/drc-20377233 (Accessed 05.05.2024)

Gillian S. Townend*, Theresa E. Bartolotta, Anna Urbanowicz, Helena Wandin, Leopold M.G. Curfs (2020) Rett Syndrome Communication Guidelines: A handbook for therapists, educators and families

Rett Finland Rett Syndrome https://www.rettsyndrome.eu/member-associations/finland/ (Accessed 20.08.2023)

Michael Zats, Feature extraction https://python.plainenglish.io/the-art-of-feature-engineering-techniques-for-creating-better-machine-learning-models-in-python-acd9a2b78f1e/ (Accessed 12.01.2024)

Rett Europe Rett Syndrome https://www.rettsyndrome.eu/rett-syndrome/frequently-asked-questions/ (Accessed 20.12.2023)

Lobo, R. A., Vieira, I. C., Medeiros, A. M., Brito, J. S., Souza, B. S., & Araújo, C. G. (2021). Insights Into Rett Syndrome: Molecular and Functional Data, Genetics, and Epigenetic Mechanisms. Frontiers in Genetics, 12, 620859. https://doi.org/10.3389/fgene.2021.620859/ (Accessed 04.12.2023)

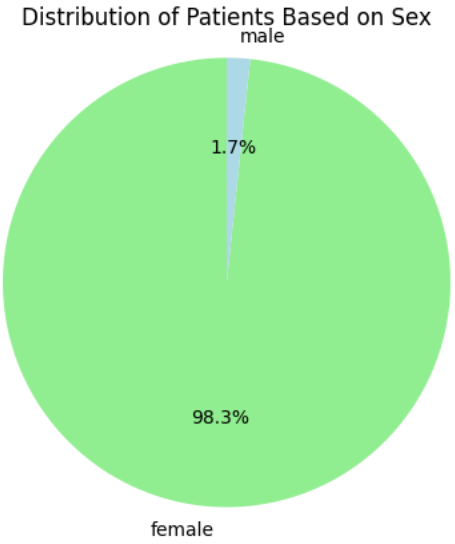
EURORDIS – Rare Diseases Europe Introduction to genetics and genomics https://openacademy.eurordis.org/courses/introduction-to-genetics-and-genomics/ (Accessed 02.05.2024)

Deep Genomics: Genetics https://www.itransition.com/machine-learning/healthcare/ (Accessed 12.05.2024)

Genomic analysis in Chilean patients with suspected Rett syndrome: Diagnostics https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2024.1278198/full/Accessed (12.05.2024)

7 Appendices

Appendices1 Demographics pie charts of Rett syndrome.



Appendices2 Favourite activities done by Rett children.